

RADIO LABELING WITH PREASSIGNED FREQUENCIES*

HANS L. BODLAENDER[†], HAJO BROERSMA[‡], FEDOR V. FOMIN[§],
ARTEM V. PYATKIN[¶], AND GERHARD J. WOEGINGER[‡]

Abstract. A radio labeling of a graph G is an assignment of pairwise distinct, positive integer labels to the vertices of G such that labels of adjacent vertices differ by at least 2. The radio labeling problem (RL) consists in determining a radio labeling that minimizes the maximum label that is used (the so-called span of the labeling). RL is a well-studied problem, mainly motivated by frequency assignment problems in which transmitters are not allowed to operate on the same frequency channel. We consider the special case where some of the transmitters have preassigned operating frequency channels. This leads to the natural variants P-RL(l) and P-RL($*$) of RL with l preassigned labels and an arbitrary number of preassigned labels, respectively.

We establish a number of combinatorial, algorithmical, and complexity-theoretical results for these variants of radio labeling. In particular, we investigate a simple upper bound on the minimum span, yielding a linear time approximation algorithm with a constant additive error bound for P-RL($*$) restricted to graphs with girth ≥ 5 . We consider the complexity of P-RL(l) and P-RL($*$) for several cases in which RL is known to be polynomially solvable. On the negative side, we prove that P-RL($*$) is NP-hard for cographs and for k -colorable graphs where a k -coloring is given ($k \geq 3$). On the positive side, we derive polynomial time algorithms solving P-RL($*$) and P-RL(l) for graphs with bounded maximum degree, and for solving P-RL(l) for k -colorable graphs where a k -coloring is given.

Key words. radio labeling, frequency assignment, graph algorithms, computational complexity, approximation algorithm, cograph, k -coloring

AMS subject classifications. 05C15, 68R10, 05C78, 68W25, 94C15

DOI. 10.1137/S1052623402410181

1. Introduction. The FREQUENCY ASSIGNMENT problem (FAP) is a general framework focused on point-to-point communication, e.g., in *radio* or *mobile telephone* networks. One of its main threads asks for an assignment of frequencies or frequency channels to transmitters while keeping interference at an acceptable level and making use of the available frequencies in an efficient way. Interference constraints are usually related to the use of the same or similar frequencies at locations within a certain distance (or transmitters within a certain reach) from each other. Due to the scarce resources and the increasing use of frequencies in modern wireless technology, the available frequencies should be used as efficiently as possible. There is usually a trade-off between avoiding interference and the efficient use of frequencies. We will not go deeper into the technical details here.

*Received by the editors June 27, 2002; accepted for publication (in revised form) February 23, 2004; published electronically October 1, 2004. The work of the second, third, and fourth authors is partly sponsored by NWO-grant 047.008.006. The work of the third author is supported by Norges forskningsråd projects 160233/V30 and 160778/V30. The work of the first and second authors is partly sponsored by EC contract IST-1999-14186: Project ALCOM-FT (Algorithms and Complexity–Future Technologies). Part of the work was done while the third and fourth authors were visiting the University of Twente.

<http://www.siam.org/journals/siopt/15-1/41018.html>

[†]Institute of Information and Computing Sciences, Utrecht University, 3508 TA Utrecht, The Netherlands (hansb@cs.uu.nl).

[‡]Faculty of Mathematical Sciences, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands (broersma@math.utwente.nl, g.j.woeginger@math.utwente.nl).

[§]Department of Informatics, University of Bergen, N-5020 Bergen, Norway (fomin@ii.uib.no).

[¶]Sobolev Institute of Mathematics, Novosibirsk 630090, Russia (artem@math.nsc.ru).

Graph theoretical issues come into play since possible interference between transmitters is usually modeled by a so-called *interference graph*. Each vertex of the interference graph represents a transmitter. If simultaneous broadcasting of two transmitters may cause an interference, then they are connected by an edge in the interference graph. The frequency channels are usually labeled by positive integers. Frequency channels with “close” labels are assumed to be “close” in the spectrum or expected to be “more likely” to cause interference.

Regarding the assumption that a pair of “close” transmitters should be assigned different frequencies or frequency channels, the FAP is equivalent to the problem of labeling the interference graph with some constraints on the labeling. In many cases the related labeling problems are variants on what is known as the *vertex coloring problem* in graph theory.

However, Hale [14] observed that the signal propagation may affect the interference even in distant regions (but with decreasing intensity). Hence, not only “close” transmitters should get different frequencies, but also frequencies used at some distance should be appropriately separated. This leads to a more detailed and complicated modeling of FAP in terms of *distance constrained labeling* of the interference graph. (See, for instance, the book [16] by Leese.)

In some applications the transmitters are not allowed to operate on the same frequency channel (for example, when every transmitter covers the whole area) while “close” transmitters should use channels with sufficient separation. In this case non-reusable frequency channels should be assigned to transmitters in a proper way. This leads to the so-called RADIO LABELING problem (RL), i.e., to the problem of assigning distinct labels to the vertices of a graph such that adjacent vertices get labels (positive integers) that differ by at least two. The purpose of RL is to find such a radio labeling with the smallest maximum label.

In this paper we initiate the investigation of two versions of this problem in which some of the transmitters (like military and governmental stations) already have preassigned labels corresponding to frequency channels which one is not allowed to change. Then the problem boils down to determining a radio labeling extending a given prelabeling in a “best possible” way. In this paper we consider some algorithmical, complexity-theoretical, and combinatorial aspects of these versions of the problem.

We do not want to claim that the results in what follows have immediate practical relevance.

1.1. Definitions and preliminary observations. We denote by $G = (V, E)$ a finite undirected and simple graph. The *girth* of G is the length of a shortest cycle in G . For every nonempty $W \subseteq V$, the subgraph of G induced by W is denoted by $G[W]$. A *cograph* is a graph containing no induced path on four vertices. The *(open) neighborhood* of a vertex v in a graph G is $N_G(v) := \{u \in V : \{u, v\} \in E\}$. The degree of a vertex v in G is $d_G(v) := |N_G(v)|$. The maximum degree of G is $\Delta(G) := \max_{v \in V} d_G(v)$. A graph G is *t-degenerate* if each of its nonempty subgraphs has a vertex of degree at most t . A *clique* C of a graph G is a subset of V such that all the vertices of C are pairwise adjacent in G . A nonempty subset of vertices $I \subseteq V$ is *independent* in G if no two of its elements are adjacent in G . The *complement* $\overline{G}$ of G is the graph on V with edge set $\overline{E}$ such that $\{u, v\} \in \overline{E}$ if and only if $\{u, v\} \notin E$.

A *k-coloring* of the vertices of a graph $G = (V, E)$ is a partition $I_1, I_2, \dots, I_k$ of V into independent sets (in which some of the I_j may be empty); the k sets I_j are called the *color classes* of the k -coloring. The *chromatic number* $\chi(G)$ is the minimum

value k for which a k -coloring exists. A *labeling* of (the vertex set of) G is an injective mapping $L: V \rightarrow \mathbb{N}^+$ (the set of positive integers). A labeling L of G is called a *radio labeling* of G if for any edge $\{u, v\} \in E$ the inequality $|L(u) - L(v)| \geq 2$ holds; the *span* of such a labeling L is $\max_{v \in V} L(v)$.

The RADIO LABELING problem (RL) is defined as follows: For a given graph G , find a radio labeling L with the smallest span. The name *radio labeling* was suggested by Fotakis and Spirakis in [8], but the same notion (under different names) has been introduced independently and earlier by other researchers (see, e.g., Chang and Kuo [2]). Problem RL is equivalent to the special case of the TRAVELING SALESMAN problem TSP(2,1) in which all edge weights (distances) are either one or two. The relation is as follows. For a graph $G = (V, E)$ let K_G be the complete weighted graph on V with edge weights 1 and 2 defined according to E : For every $\{u, v\} \in E$ the weight $w(\{u, v\})$ in K_G is 2 and for $\{u, v\} \notin E$ the weight $w(\{u, v\}) = 1$. The *weight* of a path in K_G is the sum of the weights of its edges. The following proposition can be found in [8, 9].

PROPOSITION 1.1. *There is a radio labeling of G with span k if and only if there is a Hamiltonian path (i.e., a path on $|V|$ vertices) of weight $k - 1$ in K_G .*

Another equivalent formulation of this problem, which was extensively studied in the literature, is the HAMILTONIAN PATH COMPLETION problem (HPC), i.e., the problem of partitioning the vertex set of a graph G into the smallest possible number of sets which are spanned by paths in G . This equivalence is expressed in the following well-known proposition.

PROPOSITION 1.2. *There is a radio labeling of G with span $\leq n + k - 1$ if and only if there is a partition of V into $\leq k$ sets, such that each of these sets induces a subgraph in $\bar{G}$ that contains a Hamiltonian path.*

Now let us turn to the versions of RL with preassigned labels. For a graph G a *radio prelabeling* (or simply *prelabeling*) L' of a subset $V' \subset V$ is an injective mapping $L': V' \rightarrow \mathbb{N}^+$ such that L' is a radio labeling of $G[V']$. We say that a labeling L of G *extends* the prelabeling L' if $L(u) = L'(u)$ for every $u \in V'$. We study the following two problems.

- **P-RL(*):** RADIO LABELING WITH AN ARBITRARY NUMBER OF PRELABELED VERTICES.
For a given graph G and a given prelabeling L' of G , determine a radio labeling of G extending L' with the smallest span.
- **P-RL(l):** RADIO LABELING WITH A FIXED NUMBER OF PRELABELED VERTICES.
For a given graph G , a subset $V' \subseteq V$ with $|V'| \leq l$, and a prelabeling $L': V' \rightarrow \mathbb{N}^+$, determine a radio labeling of G extending L' with the smallest span.

1.2. Earlier results. As we mentioned above, TSP(2,1) (which is equivalent to RL without any prelabeling) is a well-studied problem. Papadimitriou and Yannakakis [17] proved that this problem is MAX SNP-hard. Later Engebretsen [4] improved their result by showing that the problem is not approximable within $5381/5380 - \varepsilon$ for any $\varepsilon > 0$. An approximation algorithm for TSP(2,1) which finds a solution not worse than $7/6$ times the optimum solution is given in [17].

Damaschke et al. [3] proved that the HPC can be solved in polynomial time on cocomparability graphs (complements of comparability graphs). To obtain this result they used a reduction to the problem of finding the bump number of a partial order. (The bump number of a poset P and its linear extension L is the number

of neighbors in L which are comparable in P .) It was proved by Habib, Möhring, and Steiner [13] and by Schäffer and Simons [19] that the BUMP NUMBER problem can be solved in polynomial time. By Proposition 1.2, the result of Damaschke et al. [3] yields that RL is polynomial time solvable for comparability graphs. Later, this result was rediscovered by Chang and Kuo [2] but under the name of $L'(2, 1)$ -labeling and only for cographs, a subclass of the class of comparability graphs. Notice that RL is NP-hard for cocomparability graphs because the HAMILTONIAN PATH problem is known to be NP-hard for bipartite graphs which form a subclass of comparability graphs. Recently, Fotakis and Spirakis [8] proved that RL can be solved in polynomial time within the class of graphs for which a k -coloring can be obtained in polynomial time (for some fixed k). Note that, for example, this class of graphs includes the well-studied classes of planar graphs and graphs with bounded treewidth.

We are not aware of any existing results concerning the prelabeling versions P-RL($*$) and P-RL(l) of RL.

We complete this subsection by mentioning some results concerning the related notion of radio coloring (also known as $L(2, 1)$ -labeling, $\lambda_{2,1}$ -coloring, and $\chi_{2,1}$ -labeling). A *radio coloring* of a graph $G = (V, E)$ is a function $f: V \rightarrow \mathbb{N}^+$ such that $|f(u) - f(v)| \geq 2$ if $\{u, v\} \in E$ and $|f(u) - f(v)| \geq 1$ if the distance between u and v in G is 2. Here the *distance* between two vertices u and v in a connected graph G is the smallest number of edges in a path of G between u and v . Therefore, the difference between radio coloring and radio labeling is that, in a radio coloring, vertices at distance at least three may have equal labels (or colors). The notion of radio coloring was introduced by Griggs and Yeh [12] under the name $L(2, 1)$ -labeling. As with radio labeling the span of a radio coloring f of G is $\max_{v \in V} f(v)$.

The problem of determining a radio coloring with minimum span has received a lot of attention. For various graph classes the problem was studied by Sakai [18], Bodlaender et al. [1], van den Heuvel, Leese, and Shepherd [15], and others. NP-hardness results for this RADIO COLORING problem (RC) restricted to planar, split, or cobipartite graphs were obtained by Bodlaender et al. [1]. Fixed-parameter tractability properties of RC have been discussed by Fiala, Kratochvíl, and Kloks [6]. Fiala, Fishkin, and Fomin [5] have studied on-line algorithms for RC. For only very few graph classes the problem is known to be polynomially solvable. Chang and Kuo [2] obtained a polynomial time algorithm for RC restricted to trees and cographs. The complexity of RC even for graphs of treewidth 2 is a long-standing open question. An interesting direction of research was initiated by Fiala, Kratochvíl, and Proskurowski [7]. They considered a precolored version of RC, i.e., a version in which some colors are preassigned to some vertices. They proved that RC with a given precoloring can be solved in polynomial time for trees. Recently Golovach [11] proved that RC with a given precoloring is NP-hard for graphs of treewidth 2.

1.3. Our results and organization of the paper. We study algorithmical, complexity-theoretical, and combinatorial aspects of radio labeling with prelabeled vertices. In section 2 we give some simple combinatorial bounds for the minimum span of such labelings: We introduce an easy-to-compute lower bound M on the minimum span of a labeling extending a prelabeling, and we show that there always exists such a radio labeling

- with span $\leq \lfloor (7M - 2)/3 \rfloor$, for arbitrary graphs;
- with span $\leq \lfloor (5M + 2)/3 \rfloor$, for graphs of girth at least 4;
- with span $\leq M + 3$, for graphs of girth at least 5.

All these bounds are best possible. The third bound is even best possible for paths, i.e., for graphs of infinite girth. In section 3 we derive similar results for t -degenerate graphs. In section 4 we obtain polynomial time algorithms for graphs with bounded degree; these algorithms are based on the results of section 2.

Sections 5 and 6 are devoted to the algorithmical study of the radio labeling problem with preassigned labels. We study these problems restricted to graphs for which a k -coloring is given and restricted to cographs, two graph classes for which the radio labeling problem without prelabeling is known to be solvable in polynomial time. Known and new results on these radio labeling problems are summarized in the following table.

	Graphs with a bounded Δ	Graphs with a given k -coloring	Cographs
RL	P [8]	P [8]	P [3, 2]
P-RL(l)	P [*]	P [*]	???
P-RL(*)	P [*]	NP for $k \geq 3$ [*]	NP [*]

In this table, an entry P denotes solvable in polynomial time, NP denotes NP-hard, [*] denotes a contribution from this paper, and ??? marks an open problem.

For the results in the middle column, we assume that k is a fixed integer that is not part of the input. Note that the class of graphs with a given k -coloring contains important and well-studied graph classes such as the class of planar graphs and the class of graphs with bounded treewidth.

2. Upper bounds for the minimum span. Let $G = (V, E)$ denote a graph on n vertices, and let $V' \subseteq V$ and $L': V' \rightarrow \mathbf{N}^+$ be a fixed subset of V and a prelabeling for V' , respectively. We define the parameter M , which will be very useful in the rest of the paper:

$$M := \max \left\{ n, \max_{v \in V'} L'(v) \right\}.$$

Clearly, M is straightforward to compute if G and L' are known, and clearly, M is a lower bound on the span of any radio labeling in G extending the prelabeling L' of G . A natural question is how far M can be away from the minimum span of such a labeling. We will show that the answer to this question heavily relies on the girth of the graph G .

THEOREM 2.1. *Let L' be a prelabeling of a graph G . Then there is a radio labeling in G extending L'*

- (a) *with $\text{span} \leq \lfloor (7M - 2)/3 \rfloor$;*
- (b) *with $\text{span} \leq \lfloor (5M + 2)/3 \rfloor$ if G has girth at least 4;*
- (c) *with $\text{span} \leq M + 3$ if G has girth at least 5.*

All these bounds are best possible. The third bound is even best possible for the class of paths.

Proof. Let us start the proof of Theorem 2.1 by showing that all the stated bounds indeed are best possible: For (a), let x be a positive integer and y an integer such that $M = 3x + y$ and $-1 \leq y \leq 1$. We consider the complete graph on M vertices in which x vertices are prelabeled with labels $2, 5, 8, \dots, 3x - 1$, whereas the remaining $2x + y$ vertices are unlabeled. Since we cannot use the labels $1, 2, \dots, 3x$ at the unlabeled vertices, and the labels at these vertices have to differ by at least two, the span of any radio labeling extending the prelabeling is at least $3x + 2(2x + y) - 1 = 7x + 2y - 1 = \lfloor (7M - 2)/3 \rfloor$.

For (b), let x be a positive integer and y be an integer such that $M = 3x + y$ and $0 \leq y \leq 2$. We consider the complete bipartite graph $K_{x+1, 2x+y-1}$ on M vertices. The $x + 1$ vertices in the first part of the bipartition are prelabeled with the labels $2, 5, 8, \dots, 3x - 1$ and with $3x + y$, whereas the $2x + y - 1$ vertices in the second part of the bipartition are unlabeled. Note that the prelabeling forbids the labels $1, \dots, 3x + y + 1$ for the unlabeled vertices. Thus the span of any radio labeling extending the prelabeling is at least $5x + 2y = \lfloor (5M + 2)/3 \rfloor$.

Finally, for (c) let $M \geq 7$, and let x and y be integers such that $M = 2x + y + 7$ and $0 \leq y \leq 1$. We consider the path $v_1 - v_2 - \dots - v_M$. The prelabeling assigns $L'(v_k) = 2k + 7$ for $1 \leq k \leq x$, and $L'(v_{x+7+k}) = 2k + 6$ for $1 \leq k \leq x + y$. Furthermore, we have $L'(v_{x+1}) = 7$, $L'(v_{x+3}) = 3$, $L'(v_{x+5}) = 5$, and $L'(v_{x+7}) = 1$. Then all but the three vertices v_{x+2} , v_{x+4} , v_{x+6} are prelabeled. Since we cannot use the labels 2, 4, and 6 at these three vertices, every radio labeling of this path extending the prelabeling L' has a span of at least $M + 3$.

The greedy preprocessing step. In several of our proofs, we use the same preprocessing step. We denote this step as *greedy preprocessing*.

Greedy preprocessing is done as follows. We extend the prelabeling by assigning labels from $\{1, \dots, M\}$ to unlabeled vertices, such that all labels are different and such that adjacent vertices have labels that differ by at least two. This preprocessing step terminates when we get stuck: Then either all vertices have been labeled or, for every unused label $c \in \{1, \dots, M\}$ and for every unlabeled vertex v , v is adjacent to a vertex labeled with $c - 1$ or v is adjacent to a vertex labeled with $c + 1$. Note that this greedy preprocessing step does not change the value of M .

Blocked labels. Let $L': V' \rightarrow \mathbb{N}^+$ be a prelabeling of a graph G . We call a label c *blocked* if either it is used in the prelabeling or every vertex of $V \setminus V'$ is adjacent to a vertex labeled by $c - 1$ or $c + 1$. Thus label c cannot be used for extending of L' .

Proof of Theorem 2.1(a). It is sufficient to consider the case where G is the complete graph on n vertices. Assume that there are $l = |V'|$ prelabeled vertices and that we start with the available labels $1, 2, \dots, N = \lfloor (7M - 2)/3 \rfloor$. Hence there are at most $\min\{3l, M + 1\}$ blocked labels. Order the remaining available (nonblocked) labels increasingly, and assign the labels at the odd positions in the ordering to the remaining $n - l$ vertices. Note that we need at most $\min\{3l, M + 1\} + 2(n - l) - 1$ labels. If $3l \leq M + 1$, then this number is at most

$$3l + 2(n - l) - 1 = 2n + l - 1 \leq 2M + \frac{M + 1}{3} - 1 = \frac{7M - 2}{3}.$$

If $3l \geq M + 1$, then this number is at most

$$M + 1 + 2(n - l) - 1 = M + 2n - 2l \leq M + 2M - 2\frac{M + 1}{3} = \frac{7M - 2}{3}.$$

Therefore, in both cases we obtain a feasible radio labeling with span at most $\lfloor (7M - 2)/3 \rfloor$.

Proof of Theorem 2.1(b). We start with $N = \lfloor (5M + 2)/3 \rfloor$ available labels $1, 2, \dots, N$, perform the greedy preprocessing described above, and from now on consider the labeling thus obtained. If all vertices are labeled, then we are done and there is nothing to show. Otherwise, consider some fixed label $c \in \{1, \dots, M\}$ that is not used in this labeling. Then every unlabeled vertex v must be adjacent to a vertex labeled with $c + 1$ or to a vertex labeled with $c - 1$. Denote the set of unlabeled vertices

adjacent to $c - 1$ by A , and denote the set of unlabeled vertices that are adjacent to $c + 1$ but not to $c - 1$ by B . Then A must be an independent set. (Any edge in A together with the vertex labeled with $c - 1$ would induce a triangle and bring the girth of G down to 3.) Also B must be an independent set.

Assume that l vertices are labeled. Since every label blocks at most two other labels, there are at most $\min\{3l, M + 1\}$ blocked labels and we have at least

$$\lfloor (5M + 2)/3 \rfloor - \min\{3l, M + 1\} \geq n - l$$

available labels for the remaining $n - l$ vertices. The displayed inequality can be seen as follows. If $3l \leq M + 1$, then $\lfloor (2M + 2)/3 \rfloor \geq 2l$. Adding $M - 3l \geq n - 3l$ yields the desired inequality in this case. If $3l \geq M + 2$, then $n - l \leq M - \lceil (M + 2)/3 \rceil$. Together with $M - \lceil (M + 2)/3 \rceil \leq \lfloor (5M + 2)/3 \rfloor - M - 1$, we get the desired inequality also for this second case.

Now we distinguish three subcases. In the first subcase, there are vertices $a \in A$, $b \in B$, such that a is not adjacent to b . We assign the $|A|$ smallest of the $n - l$ available labels to the vertices in A and the $|B|$ largest of these labels to the vertices in B . This is done in such a way that vertex a receives the largest label in A , and such that vertex b receives the smallest label in B . This gives a radio labeling with span $\leq \lfloor (5M + 2)/3 \rfloor$.

In the second subcase, either A or B is empty. In this case, we use the $n - l$ available labels on the $n - l$ vertices in A or B .

In the third subcase, we assume that none of A and B is empty and that they span a complete bipartite graph, i.e., each vertex in A is adjacent to each vertex in B . Consider an arbitrary label $d \in \{1, \dots, M\}$ that is not used in the prelabeling. Then every unlabeled vertex $v \in A \cup B$ must be adjacent to the vertex labeled with $d + 1$ or to the vertex labeled with $d - 1$. There are only two possibilities for this: Either all vertices in A are adjacent to the vertex labeled with $d + 1$ and all vertices in B are adjacent to the vertex labeled with $d - 1$, or all vertices in A are adjacent to the vertex labeled with $d - 1$ and all vertices in B are adjacent to the vertex labeled with $d + 1$. As a consequence, there are at most $\min\{2l, M + 1\}$ blocked labels in this subcase, and at least

$$\lfloor (5M + 2)/3 \rfloor - \min\{2l, M + 1\} \geq n - l + 1$$

available labels for the remaining $n - l$ vertices. We assign the $|A|$ smallest of these labels arbitrarily to the vertices in A and the $|B|$ largest of these labels arbitrarily to the vertices in B .

Proof of Theorem 2.1(c). We start with the greedy preprocessing described above, and from now on consider the labeling thus obtained. Denote by $C = \{c_1, \dots, c_k\}$ with $c_1 < c_2 < \dots < c_k$ the set of all unused labels from $\{1, \dots, M\}$ in the labeling obtained after the preprocessing. Denote by U the set of unlabeled vertices. It is clear that $|U| \leq |C|$. We prove that the labeling obtained by the greedy preprocessing can always be extended to a radio labeling of G using at most three additional labels $M + 1$, $M + 2$, and $M + 3$.

Without loss of generality, we assume that after the greedy preprocessing there is a vertex of G labeled with M . (If this is not the case, the same proof produces a span $\leq M + 2$.) For the sake of convenience we shall identify each labeled vertex with its label. If $|U| = 1$, we just label the only unlabeled vertex with label $M + 2$. If $|U| = 2$, then either the two unlabeled vertices are nonadjacent and we can label them with

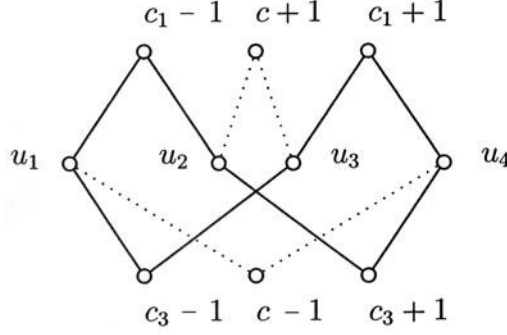


FIG. 1. A step in the proof of Theorem 2.1(c).

$M + 2$ and $M + 3$, or at least one of them is not adjacent to M (since G contains no 3-cycles), and we label this vertex with $M + 1$ and the other vertex with $M + 3$.

From now on we assume $|U| \geq 3$. If C contains two consecutive labels d and $d + 1$, then every vertex in U must be adjacent to the vertices labeled with $d - 1$ and $d + 2$. This yields a cycle of length four in G and contradicts the assumption on the girth of G . Therefore, there are no consecutive labels in C , and in particular $c_3 - 1 > c_1 + 1$.

Next, we first discuss the case $|U| \geq 4$. Then $|C| \geq 4$ and $c_3 < M$. Suppose for the sake of contradiction that $c_1 = 1$. Then all vertices in U are adjacent to label 2, and at least two vertices of U are adjacent to $c_3 - 1$ or to $c_3 + 1$. This would yield a cycle of length four. This contradiction shows $c_1 \geq 2$. There cannot be more than two vertices of U adjacent to $c_1 - 1$ because otherwise at least two of them would be adjacent to $c_3 + 1$ or to $c_3 - 1$, and we again would obtain a 4-cycle. Similarly, we see that at most two vertices of U are adjacent to $c_1 + 1$. As a consequence, each of the vertices $c_1 - 1$ and $c_1 + 1$ has exactly two neighbors in U . This implies that $|U| = 4$ and that each vertex of U is adjacent to exactly one of $c_1 - 1$ and $c_1 + 1$. Denote the four vertices in U by u_1, u_2, u_3, u_4 such that u_1 and u_2 are the neighbors of $c_1 - 1$, and u_3 and u_4 are the neighbors of $c_1 + 1$. Moreover, we may assume that u_1 and u_3 are adjacent to $c_3 - 1$ and that u_2 and u_4 are adjacent to $c_3 + 1$; see Figure 1.

For every label $c \in C \setminus \{c_1, c_3\}$ we have that $|\{c - 1, c + 1\} \cap \{c_1 - 1, c_1 + 1, c_3 - 1, c_3 + 1\}| = 0$. Otherwise, we either obtain a 4-cycle or can use the label c at one of the vertices in U ; as an example, consider the case that $c - 1 = c_1 + 1$: Then we can use the label c at u_1 or u_2 unless both are adjacent to $c_1 + 3$, yielding a 4-cycle with $c_1 - 1$; the other cases are similar. However, then there are only two possibilities for the vertices in U to be the neighbor of $c - 1$ or $c + 1$ without creating a 4-cycle: u_1 and u_4 should be adjacent to one of these labels and u_2 and u_3 to the other. Since $|C \setminus \{c_1, c_3\}| \geq 2$ and C does not contain consecutive integers, we obtain a 4-cycle in G , which is a contradiction. Therefore, the case $|U| \geq 4$ cannot occur at all.

We are left with the case that $|U| = 3$. By similar arguments as above, we may assume that $c_1 - 1$ is adjacent to u_1 , that $c_1 + 1$ is adjacent to u_2 and u_3 , that $c_3 + 1$ is adjacent to u_1 and u_2 , and that $c_3 - 1$ is adjacent to u_3 (the other cases are analogous). The girth condition implies that the graph induced by the vertices $\{u_1, u_2, u_3\}$ has at most one edge, the edge $\{u_1, u_3\}$. If such an edge exists, then M is nonadjacent to one of these two vertices, and we label this vertex with $M + 1$ and the other one with $M + 3$; in that case the vertex u_2 is labeled with $M + 2$. If there is no such edge, then M is nonadjacent to at least one of the vertices u_i (otherwise we obtain a

4-cycle) and we label this vertex with $M + 1$ and the other two vertices with $M + 2$ and $M + 3$. This completes the proof of Theorem 2.1(c). $\square$

Note that the proof of Theorem 2.1(c) yields a linear time approximation algorithm determining a radio labeling extending a prelabeling in graphs with girth ≥ 5 with a span that is at most an additive 3 away from the minimum span.

3. Graphs with bounded degeneracy. In this section, we apply the techniques from the proof of Theorem 2.1 to t -degenerate graphs, i.e., graphs with the property that each of their nonempty subgraphs has a vertex of degree at most t . We will also use the easy fact that a t -degenerate graph on n vertices has at most $t(n - t) + t(t - 1)/2$ edges. First we consider the case without any prelabeling and obtain the following result, which is easy to prove. This result will be used in the case where a prelabeling is assumed.

LEMMA 3.1. *If G is a t -degenerate graph on n vertices, then it has a radio labeling with $\text{span} \leq n + 2t$.*

Proof. Let G be a counterexample to Lemma 3.1 with the least number of vertices. Remove a vertex v of minimum degree from G and consider a radio labeling with $\text{span} \leq n - 1 + 2t$ in $G - v$ (which exists by the choice of G). If we started with the $n + 2t$ available labels $1, 2, \dots, n + 2t$, there are $2t + 1$ unused labels left. Since v has degree at most t and each neighbor of v forbids the use of at most two labels at v , we can choose a suitable label for v . $\square$

We now turn to the variant in which a prelabeling is assumed. We obtain an upper bound for the minimum span of a radio labeling extending the prelabeling in a t -degenerate graph, depending on M and t only.

THEOREM 3.2. *If G is a t -degenerate graph and L' is a prelabeling of G , then there exists a radio labeling extending L' with $\text{span} \leq M + (4 + \sqrt{3})t + 1$.*

Proof. We again start with the greedy preprocessing step as described in the preceding section, and with available labels $1, 2, \dots, M$. Denote by U the set of unlabeled vertices after this step. We only consider the case where $|U| = 2p$ for some positive integer p ; the case with odd $|U|$ is similar and left to the reader. Denote by $C = \{c_1, \dots, c_{2p}\}$ with $c_1 < \dots < c_{2p}$ the set of the first $2p$ unused labels. It is clear that C cannot contain three consecutive labels $c - 1, c, c + 1$ (otherwise, we can use the label c at some vertex of U). Thus for every i we have $c_i + 1 < c_{i+2} - 1$. Consider the set C' of vertices labeled by $c_i - 1$ and $c_i + 1$ for every odd $i = 1, 3, \dots, 2p - 1$. Then C' contains at most $2p$ vertices and each unlabeled vertex of U must be adjacent to a vertex labeled by $c_i - 1$ or $c_i + 1$ for every odd $i = 1, 3, \dots, 2p - 1$. The graph induced by the set $U \cup C'$ has at most $4p$ vertices. Denote the number of edges of this graph by x . We have $2p^2 \leq x \leq (4p - t)t + (t - 1)t/2 = 4pt - t^2/2 - t/2$. Thus $(p - t)^2 \leq 3t^2/4 - t/4 \leq 3t^2/4$, yielding $2p \leq (2 + \sqrt{3})t$. Using Lemma 3.1 and the fact that $G[U]$ is t -degenerate, we can label the unlabeled vertices of U with the labels $\{M + 2, \dots, M + \lfloor (4 + \sqrt{3})t \rfloor + 1\}$. This completes the proof. $\square$

The above results imply a polynomial time approximation algorithm for solving the radio labeling problem (with prelabeling) in t -degenerate graphs. It is possible that the bound in Theorem 3.2 can be improved considerably. We pose the following conjecture.

CONJECTURE. *If G is a t -degenerate graph and L' is a prelabeling of G , then there exists a radio labeling extending L' with $\text{span} \leq M + 3t$.*

The upper bound in the above conjecture cannot be improved. For $t = 1$ this is clear from earlier examples. (See the discussion on case (c) of Theorem 2.1.) Let us show it here for general t by the following example. Let V be the union of disjoint sets

A , B , C , and D , with $|A| = |B| = |C| = t$, $|D| = t + 1$. Let A be a clique, and assume that each vertex of $B \cup C$ is adjacent to every vertex from A . Finally assume $B \cup D$ induces a complete bipartite graph with bipartition (B, D) . Let G be the graph on V with the described edges. One easily checks that G is a t -degenerate graph. Prelabel the vertices of C with labels $3, 7, \dots, 4t - 1$ and of D with labels $1, 5, \dots, 4t + 1$. Then $M = 4t + 1$ and none of the even labels less than M can be used for labeling $A \cup B$. Since each vertex of A is adjacent to all other vertices of $A \cup B$, we lose t labels more, and the largest label we have to use is at least $7t + 1 = M + 3t$.

4. Graphs with a bounded maximum degree. We now turn to graphs with a bounded maximum degree. Using similar proof techniques as in the previous section, we will prove that P-RL(*) is polynomially solvable within this class of graphs. We first prove the next technical result on t -degenerate graphs.

THEOREM 4.1. *Let $G = (V, E)$ be a t -degenerate graph with maximum degree Δ and let $V' \subseteq V$ be the set of vertices that is prelabeled by L' . If the number $p = |V \setminus V'|$ of unlabeled vertices satisfies $p \geq 4\Delta(t + 1)$, then L' can be extended to a radio labeling of G with span M .*

Proof. Let $H = G[V \setminus V']$ be the graph induced by the unlabeled vertices, and start with the available labels $1, 2, \dots, M$. As long as H has edges we will apply the following labeling procedure. Consider a vertex v of minimal positive degree in H . It has at most t neighbors. We can label them all with unused labels because there are at most $2\Delta - 2$ blocked labels among the unused labels and the number of available labels in every step is at least the number of unlabeled vertices at that moment, which is always at least 4Δ . Adapt H by removing the neighbors of v . In each step we reduce the number of unlabeled vertices by at most t and increase the number of isolated vertices by one. Since $p \geq 4\Delta(t + 1)$, we have $q \geq 4\Delta$ isolated vertices in H when H becomes edgeless.

We now show that the set Q of the remaining q isolated vertices of H can be labeled with unused labels from the set $\{1, \dots, M\}$. Denote the set of such labels by C (it is clear that $|C| \geq q$) and consider the auxiliary bipartite graph G' with vertex partition $Q \cup C$ where an edge $\{v, c\}$ exists if and only if the vertex v of Q can be labeled with the label c . It is sufficient to show that G' has a matching saturating all vertices of Q . Suppose that there is no such matching. Then by standard matching theory there is a set $A \subset Q$ such that $|A| = a$ and $|N(A)| \leq a - 1$ (where $N(A)$ is the neighborhood of A in G'). Let $B = C \setminus N(A)$. We have $|B| \geq q - a + 1$. Note that each label could be forbidden for at most 2Δ vertices and, vice versa, that for every vertex at most 2Δ labels could be forbidden. Therefore, since there are no edges between A and B , we have $a \leq 2\Delta$ and $q - a + 1 \leq |B| \leq 2\Delta$, but this implies $q \leq 4\Delta - 1$, which is a contradiction. This completes the proof. $\square$

From Theorem 4.1 we easily obtain the following complexity result for graphs with a bounded maximum degree.

COROLLARY 4.2. *Let k be a fixed positive integer. For every graph G with maximum degree $\Delta \leq k$ and prelabeling L' , P-RL(*) can be solved in polynomial time.*

Proof. Each graph with maximum degree Δ is clearly Δ -degenerate. If at most $4\Delta(\Delta + 1)$ vertices of G are not prelabeled by L' , then one can use a brute force algorithm to find a radio labeling extending L' with a minimum span, e.g., by checking all admissible labelings. The time complexity of such a brute force algorithm is $O(n^{4\Delta(\Delta+1)})$. If more than $4\Delta(\Delta + 1)$ vertices are unlabeled, then by Theorem 4.1 there is a radio labeling extending L' with span M (which is clearly the minimum), and from the proof of Theorem 4.1 it is not difficult to check that such a labeling can

be found in polynomial time. $\square$

The above corollary shows that $P\text{-RL}(l)$ and $P\text{-RL}(*)$ have the same complexity behavior as RL for graphs with a bounded maximum degree, i.e., all three of the problems can be solved in polynomial time. This picture changes if we restrict ourselves to graphs which are k -colorable and for which a k -coloring is given (as part of the input) for some fixed positive integer k . This is the topic of the next section.

5. Graphs with a bounded chromatic number. In this section we concentrate on radio labeling algorithms for graphs with a bounded chromatic number, in particular for the case where a k -coloring of the graph (for a fixed constant k) is provided as part of the input.

Related to Proposition 1.1 we discussed the useful equivalence between RL and $\text{TSP}(2,1)$. We now adapt this equivalence to capture the restrictions of the prelabeling problem. Let L be a labeling of a graph $G = (V, E)$ on n vertices. The path $P = (v_1, v_2, \dots, v_n)$ corresponding to L visits the vertices by increasing labels, i.e., for all $1 \leq a < b \leq n$ we have $L(v_a) < L(v_b)$. P is a path in the complete graph K_G ; its weight $w(P)$ is measured according to the edge weights w in K_G as introduced in the paragraph preceding Proposition 1.1.

LEMMA 5.1. *Let L' be a prelabeling of a subset $V' \subseteq V$ of a graph $G = (V, E)$, and let $P = (v_1, v_2, \dots, v_n)$ be an ordering of the vertices of G . Then the path P in K_G corresponds to some extension L of L' to V if and only if the following two conditions are satisfied:*

- (T1) *For any $1 \leq a < b \leq n$ with $v_a, v_b \in V'$, the weight $w(v_a, v_{a+1}, \dots, v_b)$ of the subpath from v_a to v_b is at most $L'(v_b) - L'(v_a)$.*
- (T2) *Let c be the smallest index with $v_c \in V'$. If $c \neq 1$, then the weight $w(v_1, v_2, \dots, v_c)$ of the subpath from v_1 to v_c is at most $L'(v_c) - 1$.*

For any path P that satisfies (T1) and (T2), a labeling L extending L' with the smallest possible span can be computed in polynomial time $O(n)$.

Proof. Suppose P corresponds to a labeling L that extends L' . Let $a < b$, $v_a, v_b \in V'$. Consider the path $v_a, v_{a+1}, \dots, v_b$. By induction, we see that for all i with $0 \leq i \leq b-a$, $L(v_{a+i})$ is at least the sum of $L(v_a)$ and the weight $w(v_a, v_{a+1}, \dots, v_{a+i})$ of the subpath from v_a to v_{a+i} . Therefore, $L(v_b) \geq L(v_a) + w(v_a, v_{a+1}, \dots, v_b)$, and (T1) follows. In a similar way, it can be shown that (T2) must hold.

The “if”-statement remains to be proved. Consider a path P that satisfies (T1) and (T2). We construct the following labeling L : If $v_1 \in V'$, then we set $L(v_1) = L'(v_1)$, and if $v_1 \notin V'$, then we set $L(v_1) = 1$. For $i \geq 2$ and $v_i \in V'$, we set $L(v_i) = L'(v_i)$. Finally for $i \geq 2$ and $v_i \notin V'$, we set $L(v_i) = L(v_{i-1}) + 1$ if $\{v_i, v_{i-1}\} \notin E$ and we set $L(v_i) = L(v_{i-1}) + 2$ otherwise.

It can be seen that the above procedure in fact computes a labeling L with the smallest possible span for P among all labelings that extend L' . $\square$

Now let $G = (V, E)$ be a graph with a given k -coloring with color classes $I_1, I_2, \dots, I_k$. Let L' be a prelabeling of a subset $V' \subseteq V$ with $|V'| = l$. Let L be a radio labeling of G that extends L' , and let $P = (v_1, v_2, \dots, v_n)$ be the path corresponding to L . Consider two consecutive vertices v_a and v_{a+1} along this path with $v_a \in I_i$ and $v_{a+1} \in I_j$. If $i = j$, then v_a and v_{a+1} form a *monochromatic* edge. If $i \neq j$, then v_a and v_{a+1} form a *bichromatic* edge of type (i, j) .

LEMMA 5.2. *Let $G = (V, E)$ be a graph with a k -coloring with color classes $I_1, I_2, \dots, I_k$. Let L' be a prelabeling of a subset $V' \subseteq V$ with $|V'| = l$. Then there exists a radio labeling L of G extending L' with the smallest possible span, such that its corresponding path P satisfies the following two conditions:*

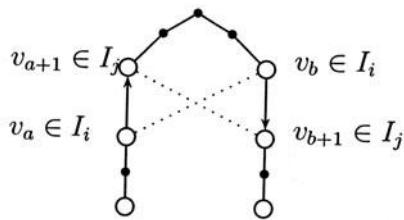


FIG. 2. A step in the proof of property (P1) in Lemma 5.2.

(P2) P contains at most $(l+1)k(k-1)$ bichromatic edges.

Suppose that this path P violates property (P1). Then there exist two bichromatic edges $\{v_a, v_{a+1}\}$ and $\{v_b, v_{b+1}\}$ of type (i, j) with $a < b$ and $i \neq j$, such that none of the vertices on the subpath $(v_{a+1}, \dots, v_b)$ are pre-labeled; see Figure 2. We replace the subpath $(v_a, v_{a+1}, v_{a+2}, \dots, v_{b-1}, v_b, v_{b+1})$ of P by $(v_a, v_b, v_{b-1}, v_{b-2}, \dots, v_{a+1}, v_{b+1})$, and thus produce a new path P^* . Since the edges $\{v_a, v_b\}$ and $\{v_{a+1}, v_{b+1}\}$ are monochromatic, whereas the edges $\{v_a, v_{a+1}\}$ and $\{v_b, v_{b+1}\}$ were not, the new path P^* has more monochromatic edges than P . Moreover, if we compute a labeling L^* for P^* as indicated in the proof of Lemma 5.1, then the span of L^* is at most the span of L . This contradicts our choice of the labeling L . Hence, P satisfies property (P1).

Property (P2) is a simple quantitative consequence of property (P1): The pre-labeled vertices split P into $l + 1$ subpaths, and any such subpath contains at most one bichromatic edge of type (i, j) . Since there are only $k(k - 1)$ possible types of bichromatic edges, the bound in property (P2) follows. $\square$

THEOREM 5.3. *Let $G = (V, E)$ be a graph with a given k -coloring with color classes $I_1, I_2, \dots, I_k$. Let L' be a prelabeling of a subset $V' \subseteq V$ with $|V'| = l$. Then a radio labeling L of G extending L' with the smallest possible span can be computed in time $O(n^{4(l+1)k(k-1)})$.*

Proof. The proof is based on Lemma 5.2. The idea is to use a brute force algorithm that enumerates all possible subsets of bichromatic edges as described in property (P2) and then tries to extend them to a radio labeling. This is done in the following way.

1. Compute all possible subsets S of at most $(l+1)k(k-1)$ bichromatic edges.
2. For each such subset S , compute all possible functions $f : S \rightarrow \{1, \dots, n-1\}$. The interpretation of $f(v_a, v_{a+1}) = m$ is that v_a is the m th vertex and that v_{a+1} is the $(m+1)$ th vertex on the path.
3. Compute all possible functions $g : V' \rightarrow \{1, \dots, n\}$. The interpretation of $g(v_c) = m$ is that v_c is the m th vertex on the path.
4. For each subset S and all functions f and g , determine whether there exists a compatible path. If such a path exists, then compute a corresponding labeling with the smallest possible span according to Lemma 5.1.
5. Output the detected labeling with the minimum span.

There are $O(n^{2(l+1)k(k-1)})$ possible subsets S in the first step, and there are $O(n^{(l+1)k(k-1)})$ possible functions f in the second step. Furthermore, there are $O(n^l)$ possible functions g in the third step. Below, we will show that the fourth step

can be performed in $O(n^2)$ time. This yields an overall time complexity of at most $O(n^{4(l+1)k(k-1)})$.

How do we check compatibility in the fourth step? If any vertex is assigned to two or more distinct positions (for instance, since it is incident with two or more bichromatic edges in S), then there clearly exists no compatible path. Also, if there are two or more distinct vertices assigned to the same position (for instance, by f and by g), then there clearly exists no compatible path. Therefore, from now on we will assume that every vertex goes to at most one position, and any position receives at most one vertex. Then the functions f and g fully determine the positions of all prelabeled vertices and the positions of all vertices on bichromatic edges in the corresponding path. What about the empty positions? Consider a maximal piece Z of empty positions, and let z_ℓ and z_r be the vertices assigned to the position immediately to the left of Z and immediately to the right of Z , respectively. Then z_ℓ and z_r must belong to the same color class I_j . Otherwise, there is a missing bichromatic edge, and there cannot be a compatible path. Therefore, we will assume that z_ℓ and z_r belong to the same color class I_j ; then the vertices that go into positions in Z must all be in I_j . This determines the color class for all empty positions.

We randomly assign the remaining vertices to the empty positions, subject to the condition that every empty position receives a vertex of the right color class. If no such assignment exists, then there cannot be a compatible path. Otherwise, all these random assignments yield a path with the same sequence of edge weights w in K_G . We finally check whether conditions (T1) and (T2) of Lemma 5.1 are satisfied by the resulting path and use this lemma to compute a corresponding labeling with the smallest possible span. $\square$

For each of the graph classes in the following corollary, it is possible to construct a vertex coloring with a constant number of colors in polynomial time. Hence we have the following.

COROLLARY 5.4. *The radio labeling problem P-RL(l) is polynomially solvable*

- *on the class of planar graphs;*
- *on any class of graphs of bounded treewidth;*
- *on the class of bipartite graphs.*

The above results show that P-RL(l) is solvable in polynomial time for graphs with a bounded chromatic number and a given coloring. This result does not carry over to the more general labeling problem P-RL($*$) where the number of prelabeled vertices is part of the input. We next show that P-RL($*$) is NP-hard even when restricted to 3-colorable graphs with a given 3-coloring; this result then easily generalizes to k -colorable graphs ($k \geq 4$) with a given k -coloring. We use a polynomial time transformation from the following problem.

PARTITION INTO TRIANGLES

Instance: A graph $G = (V, E)$ with $|V| = 3q$ for a positive integer q .

Question: Is there a partition of V into triangles, i.e., into $V_1, V_2, \dots, V_q$ such that $G[V_i] = K_3$?

This problem remains NP-hard even when the graph $G = (V, E)$ is 3-colorable and a partition of V into independent sets I_1, I_2, I_3 with $|I_1| = |I_2| = |I_3|$ is given. This can be proved by replacing EXACT COVER BY 3-SETS in the reduction on pages 68–69 of Garey and Johnson [10] by 3-DIMENSIONAL MATCHING.

THEOREM 5.5. *For any fixed $k \geq 3$, problem P-RL($*$) is NP-hard even when the input is restricted to graphs with a given k -coloring.*

Proof. Let us first prove the theorem for $k = 3$. Let $G = (V, E)$ be a graph and let I_1, I_2, I_3 , $|I_1| = |I_2| = |I_3| = q$ be a partition of V into three independent sets. We

construct a graph $F = (V_F, E_F)$ and a prelabeling L' of F such that there is a radio labeling of F with span $\leq K$ extending L' if and only if V can be partitioned into triangles.

The graph F is obtained as follows: V_F is the disjoint union of four independent sets V_1^A, V_1^B, V_2, V_3 of cardinality q . (The vertices of V_1^A, V_1^B correspond to I_1 , the vertices of V_2 to I_2 , and those of V_3 to I_3 . Also for simplicity we denote vertices from V_F and the corresponding vertices from V by the same letters.)

- (1) $E_F = \{\{u, v\} : u \in V_1^A, v \in V_2 \text{ and } \{u, v\} \notin E\} \cup,$
- (2) $\{\{u, v\} : u \in V_1^B, v \in V_3 \text{ and } \{u, v\} \notin E\} \cup,$
- (3) $\{\{u, v\} : u \in V_2, v \in V_3 \text{ and } \{u, v\} \notin E\} \cup,$
- (4) $\{\{u, v\} : u \in V_1^A, v \in V_3\} \cup,$
- (5) $\{\{u, v\} : u \in V_1^B, v \in V_2\}.$

The vertices of V_1^A are prelabeled with integers $4i - 3$ and the vertices of V_1^B are prelabeled with $4i$, where $1 \leq i \leq q$.

Notice that F is 3-colorable because the sets $V_1^A \cup V_1^B, V_2, V_3$ are independent. We claim that there is a radio labeling with span at most $K = 4q$ in F extending the prelabeling if and only if the vertex set V of G can be partitioned into q triangles.

If there is such a radio labeling in F with span K , then every label from $\{1, 2, \dots, K\}$ should be used in this labeling. Then for every i with $1 \leq i \leq q$, the vertices labeled with $4i - 3$ and $4i$ are in V_1^A and V_1^B (because of the prelabeling). Then by (4), the vertices labeled with $4i - 2$ should be in V_2 , and by (5), the vertices labeled with $4i - 1$ should be in V_3 . Then by (1), (2), and (3), the vertices in G corresponding to vertices labeled with $4i - 3, 4i - 2, 4i - 1, 4i$ induce a triangle in G . Hence, V can be partitioned into q triangles.

For the converse, consider a partition into triangles of V in G . For every triangle with vertex set $\{v_1, v_2, v_3\}$, $v_i \in I_i$, $1 \leq i \leq 3$, the corresponding four vertices $v_1^A \in V_1^A$, $v_1^B \in V_1^B$, $v_2 \in V_2$, $v_3 \in V_3$ in F can be labeled with four consecutive labels. The vertices v_1^A, v_1^B are prelabeled with $4j - 3$ and $4j$ for some j ; by (1), vertex v_2 can be labeled with $4j - 2$, and by (2) and (3), vertex v_3 can be labeled with $4j - 1$. Therefore, F has a radio labeling extending the prelabeling with at most $4q = K$ labels. This settles the case for 3-colorable graphs with a given 3-coloring.

To prove the theorem for any fixed $k \geq 3$, we add to the graph F two disjoint cliques of size k and prelabel the vertices of one clique with labels $\{4q + 1, 4q + 3, \dots, 4q + 2k - 1\}$ and of the other clique with $\{4q + 2, 4q + 4, \dots, 4q + 2k\}$. Then the new graph is k -colorable, and using the previous arguments it is clear that this graph has a radio labeling with $4q + 2k$ labels extending the prelabeling if and only if the graph G has a partition into triangles. This completes the proof. $\square$

6. NP-completeness results for cographs. We now turn to the last class of graphs for which RL is known to be polynomially solvable, namely the class of cographs, i.e., graphs without an induced path on four vertices. Using an easy reduction from 3-PARTITION we show that P-RL(*) is NP-hard for cographs.

THEOREM 6.1. *Problem P-RL(*) is NP-hard for cographs.*

Proof. We use a reduction from the problem 3-PARTITION stated below to P-RL(*).

3-PARTITION

Instance: A set A of nonnegative integers $a_1, \dots, a_{3m}$ and a bound B , such that for all i with $1 \leq i \leq 3m$, $(B + 1)/4 < a_i < B/2$ and $\sum_{1 \leq i \leq 3m} a_i = mB$.

Question: Can A be partitioned into m disjoint sets $A_1, A_2, \dots, A_m$ such that $\sum_{a_i \in A_j} a_i = B$ for every j with $1 \leq j \leq m$?

3-PARTITION is NP-complete in the strong sense (Problem SP15 in Garey and Johnson [10]).

Let the set $A = \{a_1, \dots, a_{3m}\}$ and the bound B be an instance of 3-PARTITION. Consider a complete $(3m+1)$ -partite graph G_A with vertex partition $V = V_0 \cup V_1 \cup \dots \cup V_{3m}$, where $|V_0| = m-1$ and $|V_i| = a_i$ for every i with $i = 1, \dots, 3m$. This means there is an edge between a vertex $v \in V_i$ and a vertex $w \in V_j$ if and only if $i \neq j$. Clearly, G_A is a cograph because it obviously does not contain a path on four vertices as an induced subgraph. Prelabel the vertices $v_1, \dots, v_{m-1}$ of V_0 using the labels $L'(v_j) = j(B+5) - 1$, where $j = 1, \dots, m-1$ and let $K = m(B+5) - 3$. We show that the prelabeling L' of G_A can be extended to a radio labeling with span $\leq K$ if and only if A can be 3-partitioned.

If there is a 3-partition $A_1, \dots, A_m$, then each of the sets A_i contains exactly three elements a_{i_1} , a_{i_2} , and a_{i_3} . Consider then for every $i = 1, \dots, m$ the following labeling:

- Vertices of the set V_{i_1} are labeled with $\{(i-1)(B+5) + 1, \dots, (i-1)(B+5) + a_{i_1}\}$;
- vertices of the set V_{i_2} are labeled with $\{(i-1)(B+5) + a_{i_1} + 2, \dots, (i-1)(B+5) + a_{i_1} + a_{i_2} + 1\}$;
- vertices of the set V_{i_3} are labeled with $\{(i-1)(B+5) + a_{i_1} + a_{i_2} + 3, \dots, (i-1)(B+5) + a_{i_1} + a_{i_2} + a_{i_3} + 2\}$.

Since $a_{i_1} + a_{i_2} + a_{i_3} = B$, this produces a radio labeling extending the prelabeling L' with span $\leq K$.

For the converse, suppose that a radio labeling of G_A extending the prelabeling L' with span $\leq K$ exists. For $j = 1, \dots, m$, let $B_j = \{(j-1)(B+5) + 1, \dots, j(B+5) - 3\}$ where $j = 1, \dots, m$. We call B_j a *range*. Vertices in $V_1 \cup \dots \cup V_{3m}$ must get a label in a range B_j , $1 \leq j \leq m$. Note that for each range B_j , $|B_j| = B+2$. Each range must have at least two unused labels: Labels given to vertices in different sets V_i must differ by at least two, and each $|V_i| = a_i < B/2$; thus each range contains either vertices from at least three sets V_i , and hence at least two unused labels, or vertices from at most two sets V_i , and hence at least $B+2 - 2(B/2 - 1/2) = 3$ unused labels. Since $\sum_{i=1}^{3m} |V_i| = \sum_{i=1}^{3m} a_i = mB = \sum_{j=1}^m (|B_j| - 2)$, there are exactly two unused labels in each range B_j . For each i , $1 \leq i \leq 3m$, consider the highest label ℓ_i given to a vertex in V_i . The label $\ell_i + 1$ either does not belong to a range (this is true for exactly m such labels, as there are m ranges) or is an unused label in a range. Thus, every unused label in a range is one larger than the highest label among all vertices in a set V_i , for some i , $1 \leq i \leq 3m$. As a consequence, all labels of a set V_i must be assigned to the same range. Now, let A_j be the set of values a_i such that vertices of V_i have their label in B_j . As B_j contains exactly B used labels, we have for each j , $\sum_{a_i \in A_j} a_i = B$. This completes the proof. $\square$

We do not know whether P-RL(l) is NP-hard for cographs and leave it as one of the open problems in the next section.

7. Open problems. In this paper we initiated the study of two versions of the radio labeling problem in which a prelabeling is assumed. Many questions remain open, a few of which are listed below:

- We leave the complexity of any of the variants of RADIO LABELING (RL, P-RL(l), and P-RL($*$)) for interval graphs as open problems.
- Another open problem concerns the computational complexity of P-RL(l)

for cographs. Recall that NP-hardness of $P\text{-}RL(*)$ restricted to cographs is proved in this paper and that RL is polynomial for cographs.

- Our results imply that $P\text{-}RL(l)$ is polynomial for bipartite graphs. On the other hand, we proved that $P\text{-}RL(*)$ is NP-hard for 3-partite graphs even if a 3-coloring of the graph is given. The complexity of $P\text{-}RL(*)$ for bipartite graphs is open.
- By Theorem 5.3, $P\text{-}RL(l)$ is polynomial for planar graphs and graphs of bounded treewidth. The complexity of $P\text{-}RL(*)$ for these graph classes is open.

Acknowledgment. We are grateful to Dieter Kratsch for fruitful discussions on the topic of this paper.

REFERENCES

- [1] H. L. BODLAENDER, T. KLOKS, R. B. TAN, AND J. VAN LEEUWEN, λ -coloring of graphs, in Proceedings of the 17th Annual Symposium on Theoretical Aspects of Computer Science, Lecture Notes in Comput. Sci. 1770, Springer, Berlin, 2000, pp. 395–406.
- [2] G. J. CHANG AND D. KUO, *The $L(2, 1)$ -labeling problem on graphs*, SIAM J. Discrete Math., 9 (1996), pp. 309–316.
- [3] P. DAMASCHKE, J. S. DEOGUN, D. KRATSCHE, AND G. STEINER, *Finding Hamiltonian paths in cocomparability graphs using the bump number algorithm*, Order, 8 (1992), pp. 383–391.
- [4] L. ENGBRETSSEN, *An explicit lower bound for TSP with distances one and two*, in Proceedings of the 16th Annual Symposium on Theoretical Aspects of Computer Science, Lecture Notes in Comput. Sci. 1563, Springer, Berlin, 1999, pp. 373–382.
- [5] J. FIALA, A. V. FISHKIN, AND F. V. FOMIN, *Off-line and on-line distance constrained labeling of graphs*, in Proceedings of the 9th European Symposium on Algorithms, Lecture Notes in Comput. Sci. 2161, Springer, Berlin, 2001, pp. 464–475.
- [6] J. FIALA, J. KRATOCHVÍL, AND T. KLOKS, *Fixed-parameter complexity of λ -labelings*, Discrete Appl. Math., 113 (2001), pp. 59–72.
- [7] J. FIALA, J. KRATOCHVÍL, AND A. PROSKUROWSKI, *Distance constrained labelings of precolored trees*, in Proceedings of the 7th Italian Conference on Theoretical Computer Science, Lecture Notes in Comput. Sci. 2202, Springer, Berlin, 2001, pp. 285–292.
- [8] D. A. FOTAKIS AND P. G. SPIRAKIS, *A Hamiltonian approach to the assignment of non-reusable frequencies*, in Proceedings of the 18th Conference on Foundations of Software Technology and Theoretical Computer Science, Lecture Notes in Comput. Sci. 1530, Springer, Berlin, 1998, pp. 18–29.
- [9] D. A. FOTAKIS, S. E. NIKOLETSEAS, V. G. PAPADOPOULOU, AND P. G. SPIRAKIS, *Hardness results and efficient approximations for frequency assignment problems and the radio coloring problem*, Bull. Eur. Assoc. Theor. Comput. Sci. EATCS, 75 (2001), pp. 152–180.
- [10] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman, New York, 1979.
- [11] P. GOLOVACH, *Systems of pairs of q -distant representatives, and graph colorings*, Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI), 293 (2002), pp. 5–25 (in Russian).
- [12] J. R. GRIGGS AND R. K. YEH, *Labelling graphs with a condition at distance 2*, SIAM J. Discrete Math., 5 (1992), pp. 586–595.
- [13] M. HABIB, R. H. MÖHRING, AND G. STEINER, *Computing the bump number is easy*, Order, 5 (1988), pp. 107–129.
- [14] W. K. HALE, *Frequency assignment: Theory and applications*, Proc. IEEE, 68 (1980), pp. 1497–1514.
- [15] J. VAN DEN HEUVEL, R. A. LEESE, AND M. A. SHEPHERD, *Graph labeling and radio channel assignment*, J. Graph Theory, 29 (1998), pp. 263–283.
- [16] R. A. LEESE, *Radio spectrum: A raw material for the telecommunications industry*, in Progress in Industrial Mathematics at ECMI 98, Teubner, Stuttgart, 1999, pp. 382–396.
- [17] C. H. PAPADIMITRIOU AND M. YANNAKAKIS, *The travelling salesman problem with distances one and two*, Math. Oper. Res., 18 (1993), pp. 1–11.
- [18] D. SAKAI, *Labeling chordal graphs: Distance two condition*, SIAM J. Discrete Math., 7 (1994), pp. 133–140.
- [19] A. A. SCHÄFFER AND B. B. SIMONS, *Computing the bump number with techniques from two-processor scheduling*, Order, 5 (1988), pp. 131–141.

A MULTIDIMENSIONAL FILTER ALGORITHM FOR NONLINEAR EQUATIONS AND NONLINEAR LEAST-SQUARES*

NICHOLAS I. M. GOULD[†], SVEN LEYFFER[‡], AND PHILIPPE L. TOINT[§]

Abstract. We introduce a new algorithm for the solution of systems of nonlinear equations and nonlinear least-squares problems that attempts to combine the efficiency of filter techniques and the robustness of trust-region methods. The algorithm is shown, under reasonable assumptions, to globally converge to zeros of the system, or to first-order stationary points of the Euclidean norm of its residual. Preliminary numerical experience is presented that shows substantial gains in efficiency over the traditional monotone trust-region approach.

Key words. nonlinear equations, nonlinear least-squares, nonlinear fitting, filter methods, trust-region algorithms

AMS subject classifications. 65K05, 90C30

DOI. 10.1137/S1052623403422637

1. Introduction. We analyze a filter algorithm for solving systems of nonlinear equations. More formally, we consider the problem of solving

$$(1.1) \quad c(x) = 0,$$

where c is a twice continuously differentiable function from $\mathbb{R}^n$ into $\mathbb{R}^m$. We partition the equations of (1.1) into p (not necessarily disjoint) sets $\{c_i(x)\}_{i \in \mathcal{I}_j}$ for $j = 1, \dots, p$, with $\{1, \dots, m\} = \mathcal{I}_1 \cup \mathcal{I}_2 \cup \dots \cup \mathcal{I}_p$, and define

$$(1.2) \quad \theta_j(x) \stackrel{\text{def}}{=} \|c_{\mathcal{I}_j}(x)\| \quad \text{for } j = 1, \dots, p,$$

where $\|\cdot\|$ is the ordinary Euclidean norm and where $c_{\mathcal{I}_j}$ is the vector whose components are the components of c indexed by $\mathcal{I}_j$. The point x is therefore a solution of (1.1) if and only if $\theta_j(x) = 0$ for $j = 1, \dots, p$. The quantity $\theta_j(x)$ may be interpreted as the size of the residual of the j th set of equations at the point x . We will also use the abbreviations $\theta(x) = (\theta_1(x), \dots, \theta_p(x))^T$, $\theta_k \stackrel{\text{def}}{=} \theta(x_k)$, and $\theta_{j,k} \stackrel{\text{def}}{=} \theta_j(x_k)$. In the simplest case, that is, when $p = m$ and $\mathcal{I}_j = \{j\}$, we have that $\theta_{j,k} = |c_j(x_k)|$ and $\|\theta_k\| = \|c(x_k)\|$.

We follow the classical approach for solving (1.1), which consists of minimizing a merit function involving some norm of the residual. For simplicity, we choose to consider

$$(1.3) \quad \min_{x \in \mathbb{R}^n} f(x) = \frac{1}{2} \|\theta(x)\|^2.$$

Note that this is a least-squares formulation of (1.1), which justifies the second part of the paper's title. (The least-squares are weighted with integer weights if the subsets

*Received by the editors February 7, 2003; accepted for publication (in revised form) January 1, 2004; published electronically October 1, 2004.

<http://www.siam.org/journals/siopt/15-1/42263.html>

[†]Computational Science and Engineering Department, Rutherford Appleton Laboratory, Chilton, Oxfordshire, OX11 0QX, England (n.gould@rl.ac.uk).

[‡]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439 (leyffer@mcs.anl.gov).

[§]Department of Mathematics, University of Namur, B-5000 Namur, Belgium (philippe.toint@fundp.ac.be).

$\mathcal{I}_j$ are not disjoint.) Our objective is therefore to find a (local) minimizer x_* of $f(x)$. If $f(x_*) = 0$, x_* is also a solution of (1.1).

The class of algorithms that we discuss for achieving this objective belongs to the class of trust-region methods and also to that of *filter methods* introduced by Fletcher and Leyffer in 1997 and subsequently published as [8]. Although originally intended for the solution of constrained optimization problems, we claim here that the main idea of the approach, namely, that of a filter to decide on acceptability of the successive iterates, may be extended to the context of (1.1). The question is of interest, since most of the recent contributions to the field of filter algorithms (see, for instance, Fletcher and Leyffer [8], Ulbrich, Ulbrich, and Vicente [24], Chin and Fletcher [2], Fletcher, Leyffer, and Toint [10], Gould and Toint [17], Fletcher et al. [7], and Wächter and Biegler [25]) rely on an external “restoration procedure” whose purpose is to reduce constraint infeasibilities (i.e., when only equality constraints are present, solve a problem of the type (1.1), albeit possibly approximately). The proposal by Gonzaga, Karas, and Vanti [11] goes even further as it uses such a computation as part of the step calculation at every iteration. The fact that (1.1) can itself be handled by a filter algorithm, which, as will be seen below, does not explicitly depend on any external restoration procedure, therefore significantly extends the range of applicability of the filter idea.

The question of using filter techniques for the solution of nonlinear equations has already been considered by Fletcher and Leyffer [9], but their approach is very different from the one proposed here. It relies on a quadratic programming solver to partition the constraints into a set of violated constraints and a set of satisfied or active constraints. These sets are refined from iteration to iteration. It is not clear how this approach can be extended to large-scale systems of equations where only approximate subproblem solves are available and this limits the applicability of [9] to problems that allow active-set-type solvers to be used. By contrast, our approach takes all violations into account together and does not require the revision of any subset during the algorithm. This makes it applicable to very large systems of equations that require inexact subproblem solves to be efficient.

We introduce our algorithm for problem (1.1) in section 2 and discuss its convergence in section 3. Numerical experience is reported in section 4, and some preliminary conclusions are drawn in section 5.

2. Nonlinear equations, filter, and trust regions. An efficient technique for solving (1.1) is to use the Newton or the Gauss–Newton method (or one of their many variants like secant methods), which, from a current iterate x_k (and possibly from information gathered at iterations preceding iteration k), computes a trial step s_k and thus yields a *trial point*

$$x_k^+ = x_k + s_k.$$

Unfortunately, such algorithms may not be convergent from an arbitrary starting point x_0 , and we may have to include it within a framework that guarantees this very desirable property. Our proposal is to use a suitable combination of the trust region (see Conn, Gould, and Toint [4]) and filter techniques.

The main idea of filter algorithms for constrained optimization is that new iterates of the underlying iterative algorithm can be accepted if they do not perform, compared to past iterates kept in the filter, worse on both important and typically conflicting accounts for this type of problem, namely, feasibility and low objective function value. In the context of nonlinear equations, we no longer have to consider an objective

function, but we still face conflicting purposes. Indeed, we may consider driving each of the $\{c_i(x)\}_{i=1}^m$ (or each of the $\{\theta_i(x)\}_{i=1}^p$) to zero as an independent task, which is typically conflicting with that of driving the other components of c (or θ) to zero. Thus we will consider a multidimensional filter, instead of a two-dimensional one.

2.1. The multidimensional filter. To define our filter, we first say that a point x_1 *dominates* a point x_2 whenever

$$\theta_j(x_1) \leq \theta_j(x_2) \text{ for all } j = 1, \dots, p.$$

Thus, if iterate x_{k_1} dominates iterate x_{k_2} , the latter is of no real interest to us since x_{k_1} is at least as good as x_{k_2} for each of the equation sets. All we need to do now is to remember iterates that are not dominated by other iterates using a structure called a filter. A *filter* is a list $\mathcal{F}$ of p -tuples of the form $(\theta_{1,k}, \dots, \theta_{p,k})$ such that

$$\theta_{j,k} < \theta_{j,\ell} \text{ for at least one } j \in \{1, \dots, p\}$$

for $k \neq \ell$. Filter methods propose to accept a new trial iterate x_k^+ if it is not dominated by any other iterate in the filter and x_k . In the vocabulary of multicriteria optimization, this amounts to building elements of the efficient frontier associated with the p -criteria problem of reducing infeasibility on each of the p sets of equations.

While the idea of not accepting dominated trial points is simple and elegant, it needs to be refined a little to provide an efficient algorithmic tool. In particular, we do not wish to accept a new point x_k^+ if $\theta_k^+ \stackrel{\text{def}}{=} \theta(x_k^+)$ is arbitrarily close to being dominated by another point already in the filter. To avoid this situation, we slightly strengthen our acceptability condition. More formally, we say that a new trial point x_k^+ is *acceptable for the filter* $\mathcal{F}$ if and only if

$$(2.1) \quad \forall \theta_\ell \in \mathcal{F} \quad \exists j \in \{1, \dots, p\} \quad \theta_j(x_k^+) < \theta_{j,\ell} - \gamma_\theta \delta(\|\theta_\ell\|, \|\theta_k^+\|),$$

where $\gamma_\theta \in (0, 1/\sqrt{p})$ is a small positive constant and where $\delta(\cdot, \cdot)$ is one of the following:

$$(2.2) \quad \delta(\|\theta_\ell\|, \|\theta_k^+\|) = \|\theta_\ell\|,$$

$$(2.3) \quad \delta(\|\theta_\ell\|, \|\theta_k^+\|) = \|\theta_k^+\|, \text{ or}$$

$$(2.4) \quad \delta(\|\theta_\ell\|, \|\theta_k^+\|) = \min(\|\theta_\ell\|, \|\theta_k^+\|).$$

The first of these formulas measures the margin as a function of the constraints's violation corresponding to the existing filter point; the second measures it as a function of the constraints's violation at the trial point. The third formula attempts to weaken these conditions even further by considering only the smallest of the two violations. The upper bound of $1/\sqrt{p}$ on γ_θ ensures that the right-hand side of (2.1) is always positive for some j for the choices (2.2) and (2.4) and thus that points acceptable for the filter always exist in these cases. Note that such points must exist if (2.3) is considered provided $\|\theta_\ell\| > 0$, but a small value for γ_θ clearly makes it more likely that (2.1) holds for a given θ_k^+ .

2.1.1. Adding a new point to the filter. To avoid cycling, and assuming the trial point is acceptable in the sense of (2.1), we may wish to add it to the filter, so as to avoid other iterates that are worse. The procedure is extremely simple: we simply perform the operation $\mathcal{F} \leftarrow \mathcal{F} \cup \{\theta_k\}$. But this may cause an existing filter entry θ_ℓ to be *strongly dominated* in the sense that every new entry θ_k^+ that is acceptable for the other filter points must also be acceptable for θ_ℓ . We therefore remove θ_ℓ from the filter if

$$\exists \theta_q \in \mathcal{F} \quad \forall j \in \{1, \dots, p\} \quad \theta_{j,\ell} - \gamma_\theta \|\theta_\ell\| \geq \theta_{j,q} - \gamma_\theta \|\theta_q\|$$

if we have chosen to use (2.2),

$$(2.5) \quad \exists \theta_q \in \mathcal{F} \quad \forall j \in \{1, \dots, p\} \quad \theta_{j,\ell} \geq \theta_{j,q}$$

if we have chosen to use (2.3), or

$$\exists \theta_q \in \mathcal{F} \quad \forall j \in \{1, \dots, p\} \quad \theta_{j,\ell} - \gamma_\theta \|\theta_\ell\| \geq \theta_{j,q}$$

if we have chosen to use (2.4). That this last formula is adequate results from

$$\theta_{j,k}^+ < \theta_{j,q} - \gamma_\theta \min[\|\theta_k^+\|, \|\theta_q\|] \leq \theta_{j,q} \leq \theta_{j,\ell} - \gamma_\theta \|\theta_\ell\| \leq \theta_{j,\ell} - \gamma_\theta \min[\|\theta_k^+\|, \|\theta_q\|],$$

which shows that if θ_k^+ is acceptable for θ_q , then it must be acceptable for θ_ℓ .

2.1.2. Computing a trial point. We also need to indicate how to compute the trial point $x_k^+ = x_k + s_k$ for some step s_k . This assumes we have a model $m_k(x)$ of $f(x)$ and a *trust region*

$$(2.6) \quad \mathcal{B}_k = \{x_k + s \mid \|s\|_k \leq \Delta_k\},$$

where we believe this model to be reasonably accurate. The norms $\|\cdot\|_k$ are allowed to vary from iteration to iteration provided they remain uniformly equivalent to the Euclidean norm. The convergence analysis that follows requires, as is common in trust-region methods, that this step provides, at iteration k , a *sufficient decrease* on the model: we require that

$$(2.7) \quad m_k(x_k) - m_k(x_k + s_k) \geq \kappa_{\text{mdc}} \|g_k\| \min \left[\frac{\|g_k\|}{\beta_k}, \Delta_k \right],$$

where $g_k = \nabla m_k(x_k)$, β_k is a positive upper bound on the norm of the Hessian of m_k , and κ_{mdc} is a constant in $(0, 1)$. This is a natural requirement in the context of trust-region algorithms (see [4, p. 131], for instance), and there are efficient algorithms for achieving (2.7) in the trust-region framework for both small- and large-scale cases. This condition is also satisfied if the model is further reduced outside the trust region, as is the case in our algorithm when $\|s_k\| > \Delta_k$.

At variance with classical trust-region methods, we do not require that

$$(2.8) \quad \|s_k\|_k \leq \Delta_k$$

at every iteration to obtain our most important convergence result. In particular, this means that whenever (2.8) is not enforced and $m = n$, we might use the step $s_k = -J_k^{-1}c_k$, where J_k is the Jacobian of $c(x)$ at x_k (assumed here to be full rank).

This step leads to the root of the first-order model for $c(x)$, which is also the global minimizer of

$$(2.9) \quad m_k^{\text{GN}}(x_k + s) = \frac{1}{2} \sum_{i=1}^p \|c_{\mathcal{I}_i}(x_k) + J_{\mathcal{I}_i}(x_k)s\|^2,$$

a Gauss–Newton model for $f(x)$, where $J_{\mathcal{I}_i}$ is the Jacobian of $c_{\mathcal{I}_i}$. This model may also be minimized if $m > n$ and (2.8) is not enforced, or one could choose in this case to minimize the second-order truncation of the Taylor series for $f(x)$,

$$(2.10) \quad m_k^{\text{N}}(x_k + s) = m_k^{\text{GN}}(x_k + s) + \frac{1}{2} \sum_{i=1}^p \sum_{j \in \mathcal{I}_i} c_j(x_k) \langle s, \nabla^2 c_j(x_k)s \rangle,$$

which corresponds to a full Newton model, whenever this last model is convex.

ALGORITHM 2.1. FILTER-TRUST-REGION ALGORITHM.

Step 0: Initialization.

An initial point x_0 and an initial trust-region radius $\Delta_0 > 0$ are given, as well as constants $0 < \gamma_0 \leq \gamma_1 < 1 \leq \gamma_2$, $\gamma_\theta \in (0, 1/\sqrt{p})$, $0 < \eta_1 < \eta_2 < 1$. Assume that the sets $\{\mathcal{I}_j\}_{j=1}^p$ are also given. Compute $c_0 = c(x_0)$ and θ_0 . Set $k = 0$, define a flag **RESTRICT** to be unset and initialize the filter to the empty set.

Step 1: Test for optimality.

If $\theta_k = 0$ or $\|\nabla f(x_k)\| = 0$, stop.

Step 2: Determine a trial step.

Compute a step s_k that satisfies (2.7) and that also satisfies (2.8) if **RESTRICT** is set. Compute the trial point $x_k^+ = x_k + s_k$.

Step 3: Evaluate the residual at the trial step.

Compute $c(x_k^+)$ and $\theta_k^+ = \theta(x_k^+)$. Define

$$\rho_k = \frac{f(x_k) - f(x_k^+)}{m_k(x_k) - m_k(x_k^+)}.$$

Step 4: Test to accept the trial step.

- If x_k^+ is acceptable for the current filter:
set $x_{k+1} = x_k^+$, unset **RESTRICT**, and add θ_k^+ to the filter if either $\rho_k < \eta_1$ or (2.8) fails.
- If x_k^+ is not acceptable for the current filter:
If (2.8) holds and $\rho_k \geq \eta_1$, set $x_{k+1} = x_k^+$ and unset **RESTRICT**. Else, set $x_{k+1} = x_k$ and set **RESTRICT**.

Step 5: Update the trust-region radius.

If $\|s_k\|_k \leq \Delta_k$, update the trust-region radius by choosing

$$\Delta_{k+1} \in \begin{cases} [\gamma_0 \Delta_k, \gamma_1 \Delta_k] & \text{if } \rho_k < \eta_1, \\ [\gamma_1 \Delta_k, \Delta_k] & \text{if } \rho_k \in [\eta_1, \eta_2] \\ [\Delta_k, \gamma_2 \Delta_k] & \text{if } \rho_k \geq \eta_2; \end{cases}$$

otherwise, set $\Delta_{k+1} = \Delta_k$. Increment k by one and go to Step 1.

2.2. The algorithm. We now combine these ideas into an algorithm. The main objective of Algorithm 2.1 is to let the filter play the major role in ensuring global convergence and to fall back on the usual ℓ_2 -norm reduction algorithm if things do not go well, or if convergence occurs to a local minimizer of f which is not a zero of c .

This algorithm provides only a canvas for a practical implementation, as we have not fully specified each step. In particular, we have not detailed how the step s_k can be computed in practice to satisfy the sufficient decrease condition (2.7) and remain in the trust region when required, but the literature provides several choices of direct or iterative techniques; see Moré and Sorensen [20], Gould et al. [12], or Chapter 7 of [4], for example. Similarly, techniques to initialize and update the trust-region radius can be found in Chapter 17 of this last reference. As is usual, we say that iteration k is successful if $x_{k+1} = x_k + s_k$.

3. Global convergence. We now investigate the convergence properties of Algorithm 2.1, under the following set of assumptions:

- A1: $c(x)$ is twice continuously differentiable on $\mathbb{R}^n$.
- A2: The iterates x_k remain in a bounded domain of $\mathbb{R}^n$.
- A3: $m_k(x)$ is twice differentiable on $\mathbb{R}^n$ for all k .
- A4: For all k , $m_k(x_k) = f(x_k)$ and $g_k = \nabla m_k(x_k) = \nabla f(x_k)$.

For the purpose of the convergence analysis, we also assume that the algorithm does not terminate in Step 1.

Note that A1, A2, and A3 together imply that there is a constant $\kappa_u > 0$ such that

$$(3.1) \quad \|c(x)\| \leq \kappa_u, \quad \|\nabla^2 c_i(x)\| \leq \kappa_u, \quad \text{and} \quad \|\nabla^2 m_k(x)\| \leq \kappa_u$$

for all k and all x in the convex hull of $\{x_k\}$. The second of these inequalities then ensures that κ_u can also be chosen such that

$$\|\nabla^2 f(x)\| \leq \kappa_u.$$

We could have assumed the three conditions (3.1) independently instead of imposing A2, but we have chosen not to do so for the sake of simplicity. Assumptions A1, A3, A4, and (3.1) are typical of convergence theory for trust-region methods (see Chapters 6 and 16 of [4]).

We first investigate what happens if infinitely many values are added to the filter in the course of the algorithm.

LEMMA 3.1. *Suppose that A1 and A2 hold and that infinitely many values of θ_k are added to the filter by the algorithm. Then*

$$(3.2) \quad \lim_{k \rightarrow \infty} \|c_k\| = \lim_{k \rightarrow \infty} \|\nabla f(x_k)\| = 0.$$

Proof. Let $\{k_i\}$ index the subsequence of iterations at which $\theta_{k_i} = \theta_{k_i-1}^+$ is added to the filter. Now assume, for the purpose of a contradiction, that there exists a subsequence $\{k_j\} \subseteq \{k_i\}$ such that

$$\|\theta_{k_j}\| \geq \epsilon_1$$

for some $\epsilon_1 > 0$. Since our assumptions imply that $\{\|\theta_{k_j}\|\}$ is bounded above and below, there must exist a further subsequence $\{k_\ell\} \subseteq \{k_j\}$ such that

$$(3.3) \quad \lim_{\ell \rightarrow \infty} \theta_{k_\ell} = \theta_\infty \quad \text{with} \quad \|\theta_\infty\| \geq \epsilon_1.$$

Moreover, by definition of $\{k_\ell\}$, θ_{k_ℓ} is acceptable for the filter for every ℓ , which implies in particular that for each ℓ , there exists a $j \in \{1, \dots, p\}$ such that

$$(3.4) \quad \theta_{j,k_\ell} - \theta_{j,k_{\ell-1}} < -\gamma_\theta \delta(\|\theta_{k_{\ell-1}}\|, \|\theta_{k_\ell}\|).$$

Furthermore, this conclusion does not depend on $\theta_{k_{\ell-1}}$ still being in the filter. Indeed, if $\theta_{k_{\ell-1}}$ has been removed from the filter, it must be strongly dominated by another entry in the filter, which then guarantees that (3.4) holds. But (3.3) and our assumptions on δ imply that there exists an $\epsilon_2 > 0$ such that

$$\delta(\|\theta_{k_{\ell-1}}\|, \|\theta_{k_\ell}\|) \geq \epsilon_2$$

for all ℓ sufficiently large. Hence we deduce from (3.4) that

$$\theta_{j,k_\ell} - \theta_{j,k_{\ell-1}} < -\gamma_\theta \epsilon_2$$

for ℓ sufficiently large. But the left-hand side of this inequality tends to zero when ℓ tends to infinity because of (3.3), yielding the desired contradiction. Hence

$$(3.5) \quad \lim_{i \rightarrow \infty} \|\theta_{k_i}\| = 0.$$

Consider now any $\ell \notin \{k_i\}$ and let $k_{i(\ell)}$ be the last iteration before ℓ such that $\theta_{k_{i(\ell)}}$ was added to the filter. Since, by construction, every successful iteration where the value of θ at the trial point is not included in the filter must result in a decrease of the objective function (since $\rho_k \geq \eta_1$ on such iterations), we deduce that, for all $\ell \notin \{k_i\}$,

$$f(x_\ell) \leq f(x_{k_{i(\ell)}}),$$

and therefore that

$$\|\theta(x_\ell)\| \leq \|\theta(x_{k_{i(\ell)}})\|.$$

Combining this inequality with (3.5) and the fact that $k_{i(\ell)} \in \{k_i\}$, we obtain that

$$\lim_{k \rightarrow \infty} \|\theta_k\| = 0,$$

which implies (3.2) and completes the proof. $\square$

Let us now consider the case where only finitely many values are added to the filter in the course of the algorithm, and let us examine first the case where the number of trial points acceptable for the filter is also finite. Hence, the main test of Step 4 fails for all $k \geq k_0$, say, and the method is then identical to an ordinary trust-region method for minimizing $f(x)$, since iterations for which (2.8) fails have no effect on the current iterate nor on the trust region radius and may be ignored in the convergence theory. Our assumptions are sufficient to use the results of Chapter 6 in [4] and we may therefore deduce the following two properties.

LEMMA 3.2. *Suppose that A1–A4 hold and that only finitely many trial points x_k^+ are acceptable for the filter. Suppose furthermore that there are only finitely many successful iterations. Then the algorithm terminates in Step 1 with $x_k = x_*$, and $\nabla f(x_*) = 0$.*

LEMMA 3.3. *Suppose that A1–A4 hold, that only finitely many trial points x_k^+ are acceptable for the filter, and that there are infinitely many successful iterations. Then*

$$\liminf_{k \rightarrow \infty} \|\nabla f(x_k)\| = 0.$$

We have thus covered all cases, except that where only finitely many values are added to the filter but infinitely many trial points are acceptable for the filter. In this situation, we therefore have for sufficiently large k three kinds of intertwined iterations:

- The first are iterations where x_k^+ is not acceptable for the filter but (2.8) fails. These iterations have no effect and may be ignored, as we argued before Lemma 3.2.
- The second are iterations where x_k^+ is not acceptable for the filter but (2.8) holds. These are perfectly standard trust-region iterations.
- In the third kind, x_k^+ is acceptable for the filter, but θ_k^+ is not added to it because $\rho_k \geq \eta_1$ and (2.8) holds. These are again standard (successful) trust-region iterations.

The sequence of iterations can therefore be viewed as resulting from a standard trust-region algorithm, and known convergence results are again applicable. As a consequence, the conclusions of Lemmas 3.2 and 3.3 are also valid in this context.

We summarize our results in the next theorem.

THEOREM 3.4. *Suppose that A1–A4 hold. Then there exists a subsequence $\{k_i\}$ such that $\liminf_{i \rightarrow \infty} \|\nabla f(x_{k_i})\| = 0$. Moreover, if infinitely many values are added to the filter, then $\lim_{k \rightarrow \infty} \|c(x_k)\| = 0$.*

Note that this theorem does not guarantee that a subsequence of $\{\|c_k\|\}$ converges to zero in all cases, but only that a stationary point of $f(x)$ is reached. This is expected because it may happen that the equations (1.1) have no solution or that the iterates get trapped near a local minimum of $f(x)$ with positive value. Such an outcome might be troublesome if one aims at finding a root of (1.1), as opposed to a locally minimum residual, but avoiding this situation requires global optimization techniques, which are outside the scope of this paper.

The convergence theory of trust-region algorithms also implies (see [4, Theorem 6.4.6]) that the limit inferior in Lemma 3.3 can be replaced by a true limit provided

$$(3.6) \quad \|s_k\|_k \leq \kappa_\Delta \Delta_k \quad \text{for all } k \geq k_0,$$

for some $k_0 > 0$ and some constant $\kappa_\Delta \geq 1$. In the usual trust-region context, this is automatically guaranteed since the trial point always lies within the trust region itself. However, this property no longer holds for the algorithm described in this paper, because unrestricted steps are possible. But we may still obtain the stronger result if we are ready to assume that (3.6) holds for some κ_Δ possibly larger than one, which might be seen as a reasonable implementation safeguard. The rest of the discussion above then remains unchanged and we deduce our final convergence result.

THEOREM 3.5. *Suppose that A1–A4 and (3.6) hold. Then either $\|\nabla f(x_k)\| = 0$ for some finite k or $\lim_{k \rightarrow \infty} \|\nabla f(x_k)\| = 0$. Moreover, if infinitely many values are added to the filter, then we have that $\lim_{k \rightarrow \infty} \|c(x_k)\| = 0$.*

4. Numerical experience.

4.1. Test problems and performance measures. We now look at the practical behavior of the new method when applied to test problems of the CUTEr collection¹ (see Bongartz et al. [1] and Gould, Orban, and Toint [13]). Table 4.1 indicates the

¹We have excluded the problems FLOSP2HH, FLOSP2HL, FLOSP2HM, FLOSP2TL, FLOSP2TM, and SEMICON1 because their solution appears to require more specific preconditioning (they all require more than 1 hour of CPU time for every option we tried) and problems CHEMRCTA and CHEMRCTB because they seemed to generate numerical overflow with all the tested methods.

TABLE 4.1
Test problem characteristics.

Problem	n	n free	m	$\ \theta_*\ $	Problem	n	n free	m	$\ \theta_*\ $
ARGLALE	400	400	800	+	DRCVITY1	4489	3969	3969	0
ARGLBLE	400	400	800	+	DRCVITY2	4489	3969	3969	0
ARGLCLE	400	400	799	+	DRCVITY3	4489	3969	3969	0
ARGTRIG	200	200	200	0	EIGENAU	2550	2550	2550	0
ARTIF	100002	100000	100000	0	EIGENB	2550	2250	2250	0
ARWHDNE	500	500	998	+	EIGENC	2652	2652	2652	0
BDVALUES	10002	10000	10000	0	INTEGREQ	1002	1000	1000	0
BRATU2D	123904	122500	122500	0	MSQRTA	4900	4900	4900	0
BRATU2DT	23104	22500	22500	0	MSQRTB	4900	4900	4900	0
BRATU3D	4913	3375	3375	0	POROUS1	5184	4900	4900	0
BROWNALE	1000	1000	1000	0	POROUS2	5184	4900	4900	0
BROYDN3D	100000	100000	100000	0	SEMICON2U	5002	5000	5000	0
BROTDNBD	100000	100000	100000	0	SPMSQRT	27001	27001	44999	0
CBRATU2D	7200	6728	6728	0	WOODSNE	80000	80000	60001	+
CBRATU3D	16000	11664	11664	0	YATP1SQ	123200	123200	123200	0
CHANDHEU	500	500	500	0	YATP2SQ	123200	123200	123200	0
CHANNEL	19200	19198	19198	0					

names and dimensions of the test problems considered. The columns $\|\theta_*\|$ in this table indicate whether the residual at the solution is zero (0) or positive (+). As is apparent from this table, the test set includes a fair mix of problems sizes and over- or underdetermined cases. All experiments reported in this section were run with the current version of the Fortran 95 *FILTRANE* package (see Gould and Toint [16]) on a Dell Latitude C840 portable computer (1.6 MHz, 1 Gbyte RAM) under the Fujitsu frt Fortran compiler with default optimization. The values $\gamma_0 = 0.625$, $\gamma_1 = 0.25$, $\gamma_2 = 2$, $\eta_1 = 0.01$, $\eta_2 = 0.9$, and

$$\gamma_\theta = \min \left[0.001, \frac{1}{2\sqrt{p}} \right]$$

were used.

In what follows, we compare several variants of the algorithm discussed above for reliability and efficiency. All variants discussed use the Euclidean norm to define the trust region (in (2.6)) and

$$\delta(\|\theta_\ell\|, \|\theta_k^+\|) = \|\theta_k^+\|.$$

This latter choice was made because it simplifies the removal of strongly dominated filter entries (see (2.5)).

Efficiency comparisons are made using the performance profiles introduced by Dolan and Moré [6]. Suppose that a given algorithmic variant i from a set $\mathcal{A}$ reports a statistic $s_{ij} \geq 0$ when run on example j from our test set $\mathcal{T}$, and suppose that the smaller this statistic, the better the variant is considered. Let

$$k(s, s^*, \sigma) = \begin{cases} 1 & \text{if } s \leq \sigma s^*, \\ 0 & \text{otherwise.} \end{cases}$$

Then, the *performance profile* of algorithm i is the function

$$p_i(\sigma) = \frac{\sum_{j \in \mathcal{T}} k(s_{ij}, s_j^*, \sigma)}{|\mathcal{T}|} \quad (\sigma \geq 1),$$

where $s_j^* = \min_{i \in \mathcal{A}} s_{ij}$. (s_{ij} is defined as infinity if variant i failed on problem j , which then implies that $k = 0$.) Thus $p_i(1)$ gives the fraction of the number of examples for which algorithm i was the most effective (according to statistics s_{ij}), $p_i(2)$ gives the fraction of the number for which algorithm i is within a factor of 2 of the best, and $\lim_{\sigma \rightarrow \infty} p_i(\sigma)$ gives the fraction of the examples for which the algorithm succeeded. We consider such a profile to be a very effective means of comparing the relative merits of our algorithmic variants.

We have chosen to compare unpreconditioned versions of the new methods, NITSOL and LANCELOT-B (see section 4.5), as no preconditioning was reasonably successful on our test problems and also because the CUTer interface to NITSOL only allows for its use without a preconditioner. It must be noted, however, that all tested methods provide facilities for preconditioning, which would typically be required to solve more difficult problems. We ran additional tests with our new technique using diagonal and band preconditioners, which confirm the findings reported below for the unpreconditioned case.

4.2. Filter versus trust region and pure Newton. We start by examining the most obvious question, namely, that of the potential added value of the filter technique compared to the more traditional trust-region approach. We therefore consider the following algorithmic variants.

GNFTR. This variant is Algorithm 2.1, where, at each iteration, the trial point is computed by approximately minimizing $m^{\text{GN}}(x_k + s)$ using the generalized Lanczos trust-region algorithm of Gould et al. [12] as implemented in the GALAHAD library [14]. This procedure is terminated at the first s for which

$$(4.1) \quad \|\nabla m^{\text{GN}}(x_k + s)\| \leq \min \left[0.1, \sqrt{\max(\epsilon_M, \|\nabla m^{\text{GN}}(x_k)\|)} \right] \|\nabla m^{\text{GN}}(x_k)\|,$$

where ϵ_M is the machine precision. The choice $\theta(x) = c(x)$ is also made, which implies that an m -dimensional filter is used. The bound (3.6) is imposed with $\kappa_\Delta = 1000$ at all iterations following the first one at which a restricted step was taken. In addition, the condition

$$f(x_k + s_k) \leq \min(10^6 f(x_0), f(x_0) + 1000)$$

is imposed for the trial point to be acceptable for the filter. Instead of the formal requirement of Step 1, the algorithm stops if

$$(4.2) \quad \|\nabla f(x_k)\| \leq 10^{-6} \sqrt{n} \quad \text{or} \quad \|c(x_k)\|_\infty \leq 10^{-6},$$

or if the number of iterations exceeds 1000, or if the computing time exceeds 1 hour. Other details of the implementation are discussed in Gould and Toint [16] and do not matter in our comparison.

GNTR. This variant is identical to GNFTR except that all trial steps are considered as unacceptable for the filter. The resulting method is therefore nothing but a basic trust-region algorithm (Algorithm BTR in [4]).

GNDD. This is a dare-devil variant identical to GNFTR except that every trial step is considered to be acceptable for the filter. Note that this implies that the filter entries need not be stored, since comparisons are no longer performed. Note also that this method does not include any mechanism for guaranteeing global convergence.

All three variants are reliable, even if they all reported failure (while still producing very good approximations to the solution) on problems ARGBLE and ARGCLE, two

very ill-conditioned linear least-squares problems. GNFTR required more than 1000 iterations for **DRCAVTY3**. GNDD was stopped after 1 hour of CPU time on **SEMICN2U** and required more than 1000 iterations on **ARWHDNE**. It furthermore failed on **YATP2SQ** because it wandered off into a region where the Hessian of the Gauss–Newton model is so ill-conditioned that negative curvature was erroneously reported in the course of the conjugate-gradients iteration, causing divergence. The good reliability of GNDD is surprising, given that Newton’s method is known to fail on some problems if not augmented by a globalization mechanism. We therefore investigated if this behavior could be caused by the starting points for the test problems being too close to the problem’s solution. To test this hypothesis, we ran all test problems again, perturbing the starting point using the transformation $x_0 \leftarrow 10x_0 + 1$, but this gave essentially the same results (except that fewer problems were suitable for comparison as local minimizers of the constraints violation were found more often). Another potential explanation of this phenomenon is simply that the CUTEr test problem collection contains too few examples of large-scale nonlinear systems where an algorithm like GNDD diverges,² although these test problems are derived from a number of different application areas including nonlinear PDEs, nonlinear ODEs, nonlinear matrix equations, and general algebraic nonlinear systems. Unfortunately, we have no real means of assessing the generality of our problem collection.³ A last possible explanation of GNDD’s remarkable reliability is that GNDD does in fact converge on a large class on nonlinear systems but that our understanding is currently too limited to explain this behavior.

The relative efficiency of these methods is illustrated by the performance profile in Figures 4.1 and 4.2. The overall performance of the filter variants is impressive. These profiles show that GNFTR is the clear winner in CPU time: the GNFTR filter variant appears to be within a factor of two of the best for approximately 95% of the test problems (it is fastest on 74%). The dominance of the GNFTR variant over the GNTR variants is even stronger if one considers the number of iterations (which, in our setting, is equal to the number of constraints evaluations and approximately⁴ equal to the number of Jacobian evaluations). The efficiency gains obtained by augmenting the trust-region algorithm with a filter-based acceptance rule for trial points thus appear to be very significant. It is also interesting to note that GNFTR is very comparable to GNDD for the problems where the latter does not fail, indicating that the filter technique may have the efficiency of a “blind” Newton’s method while preserving guaranteed convergence (in the sense of Theorem 3.4). Finally, the dominance of GNDD over GNTR for small values of σ indicates that the trust-region mechanism might impose too heavy a burden on Newton’s method to ensure its global convergence.

Given that each filter entry is an m -dimensional vector, storing many filter entries may impose large memory requirements. Fortunately, large filters appear to be exceptional, as can be seen in Figure 4.3, where we show the relation between m and the maximal filter size (over all iterations and all test examples involving m equations). With the exception of problem **SEMICN2U**, where 293 filter entries were required for $m = 5000$, all runs required less than 45, irrespective of the value of m . The memory requirements thus remain typically modest.

²GNDD also fails on a few small-scale problems from the CUTEr collection, like **HEART8**, **PFIT1**, **PFIT3**, **PFIT4**, and **RSNBRNE**.

³However, we are always pleased to include further problems that are brought to our attention.

⁴Equality does not hold because the Jacobian is not evaluated at unsuccessful iterations.

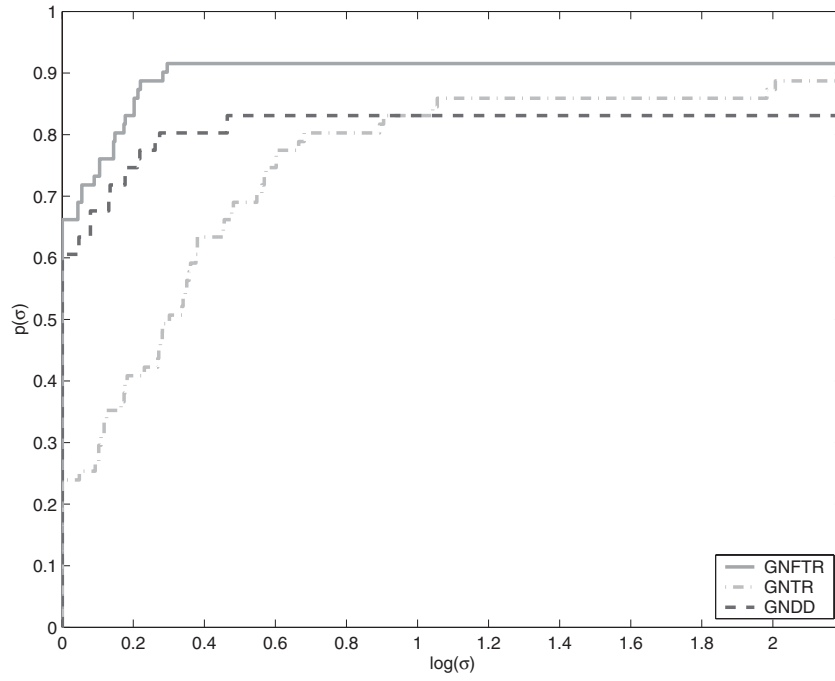


FIG. 4.1. CPU time performance profile for GNFT, GNTR, and GNDD.

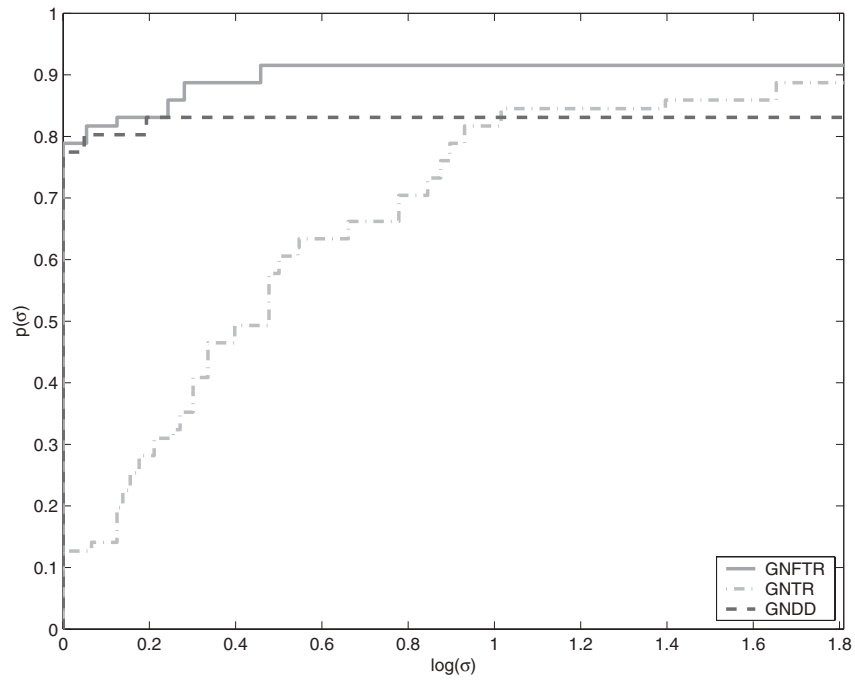


FIG. 4.2. Iteration number performance profile for GNFT, GNTR, and GNDD.

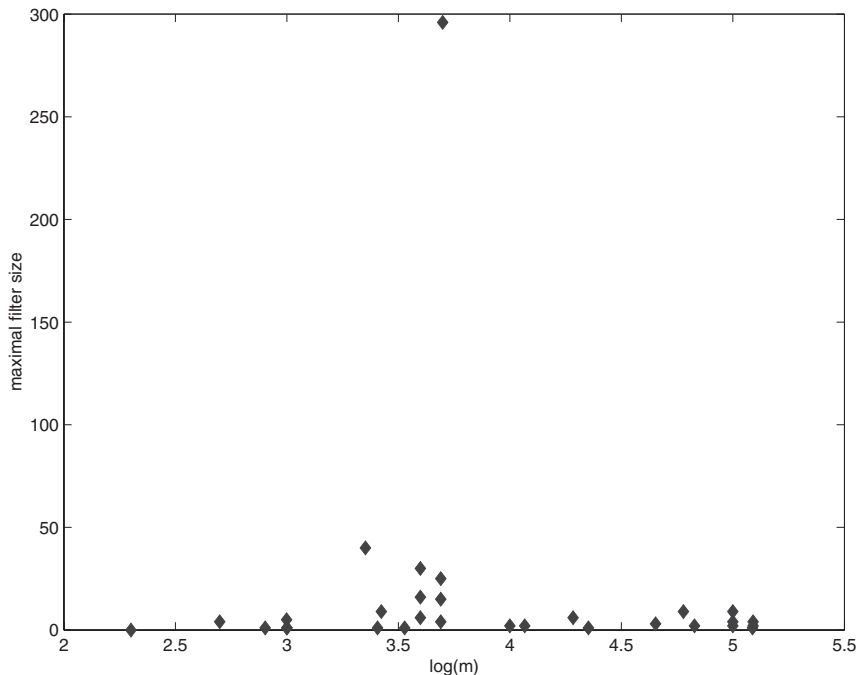


FIG. 4.3. The maximal number of entries in the filter as a function of $\log(m)$.

4.3. Model choice and subproblem accuracy. We next consider the choice of the Gauss–Newton model (2.9) and compare it to the full Newton model given by (2.10), in each case using default (see (4.1)) and full machine accuracy for the solution of the linear subproblem. An important feature of (2.10) is that, in contrast with (2.9), it need no longer be convex. While this is of no consequence in the pure trust-region context because the trust-region boundary prevents arbitrary long steps toward unbounded minima of (2.10), this causes a problem for Algorithm 2.1. Indeed, negative curvature of the model may be discovered when unrestricted steps are computed. This might result in unnecessarily long steps, which have then to be cut back, generating inefficient oscillations in the step length. We have thus added to Algorithm 2.1 the feature that unrestricted steps for which negative curvature is detected are automatically recomputed with **RESTRICT** set without even attempting to compute the constraints values at the trial point. This can be done at minimal cost by using the reentry facility in the GLTR package, which exploits previously computed Krylov spaces (see Gould et al. [12] for details). We thus introduce three new variants of Algorithm 2.1:

- NFTR is identical to GNFTTR except that the model (2.10) is used instead of (2.9) and unrestricted steps containing negative curvature are restricted before attempting to compute the constraint values at the trial point.
- GNFTTR-full is identical to GNFTTR except that (4.1) is replaced by

$$(4.3) \quad \|\nabla m^{\text{GN}}(x_k + s)\| \leq \sqrt{\epsilon_M} \|\nabla m^{\text{GN}}(x_k)\|.$$

- NFTR-full is identical to NFTR except that (4.1) is replaced by (4.3).

Comparing the reliability of GNFTTR, GNFTTR-full, NFTR, and NFTR-full, we note that, as above, all four variants stall on **ARGLBLE** and **ARGLCLE**. NFTR has not

converged in 1000 iterations on **DRCVITY3**, and this is also the case for **GNFTR**-full on **ARTIF** and **DRCVITY3** and for **NFTR**-full on **DRCVITY2**. The package was terminated after 1 hour of CPU time for **GNFTR**-full on **DRCVITY2**, **MSQRTA**, and **MSQRTB**, for **NFTR**-full on **BDVALUES**, **EIGENC**, **MSQRTA**, **MSQRTB**, **DRCVITY3**, **SEMICN2U**, and **YATP2SQ**. The variants using the full Newton model therefore seem less reliable than the Gauss–Newton ones on our test set. Furthermore, they are less efficient, as is apparent in the performance profiles of Figures 4.4 and 4.5. But these conclusions must be tempered by the observation that many of our test problems are actually such that $\|c(x_*)\| = 0$, therefore implicitly favoring the Gauss–Newton model, as it is well known that the full Newton model is typically more efficient when the solution of the problem occurs at a point x_* for which $\|c(x_*)\|$ is strictly positive.

Occasionally, however, there is considerable benefit from the Newton model. The most extreme example in this direction is that of the **ARWHDNE** problem, which is also the problem for which $\|c(x_*)\|$ is largest. It is detailed in Table 4.2.

TABLE 4.2

Detailed performance of the GNFTR, GNFTR-full, NFTR, and NFTR-full variants on problem ARWHDNE.

	CPU(s)	iter.	CG iter.
GNFTR	0.2	178	240
GNFTR-full	0.2	179	273
NFTR	0.1	6	11
NFTR-full	0.1	6	12

Adaptive procedures to choose between the Gauss–Newton and the Newton model have been studied (see Dennis, Gay, and Welsch [5], Toint [22], or Lukšan [19], for instance), but their application to our framework is beyond the scope of this paper and is postponed to [16].

4.4. Equation grouping. We finally examine the impact of varying p , the dimension of the filter space, by defining sets of equations as explained in the introduction. We have chosen to compare our initial **GNFTR** variant (in which every equation has its own subset) with three different alternative strategies for grouping equations in k disjoint subsets (groups), with $k = n/10$, 10, and 1. In our experiments, we simply assigned the i th equation to the group

$$s(i) = \begin{cases} \text{mod}(i, k) & \text{if } \text{mod}(i, k) > 0, \\ k & \text{otherwise.} \end{cases}$$

Observe that if all equations are merged into a single group, (2.1) and (2.3) reduce to

$$\|c(x_k^+)\| < \frac{1}{1 + \gamma_\theta} \|c(x_{\ell(k)})\|,$$

where $\ell(k)$ is the last entry added to the filter before or at iteration k . It follows from the design of the algorithm that every successful iteration between $\ell(k)$ and k must produce a successful trust-region step, since otherwise the corresponding trial step would have been added to the filter in Step 4. This method may thus be seen as a particular nonmonotone filter-trust-region strategy with variable memory, in the spirit of the proposals by Ke and Han [18], Xiao and Chu [26], and Toint [23].

Not surprising, all variants fail on the ill-conditioned problems **ARGLBLE** and **ARGLCLE**. The computation was terminated after 1000 iterations on **ARWHDNE** for the

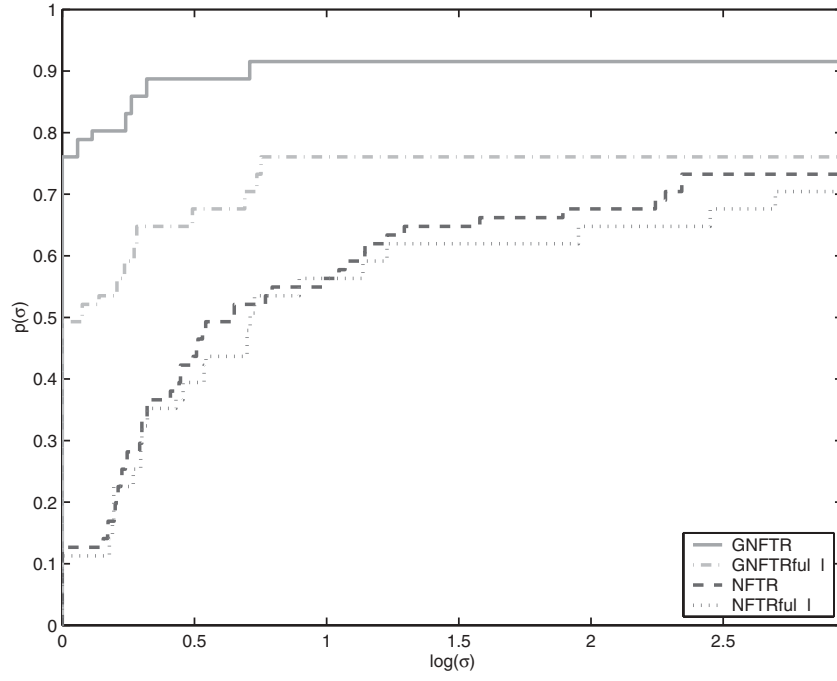


FIG. 4.4. CPU time performance profile for GNFTTR, NFTR, GNFTTR-full, and NFTR-full.

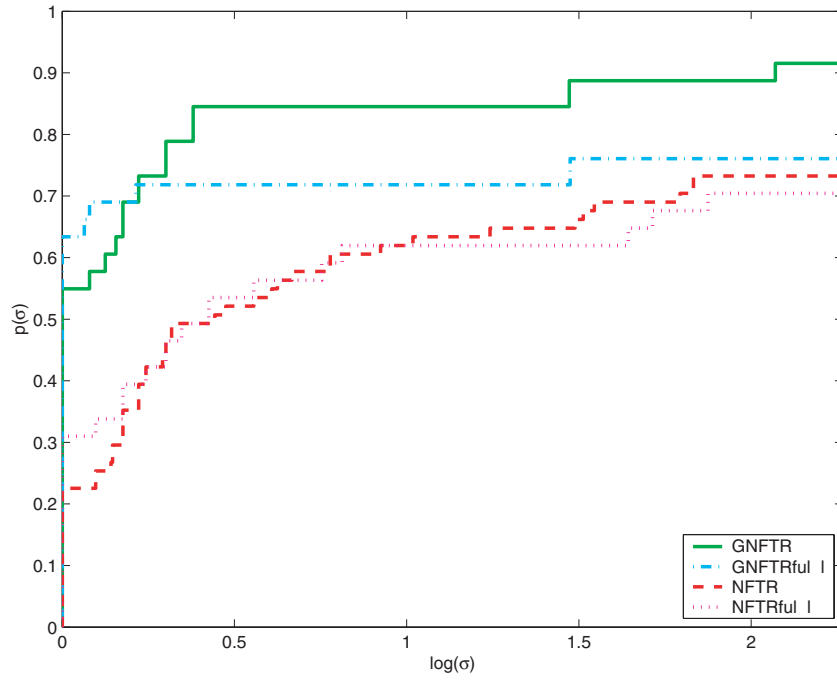


FIG. 4.5. Iterations performance profile for GNFTTR, NFTR, GNFTTR-full, and NFTR-full.

case of 10 groups, while it was terminated after 1 hour of CPU time on **DRCAVITY3** for the cases of 1, 10, and $n/10$ groups.

The performance profiles in Figures 4.6 and 4.7 show that the variants differ, but to a lesser degree than in the comparisons discussed above.

For each of the three criteria, we note that a higher number of groups (a higher value of p) seems to result in better performance. This effect is mostly due to the fact that the set of unacceptable values in a higher dimensional filter space is typically smaller than for a low filter dimension, which then allows more iterates to be accepted.

It is interesting to note, in Figure 4.8, that the size of the filter tends to decrease with the number of groups, mostly because more filter entries are removed from the filter (using (2.5)) when there are fewer groups and thus fewer filter dimensions. Thus the memory requirements for storing the filter entries decrease even faster than simply implied by the reduction in the dimension of each filter entry.

We finally observe that the performance of **GNFTR** with a single group remains excellent, which indicates that a unidimensional filter may provide an attractive alternative to other nonmonotone trust-region methods.

We also attempted to partition the equations in different groups such that the $\theta_i(x_0)$ are balanced in size, but this strategy did not produce results significantly different from those obtained without balancing.

4.5. Comparison with NITSOL and LANCELOT-B. Finally, we attempted to compare **GNFTR** with two existing packages for solving nonlinear equations. We considered **NITSOL** (see Pernice and Walker [21]) and **LANCELOT-B** (see Conn, Gould, and Toint [3] and Gould, Orban, and Toint [14]), a program available in the **GALAHAD** library.

NITSOL uses a truncated **GMRES** iterative method to solve Newton's equations

$$J(x_k)s_k + c_k = 0,$$

followed by a backtracking linesearch to ensure global convergence. **LANCELOT-B** is a nonmonotone trust-region method. (See [23] for a description of the algorithm.) Although principally intended for the solution of nonlinear optimization problems, it can nevertheless be used to solve nonlinear systems of equations. In that case, it uses a truncated conjugate gradient method on the full Newton model (2.10), which makes it very similar in spirit to **NFTR**.

Comparing different software packages is always difficult. To make the exercise as fair as possible, we modified the default stopping criterion of **NITSOL** to reflect the factor $\sqrt{n}$ used in our new code (see (4.2)) and we set the maximum number of iterations to 1000 and the maximum allowed number of **GMRES** iterations to 100,000. We believe that this makes **NITSOL**'s performance very comparable to that of our new method. However, because **NITSOL** requires that problems be square (i.e., $m = n$), we removed problems **ARGLALE**, **ARGLBLE**, **ARGLCLE**, **ARWHDNE**, **SPMSQRT**, and **WOODSNE** from the comparisons. We also set the maximum number of iterations to 1000 for **LANCELOT-B**, but we could not ensure a strictly comparable stopping criterion, as **LANCELOT-B** terminates when each component of the gradient is below the convergence threshold. This is slightly more restrictive than (4.2), which requires that their average is below that threshold.

NITSOL reported that no further progress was possible and stopped away from the solution for problems **BROYDNBD**, **EIGENB**, **EIGENC**, **MSQRTA**, **MSQRTB**, **POROUS1**, **POROUS2**, **DRCAVITY3**, **EIGENAU**, and **YATP2SQ**. **LANCELOT-B** reported the same diagnostic for **POROUS1** and **POROUS2**, although it reduced the norm of the constraints

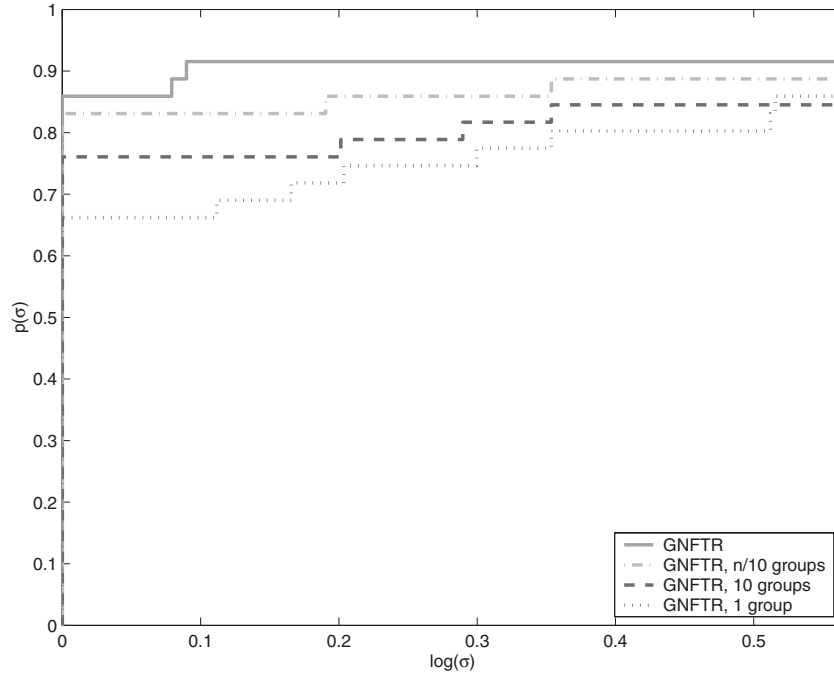


FIG. 4.6. CPU time performance profile for GNFTTR according to the number of groups.

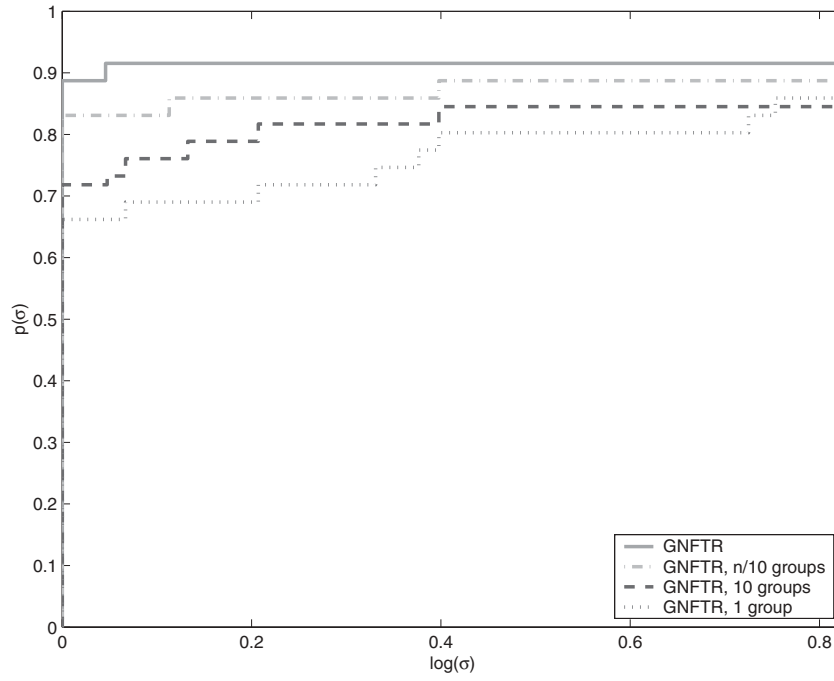


FIG. 4.7. Iterations performance profile for GNFTTR according to the number of groups.

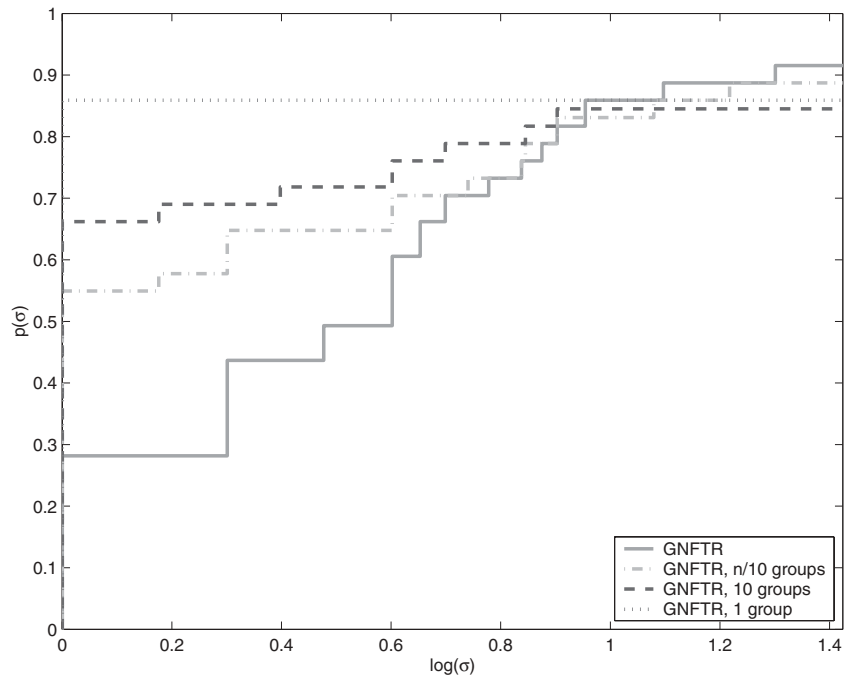


FIG. 4.8. Maximal filter size performance profile GNFTTR according to the number of groups.

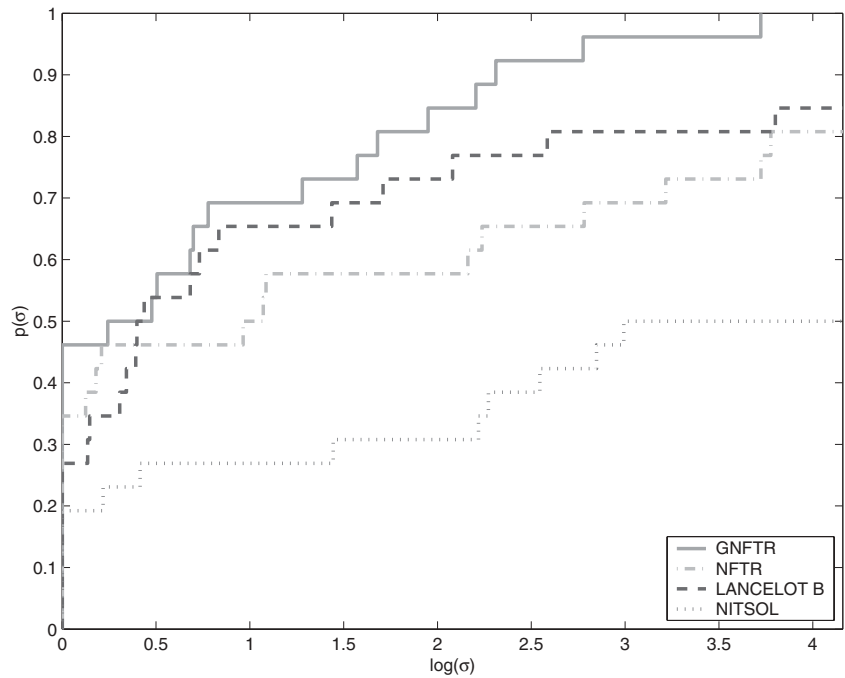


FIG. 4.9. CPU time performance profile for GNFTTR, LANCELOT-B, and NITSOL.

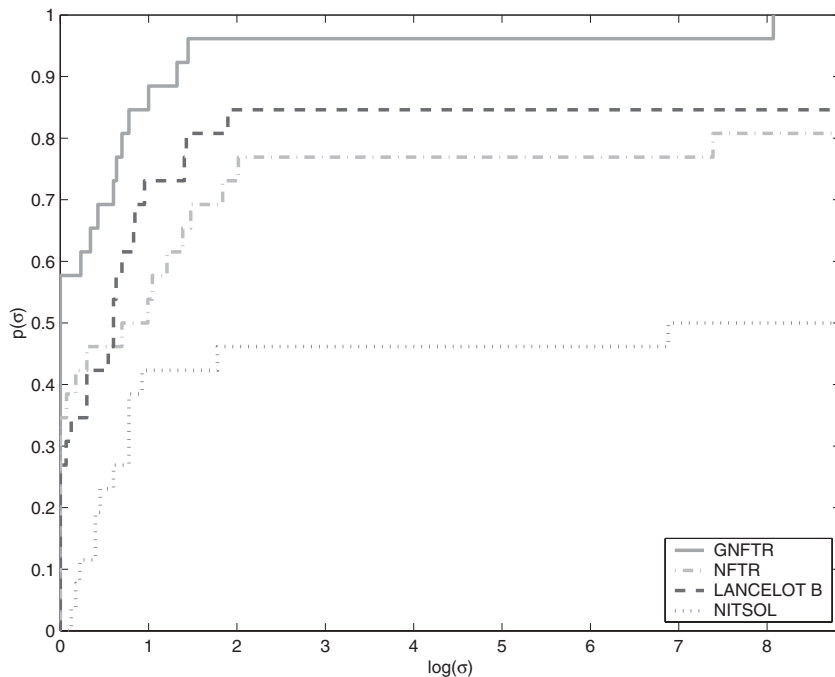


FIG. 4.10. Iterations performance profile for *GNFT*, *LANCELOT-B*, and *NITSOL*.

violation to approximately 10^{-16} . The runs were terminated after 1 hour of CPU time on problems *ARTIF*, *BRATU2D*, and *BRATU2DT* for *NITSOL* and on problems *ARTIF*, *BDVALUES*, *DRCAVTY1*, *DRCAVTY2*, *DRCAVTY3*, and *YATP2SQ* for *LANCELOT-B*.

Figures 4.9 and 4.10 give the CPU time and iteration performance profiles for the comparison. *GNFT* clearly dominates *LANCELOT-B* and *NITSOL* on our test set, both in reliability and efficiency. The comparison between *NFTR* and *LANCELOT-B* is less clear, the former dominating for small values of σ and the latter for larger ones.

5. Further comments and perspectives. We have presented a new algorithm for nonlinear equations and nonlinear least squares that blends filter and trust-region ideas. We have proved, under reasonable assumptions, that every limit point of the sequence of iterates must be a first-order stationary point of the Euclidean norm merit function. From the preliminary numerical experience presented, it appears that its efficiency is remarkably good, both in computing time and in the number of iterations and function calls. Its reliability is globally satisfying, but further work on the algorithmic heuristics is expected to bring improvements. Of course, these conclusions depend on the test problem set used, and the authors are well aware that only continued numerical investigation can confirm the suggested gains in the longer term. A Fortran 95 module, called *FILTRANE* [16], has just been introduced into the *GALAHAD* library of optimization algorithms [14] and should facilitate a wider use of the filter-trust-region algorithm.

While the analysis discussed above is highly encouraging, it represents only the first steps in the area of research that is opened by the combination of the classical trust-region method and multidimensional filters. Numerous extensions and further developments are possible, in both theory and practice. A first idea would be to

consider the values of the $c(x_k)$ themselves for inclusion in the filter, instead of their absolute values, whenever each set $\mathcal{I}_j$ contains only a single index. The filter is then no longer restrained to the positive orthant and possibly more iterates can be accepted. Interestingly, the extension of our theory to this case is straightforward. Its efficiency is under investigation. Another possibility for allowing even more iterates to be potentially acceptable is to introduce nonmonotone filter techniques similar to those discussed in Gould and Toint [15] or by modifying Algorithm 2.1 to accept x_k^+ at the next iterate if it produces a decrease in the merit function which can be judged sufficient irrespective of the step size. We could, for example, choose to accept x_k^+ whenever

$$f(x_k) - f(x_k^+) \geq \min(\epsilon, \kappa[f(x_k)]^\alpha)$$

for some positive ϵ , κ , and α . The convergence theory presented above directly extends to cover this additional condition, but its practical use remains to be explored in detail. It may also be interesting to consider norms other than the Euclidean to construct the merit function $f(x)$, the ℓ_1 and ℓ_∞ norms being obvious candidates. In this case, the subproblems consist of (approximately) solving linear programs, which, although feasible in principle, requires further research to define suitable subproblem truncation procedures.

The possibility of handling linear or nonlinear inequalities, in conjunction with equalities or alone, is also a natural development. Its theoretical aspects directly result from the present paper, since it is enough to redefine

$$\theta_j(x) = \| [c_{\mathcal{I}_j}(x)]_+ \|,$$

where the symbol $[c_k(x)]_+$ simply measures the violation of the k th constraint at x . The analysis of section 3 applies without any modification to this more general case. Practical aspects of this development are described in [16].

Finally, several strategies come to mind for coping with the (so far hypothetical) case where the memory requirements for the storage of filter entries are excessive. The first, obviously, is to use out-of-core storage. The second is to impose an upper bound on the number of filter entries and to fall back on the pure trust-region algorithm if this maximum is reached. A more sophisticated technique would be to reduce the filter space dimension by decreasing the number of equation subsets. For instance, if equations of index $i \in \mathcal{M}$ are to be merged into a new subset, the components of each filter entry whose index is $\mathcal{M}$ at iteration k , i.e., $\{\theta_{i,k}\}_{i \in \mathcal{M}}$, can be replaced by the single component

$$\sqrt{\sum_{i \in \mathcal{M}} \theta_{i,k}^2}.$$

The discussion above then suggests that the reduction in memory requirement could exceed $|\mathcal{F}| \times (|\mathcal{M}| - 1)$, as the merging process may create strongly dominated filter entries, which can then be removed. It also suggests that the deterioration in performance would remain moderate. One might also consider selectively removing filter entries that are not strongly dominated, but this last idea needs to be investigated in more detail as it is not compatible with our current proof of global convergence.

An adequate definition of equation subsets remains an interesting open question. One could, for instance, base this definition on some domain decomposition if the problem at hand arises from discretization, or one could assign different subsets to

residuals of coupled equations, if the problem has this form. Many other strategies are possible and further research is necessary to identify the better ones.

We conclude by noting that the ideas presented in this paper immediately suggest other uses for the same techniques, like the introduction of multidimensional filters in nonlinear programming (where current state-of-the-art methods use only a two-dimensional filter space) or filter techniques for the solution of nonlinear equilibrium problems (where the complementarity residual is a separate filter entry).

Acknowledgments. The authors are extremely grateful to Michael Ulbrich, Luis N. Vicente, Dominique Orban, and an anonymous referee for their useful comments on an earlier version of this paper.

REFERENCES

- [1] I. BONGARTZ, A. R. CONN, N. I. M. GOULD, AND PH. L. TOINT, *CUTE: Constrained and unconstrained testing environment*, ACM Trans. Math. Software, 21 (1995), pp. 123–160.
- [2] C. M. CHIN AND R. FLETCHER, *Convergence properties of SLP-filter algorithms that takes EQP steps*, Math. Program., 96 (2003), pp. 161–177.
- [3] A. R. CONN, N. I. M. GOULD, AND PH. L. TOINT, *LANCELOT: A Fortran Package for Large-Scale Nonlinear Optimization (Release A)*, Springer Ser. Comput. Math. 17, Springer-Verlag, Heidelberg, Berlin, New York, 1992.
- [4] A. R. CONN, N. I. M. GOULD, AND PH. L. TOINT, *Trust Region Methods*, MPS/SIAM Ser. Optim. 1, SIAM, Philadelphia, 2000.
- [5] J. E. DENNIS, D. M. GAY, AND R. E. WELSH, *An adaptive nonlinear least squares algorithm*, ACM Trans. Math. Software, 7 (1981), pp. 348–368.
- [6] E. D. DOLAN AND J. J. MORÉ, *Benchmarking optimization software with performance profiles*, Math. Program., 91 (2002), pp. 201–213.
- [7] R. FLETCHER, N. I. M. GOULD, S. LEYFFER, PH. L. TOINT, AND A. WÄCHTER, *Global convergence of a trust-region SQP-filter algorithm for general nonlinear programming*, SIAM J. Optim., 13 (2002), pp. 635–659.
- [8] R. FLETCHER AND S. LEYFFER, *Nonlinear programming without a penalty function*, Math. Program., 91 (2002), pp. 239–269.
- [9] R. FLETCHER AND S. LEYFFER, *Filter-type algorithms for solving systems of algebraic equations and inequalities*, in High Performance Algorithms and Software in Nonlinear Optimization, G. Di Pillo and A. Murli, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003, pp. 259–278.
- [10] R. FLETCHER, S. LEYFFER, AND PH. L. TOINT, *On the global convergence of a filter—SQP algorithm*, SIAM J. Optim., 13 (2002), pp. 44–59.
- [11] C. C. GONZAGA, E. KARAS, AND M. VANTI, *A globally convergent filter method for nonlinear programming*, SIAM J. Optim., 14 (2003), pp. 646–669.
- [12] N. I. M. GOULD, S. LUCIDI, M. ROMA, AND PH. L. TOINT, *Solving the trust-region subproblem using the Lanczos method*, SIAM J. Optim., 9 (1999), pp. 504–525.
- [13] N. I. M. GOULD, D. ORBAN, AND PH. L. TOINT, *CUTEr, a constrained and unconstrained testing environment, revisited*, ACM Trans. Math. Software, 29 (2003), pp. 373–394.
- [14] N. I. M. GOULD, D. ORBAN, AND PH. L. TOINT, *GALAHAD—a library of thread-safe Fortran 90 packages for large-scale nonlinear optimization*, ACM Trans. Math. Software, 29 (2003), pp. 353–372.
- [15] N. I. M. GOULD AND PH. L. TOINT, *Global Convergence of a Non-monotone Trust-Region Filter Algorithm for Nonlinear Programming*, Tech. Report 02/15, Department of Mathematics, University of Namur, Namur, Belgium, 2002.
- [16] N. I. M. GOULD AND PH. L. TOINT, *FILTRANE: A Fortran 95 Filter-Trust-Region Package for Solving Systems of Nonlinear Equalities, Nonlinear Inequalities and Nonlinear Least-Squares Problems*, Tech. Report 03/15, Rutherford Appleton Laboratory, Chilton, Oxfordshire, England, 2003.
- [17] N. I. M. GOULD AND PH. L. TOINT, *Global convergence of a hybrid trust-region SQP-filter algorithm for general nonlinear programming*, in System Modeling and Optimization XX, E. Sachs and R. Tichatschke, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003, pp. 23–54.

- [18] X. KE AND J. HAN, *A class of nonmonotone trust region algorithms for constrained optimizations*, Chinese Sci. Bull., 40 (1995), pp. 1321–1324.
- [19] L. LUKŠAN, *Hybrid methods for large sparse nonlinear least-squares*, J. Optim. Theory Appl., 89 (1996), pp. 575–595.
- [20] J. J. MORÉ AND D. C. SORESENSEN, *Computing a trust region step*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 553–572.
- [21] M. PERNICE AND H. F. WALKER, *NITSOL: A Newton iterative solver for nonlinear systems*, SIAM J. Sci. Comput., 19 (1998), pp. 302–318.
- [22] PH. L. TOINT, *On large scale nonlinear least squares calculations*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 416–435.
- [23] PH. L. TOINT, *A non-monotone trust-region algorithm for nonlinear optimization subject to convex constraints*, Math. Program., 77 (1997), pp. 69–94.
- [24] M. ULBRICH, S. ULBRICH, AND L. N. VICENTE, *A globally convergent primal-dual interior point filter method for nonconvex nonlinear programming*, Math. Program. Ser. A, 100 (2004), pp. 379–410.
- [25] A. WÄCHTER AND L. T. BIEGLER, *Global and Local Convergence of Line Search Filter Methods for Nonlinear Programming*, Tech. Report CAPD B-01-09, Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA, 2001.
- [26] Y. XIAO AND E. K. W. CHU, *Nonmonotone Trust Region Methods*, Tech. Report 95/17, Monash University, Clayton, Australia, 1995.

SEMISMOOTH NEWTON AND AUGMENTED LAGRANGIAN METHODS FOR A SIMPLIFIED FRICTION PROBLEM*

GEORG STADLER[†]

Abstract. In this paper a simplified friction problem and iterative second-order algorithms for its solution are analyzed in infinite dimensional function spaces. Motivated from the dual formulation, a primal-dual active set strategy and a semismooth Newton method for a regularized problem as well as an augmented Lagrangian method for the original problem are presented and their close relation is analyzed. Local as well as global convergence results are given. By means of numerical tests, we discuss among others convergence properties, the dependence on the mesh, and the role of the regularization and illustrate the efficiency of the proposed methodologies.

Key words. friction problem, semismooth Newton method, augmented Lagrangians, primal-dual active set algorithm

AMS subject classifications. 74M10, 65K05, 35J85, 49M15

DOI. 10.1137/S1052623403420833

1. Introduction. This paper is devoted to the convergence analysis of iterative algorithms for the solution of mechanical problems involving friction. As a model problem we consider a simplified friction problem that can be stated as the minimization of the nondifferentiable functional

$$(P) \quad \begin{cases} J(y) := \frac{1}{2}a(y, y) - (f, y)_{L^2(\Omega)} + g \int_{\Gamma_f} |\tau y(x)| dx \\ \text{over the set } Y := \{y \in H^1(\Omega) : \tau y = 0 \text{ a.e. on } \Gamma_0\}, \end{cases}$$

where $\Omega \subset \mathbb{R}^n$ is an open bounded domain with Lipschitz continuous boundary Γ , $\Gamma_0 \subset \Gamma$ is a possibly empty open set, $\Gamma_f := \Gamma \setminus \Gamma_0$, further $g > 0$, $f \in L^2(\Omega)$, τ denotes the trace operator, and $a(\cdot, \cdot)$ denotes a coercive bilinear form on $Y \times Y$. Introducing for $y \in H^1(\Omega)$ the abbreviation

$$j(y) := g \int_{\Gamma_f} |\tau y(x)| dx,$$

it is well known (cf. [12]) that (P) can be written as the elliptic variational inequality of the second kind

$$(1.1) \quad \begin{cases} \text{Find } y \in Y \text{ such that} \\ a(y, z - y) + j(z) - j(y) \geq (f, z - y)_{L^2(\Omega)} \text{ for all } z \in Y. \end{cases}$$

While usually in engineering papers finite dimensional discretizations of (P) or (1.1) and related problems are studied, little attention has been paid to their infinite dimensional counterparts, specifically to Newton-type methods. This contribution focuses on the formulation and analysis of second-order solution algorithms for (P) in a function space framework. Such an infinite dimensional analysis gives more

*Received by the editors January 7, 2003; accepted for publication (in revised form) February 11, 2004; published electronically October 1, 2004. This work was supported by the “Fonds zur Förderung der wissenschaftlichen Forschung” under “SRC 03, Optimization and Control.”
<http://www.siam.org/journals/siopt/15-1/42083.html>

[†]Institut für Mathematik, Karl-Franzens-Universität Graz, Heinrichstraße 36, A-8010 Graz, Austria (ge.stadler@uni-graz.at).

insight into the problem, which is also of significant practical importance since the performance of a numerical algorithm is closely related to the infinite dimensional problem structure. In particular, it is desirable that the numerical method can be considered as the discrete version of a well-defined and well-behaved algorithm for the continuous problem. A finite dimensional approach misses important features as, for example, the regularity of Lagrange multipliers and its consequences as well as smoothing and uniform definiteness properties of the involved operators. It is well accepted that these properties significantly influence the behavior of numerical algorithms.

In principal there are two approaches to overcome the difficulty associated with the nondifferentiability in (P). One is based on resolving the derivative of the absolute value function introducing a Lagrange multiplier; the other one is based on an appropriate smoothing of the nondifferentiable term $j(\cdot)$.

An overrelaxation method and the Uzawa algorithm are proposed in the monographs [12, 11] for the solution of (P), and convergence results for these first-order methods are given. The Uzawa method is also suggested for a variational inequality of the second kind in [14]; however, in that paper no numerical results are given. In [19] iterative techniques for the solution of friction contact problems are presented and further developed in [15]. Those methods require minimization of a nondifferentiable functional over a convex set in every iteration step, which also motivates our investigation of problem (P).

In [7, 5, 6] a generalized differentiability concept (Pang's B-differential) is used that allows application of a Newton-like method for discretizations of friction contact problems, whereas algorithm formulation and analysis are done in finite dimensional spaces and only few convergence rate results are given. The authors of those contributions report on good numerical results and, in [6], an almost mesh independent behavior of the algorithm is observed, which suggests that the finite dimensional method is induced by an infinite dimensional one. A different approach towards numerical realization of discrete elliptic variational inequalities of the second kind was followed in [26, 25], where monotone multigrid methods are employed to derive an efficient solution method.

For a smoothed variational inequality of the second kind, again in [11] the Uzawa method is proposed. More recent contributions apply classical Newton methods to the smoothed finite dimensional problems (see, e.g., [29]).

While there is a large literature on finite dimensional constrained and nondifferential optimization techniques (see, e.g., [10, 31, 30] for finite dimensional semismooth Newton methods), the systematic analysis of these methods in continuous function spaces started only rather recently [16, 33]. The methods proposed in this paper are related to the primal-dual active set strategy for the solution of constrained optimal control problems [1, 2]. This algorithm is closely related to semismooth Newton methods as shown in [16]. The papers [1, 2, 16, 22] apply these methodologies to unilateral pointwise constrained optimization problems; the convergence analysis for bilaterally constrained problems (as is the dual of (P)) involves additional problems as will come out in this contribution (see also [18, 17]). The first-order augmented Lagrangian method for nonsmooth optimization is investigated within a Hilbert space framework in [21].

This paper is organized as follows: In section 2 the dual problem for (P) and the extremality conditions are determined, and further possible generalizations of the model problem to vector-valued friction problems in linear elasticity are discussed. Section 3 is devoted to a regularization procedure for the dual formulation, the

corresponding primal problem, and the convergence of the regularized problems. In section 4 we state algorithms for the solution of the regularized and the original friction problem and investigate their relation. Section 5 analyzes these algorithms and gives local as well as global convergence results. Finally, section 6 summarizes our numerical testing.

2. The dual problem. In this section we summarize basic results for (P); further we calculate the dual problem and corresponding extremality conditions and discuss the relation between friction problems in elasticity and the model problem (P).

To simplify notation we use the trace operator $\tau_f : Y \longrightarrow L^2(\Gamma_f)$ defined by $\tau_f y = (\tau y)|_{\Gamma_f}$. Before we calculate the dual problem, we state the following existence and uniqueness result.

THEOREM 2.1. *Problem (P) or equivalently (1.1) admits a unique solution $\bar{y} \in Y$.*

Note that in [8] conditions are stated that guarantee existence and uniqueness of a solution to (P) in the case that $a(\cdot, \cdot)$ is not coercive. To get a deeper insight into problem (P) we next calculate the corresponding dual problem. We start with rewriting problem (P) as

$$(2.1) \quad \inf_{y \in Y} \{F(y) + G(\tau_f y)\}$$

with the convex functionals

$$F(y) = \frac{1}{2}a(y, y) - (f, y)_{L^2(\Omega)} \quad \text{and} \quad G(\tau_f y) = g \int_{\Gamma_f} |\tau_f y(x)| dx.$$

Following [9], the dual problem can be written as

$$(2.2) \quad \sup_{\lambda \in L^2(\Gamma_f)} \{-F^*(-\tau_f^* \lambda) - G^*(\lambda)\},$$

where $F^* : Y^* \longrightarrow \mathbb{R}$ and $G^* : L^2(\Gamma_f) \longrightarrow \mathbb{R}$ denote the convex conjugate functionals to F and G , and $\tau_f^* \in \mathcal{L}(L^2(\Gamma_f), Y^*)$ is the adjoint of τ_f . In (2.2) we already identified $L^2(\Gamma_f)$ with its dual. Calculating (2.2) explicitly results in the following dual problem.

$$(P^*) \quad \sup_{|\lambda| \leq g \text{ a.e. on } \Gamma_f} J^*(\lambda) := -\frac{1}{2}a(w(\lambda), w(\lambda)),$$

where $w(\lambda)$ denotes the unique solution of

$$(2.3) \quad a(w, v) - (f, v)_{L^2(\Omega)} + (\lambda, \tau_f v)_{L^2(\Gamma_f)} = 0 \quad \text{for all } v \in Y$$

for given $\lambda \in L^2(\Gamma_f)$. From the convexity of F and G , further from the fact that there exists a $y_0 \in Y$ such that $F(y_0) < \infty$, $G(\tau_f y_0) < \infty$, and from the continuity of G at $\tau_f y_0$, it follows that

$$\inf_{y \in Y} J(y) = - \sup_{|\lambda| \leq g \text{ a.e. on } \Gamma_f} J^*(\lambda);$$

that is, no duality gap occurs (see [9]). Existence of a solution $\bar{\lambda} \in L^2(\Gamma_f)$ for the dual problem (P*) follows from Fenchel's duality theorem. Thus, by means of duality theory we have transformed (P), the unconstrained minimization of a nondifferentiable functional, into (P*), the constrained maximization of a smooth functional.

Following duality theory the solutions $\bar{y}$, $\bar{\lambda}$ of the primal problem (P) and the dual problem (P*), respectively, are connected by the extremality conditions

$$-\tau_f^* \bar{\lambda} \in \partial F(\bar{y}), \quad \bar{\lambda} \in \partial G(\tau_f \bar{y}),$$

with ∂ denoting the subdifferential. These conditions, which also characterize primal and dual solution (cf. [9]), result in

$$(2.4) \quad a(\bar{y}, v) - (f, v)_{L^2(\Omega)} + (\bar{\lambda}, \tau_f v)_{L^2(\Gamma_f)} = 0 \text{ for all } v \in Y$$

and in the complementarity condition

$$(2.5) \quad \begin{cases} \tau_f \bar{y} \leq 0 & \text{a.e. on } \mathcal{A}_- := \{x \in \Gamma_f : \bar{\lambda} = -g \text{ a.e. on } \Gamma_f\}, \\ \tau_f \bar{y} = 0 & \text{a.e. on } \mathcal{I} := \{x \in \Gamma_f : |\bar{\lambda}| < g \text{ a.e. on } \Gamma_f\}, \\ \tau_f \bar{y} \geq 0 & \text{a.e. on } \mathcal{A}_+ := \{x \in \Gamma_f : \bar{\lambda} = g \text{ a.e. on } \Gamma_f\}. \end{cases}$$

The next lemma states an equivalent expression for condition (2.5), which will be used frequently. The equivalence follows from results in convex analysis but can also easily be verified by a direct computation.

LEMMA 2.2. *Condition (2.5) can equivalently be expressed as*

$$(2.6) \quad \tau_f \bar{y} = \max(0, \tau_f \bar{y} + \sigma(\bar{\lambda} - g)) + \min(0, \tau_f \bar{y} + \sigma(\bar{\lambda} + g))$$

for every $\sigma > 0$.

Concerning a numerical solution of (P) in [11], two methods are proposed. First, an overrelaxation method for the discretized problem is described and tested; second, Uzawa's algorithm is applied to (P) and convergence results are given. Since we focus on higher-order methods, and further since for (P*) the iterates of the algorithms presented in section 4 are not contained in spaces of square integrable functions, we introduce a regularization procedure for (P*) that allows the statement and analysis of our algorithms in infinite dimensional Hilbert spaces and will be shown to be closely related to augmented Lagrangians.

We conclude this section with a brief discussion of possible generalizations of our model problem (P) to vector-valued friction problems in linear elasticity (see [8, 24]). Let $\Omega \subset \mathbb{R}^n$ be occupied by the elastic body in its nondeformed state and denote by $\mathbf{y} \in \mathbf{Y} := Y^n$ the deformation of the body that is fixed on Γ_0 , subject to a given force on Ω and to friction conditions on Γ_f . For the material law we take, e.g., the Navier–Lamé equations and denote the corresponding bilinear form on $\mathbf{Y} \times \mathbf{Y}$ by $\mathbf{a}(\cdot, \cdot)$. Then, similarly as for (P), we get the following condition for $\mathbf{y}$:

$$(2.7) \quad \mathbf{a}(\mathbf{y}, \mathbf{v}) - (\mathbf{f}, \mathbf{v})_{L^2(\Omega)} + (\boldsymbol{\lambda}, \mathbf{v}_T)_{L^2(\Gamma_f)} = 0 \text{ for all } \mathbf{v} \in \mathbf{Y},$$

where $\mathbf{L}^2(\Omega) := (L^2(\Omega))^n$, $\mathbf{L}^2(\Gamma_f) = (L^2(\Gamma_f))^n$, $\mathbf{f} \in \mathbf{L}^2(\Omega)$ is a given force, $(\cdot)_T$ denotes the tangential part along Γ_f , and $\boldsymbol{\lambda} \in \mathbf{L}^2(\Gamma_f)$. Let us assume that the region of contact between the body and the rigid foundation is fixed (bilateral contact), i.e., the normal component $\mathbf{y}_N$ of $\mathbf{y}$ along Γ_f is zero, and let us take Tresca's law to model friction on Γ_f . This leads to the conditions $\mathbf{y}_N = 0$ and $\|\boldsymbol{\lambda}\| \leq g$ on Γ_f (here $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^n$) and to

$$(2.8) \quad \begin{cases} \mathbf{y}_T = 0 & \text{a.e. on } \mathcal{I} := \{x \in \Gamma_f : \|\boldsymbol{\lambda}\| < g \text{ a.e. on } \Gamma_f\}, \\ \mathbf{y}_T = \frac{1}{g} \|\mathbf{y}_T\| \boldsymbol{\lambda} & \text{a.e. on } \mathcal{A} := \{x \in \Gamma_f : \|\boldsymbol{\lambda}\| = g \text{ a.e. on } \Gamma_f\}. \end{cases}$$

The above complementarity system models the assumption that slip begins if a certain magnitude of friction traction is exceeded. Note the similarity between (2.4), (2.5) and (2.7), (2.8). In case of planar elasticity (i.e., $n = 2$), one has that $\mathbf{y}_T = y\mathbf{t}$, where $\mathbf{t}$ is the unit tangential vector along Γ_f and y is a scalar-valued function. This allows us to replace (2.8) by (2.5); i.e., this paper covers friction in planar elasticity. For $n \geq 3$, however, compared to (2.5) the expression in (2.8) includes an inherent nonlinearity on $\mathcal{A}$. We plan to extend the results presented in this paper for the simplified model problem (P) to elasticity problems with friction for $n \geq 3$ in our future research.

3. Regularization. In this section we introduce a regularization procedure to overcome the difficulty associated with the nondifferentiability of the functional J in (P). Therefore we consequently utilize results from duality theory and discuss relations between the regularization and the primal and dual problem.

For the term $\int_{\Gamma_f} |\tau y(x)| dx$ in (P), which involves the absolute value function, many ways to construct sequences of differentiable approximations are possible (cf., e.g., [13, 20]). While these regularizations are mainly motivated by the primal problem, our approximation is motivated by considering the dual problem and by results in the context of semismooth Newton methods [16, 22, 33] and augmented Lagrangians [21]. In the corresponding primal problem the regularization turns out to be a very natural one that is related to those used in [11, 14, 28].

This section is organized as follows. After presenting the regularization for the dual problem, we calculate the corresponding primal problem and the optimality system, argue the connection with [11, 14, 22, 28], and investigate the convergence as the regularization parameter tends to infinity.

3.1. Regularization for (P*). For fixed $\gamma > 0$ and $\hat{\lambda} \in L^2(\Gamma_f)$, we consider

$$(P_\gamma^*) \quad \sup_{|\lambda| \leq g \text{ a.e. on } \Gamma_f} J_\gamma^*(\lambda) := -\frac{1}{2}a(w(\lambda), w(\lambda)) - \frac{1}{2\gamma}\|\lambda - \hat{\lambda}\|_{L^2(\Gamma_f)}^2 + \frac{1}{2\gamma}\|\hat{\lambda}\|_{L^2(\Gamma_f)}^2,$$

where again $w(\lambda)$ denotes the solution to (2.3) for given λ . This regularized problem, which has the form of the auxiliary problem in proximal point methods, is obtained from (P*) by adding

$$(3.1) \quad -\frac{1}{2\gamma}\|\lambda - \hat{\lambda}\|_{L^2(\Gamma_f)}^2 + \frac{1}{2\gamma}\|\hat{\lambda}\|_{L^2(\Gamma_f)}^2$$

to the objective functional. Standard arguments show that (P_γ^*) admits a unique solution λ_γ for every $\gamma > 0$. The second term in (3.1), which is a constant, can be neglected from the optimizational point of view; it has been introduced to get a simple connection with the corresponding primal problem (see Theorem 3.2).

In what follows we shall use $e : Y \times L^2(\Gamma_f) \longrightarrow Y^*$ defined by

$$\langle e(y, \lambda), z \rangle_{Y^*, Y} = a(y, z) - (f, z)_{L^2(\Omega)} + (\lambda, \tau_f z)_{L^2(\Gamma_f)}.$$

This allows us to write (2.3) as $e(y, \lambda) = 0$ in Y^* . We now derive the first-order optimality conditions for the constrained optimization problem (P_γ^*) using Lagrange multipliers.

THEOREM 3.1. *Let $\lambda_\gamma \in L^2(\Gamma_f)$ be the unique solution of (P_γ^*) . Then there exist $y_\gamma \in Y$ and $\xi_\gamma \in L^2(\Gamma_f)$ such that*

$$(3.2a) \quad e(y_\gamma, \lambda_\gamma) = 0 \text{ in } Y^*,$$

$$(3.2b) \quad \tau_f y_\gamma + \gamma^{-1}(\hat{\lambda} - \lambda_\gamma) - \xi_\gamma = 0 \text{ in } L^2(\Gamma_f),$$

$$(3.2c) \quad \xi_\gamma - \max(0, \xi_\gamma + \sigma(\lambda_\gamma - g)) - \min(0, \xi_\gamma + \sigma(\lambda_\gamma + g)) = 0 \text{ in } L^2(\Gamma_f)$$

hold for every $\sigma > 0$.

Note that (3.2c) is equivalent to

$$(3.3) \quad \begin{cases} \xi_\gamma \leq 0 & \text{a.e. on } \mathcal{A}_{\gamma,-} & := \{x \in \Gamma_f : \lambda_\gamma = -g \text{ a.e. on } \Gamma_f\}, \\ \xi_\gamma = 0 & \text{a.e. on } \mathcal{I}_\gamma & := \{x \in \Gamma_f : |\lambda_\gamma| < g \text{ a.e. on } \Gamma_f\}, \\ \xi_\gamma \geq 0 & \text{a.e. on } \mathcal{A}_{\gamma,+} & := \{x \in \Gamma_f : \lambda_\gamma = g \text{ a.e. on } \Gamma_f\}. \end{cases}$$

3.2. Corresponding primal problem. Next we turn our attention to the primal formulation of problem (P_γ^*) . For $\alpha \in \mathbb{R}$ we define

$$h(x, \alpha) = \begin{cases} \frac{g}{\gamma} |\gamma x + \alpha| - \frac{g^2}{2\gamma} & \text{if } |\gamma x + \alpha| > g, \\ \frac{1}{2\gamma} (\gamma x + \alpha)^2 & \text{if } |\gamma x + \alpha| \leq g. \end{cases}$$

The function h is a continuously differentiable smoothing of the absolute value function. We can now define problem (P_γ) :

$$(P_\gamma) \quad \min_{y \in Y} J_\gamma(y) := \frac{1}{2} a(y, y) - (f, y)_{L^2(\Omega)} + \int_{\Gamma_f} h(\tau_f y(x), \hat{\lambda}(x)) dx.$$

Note that the functional J_γ is uniformly convex and continuously differentiable. The next theorem clarifies the connection between (P_γ) and (P_γ^*) . The proof uses standard arguments from duality theory.

THEOREM 3.2. *Problem (P_γ^*) is the dual problem of (P_γ) and we have $J_\gamma^*(\lambda_\gamma) = -J_\gamma(y_\gamma)$, where λ_γ and y_γ denote the solutions of (P_γ^*) and (P_γ) , respectively. Furthermore, if one introduces the variable $\xi_\gamma := \tau_f y_\gamma + \gamma^{-1}(\hat{\lambda} - \lambda_\gamma) \in L^2(\Gamma_f)$, the extremality conditions yield (3.2a)–(3.2c) and these conditions are sufficient for λ_γ and y_γ to be the solution of (P_γ^*) and (P_γ) , respectively.*

We next discuss by what means the above regularization is related to those used in other papers. First consider the case that $\hat{\lambda} \equiv 0$. Then the smoothing of the absolute value function in (P_γ) results in

$$h(x) = \begin{cases} g|x| - \frac{g^2}{2\gamma} & \text{if } |x| \geq \frac{g}{\gamma}, \\ \frac{\gamma}{2} x^2 & \text{if } |x| < \frac{g}{\gamma}. \end{cases}$$

This approximation of the absolute value function has also been studied and used in [11, 14, 28] for the numerical solution of related problems. Let us now argue the relation of the above regularization to the one in [22], where $\hat{\lambda}$ is now arbitrary. For this purpose we choose $\sigma := \gamma^{-1}$ in the complementarity condition (3.2c) and eliminate the variable ξ_γ using (3.2b). This gives

$$(3.4) \quad \tau_f y_\gamma + \gamma^{-1}(\hat{\lambda} - \lambda_\gamma) - \max(0, \tau_f y_\gamma + \gamma^{-1}(\hat{\lambda} - g)) - \min(0, \tau_f y_\gamma + \gamma^{-1}(\hat{\lambda} + g)) = 0.$$

One can see that the specific choice of $\sigma := \gamma^{-1}$ results in eliminating the variable λ_γ in the max- and min-function, which is of interest regarding semismooth Newton methods, as will become clear in the next section. In [22] a formulation related to (3.4) was successfully used to construct an effective algorithm for unilaterally constrained variational problems of the first kind. However, in the case of (P_γ^*) , which is a bilaterally constrained optimization problem, (3.4) may mislead us to an algorithm, which is less efficient. Obviously, from the theoretical point of view, the two formulations are equivalent, but the splitting of (3.4) into (3.2b) and (3.2c) contains the additional parameter σ and thus motivates a slightly different algorithm, as will be discussed in section 4. Next we investigate the convergence as $\gamma \rightarrow \infty$.

3.3. Convergence of the regularized problems. We conclude this section with a convergence result with respect to the regularization parameter γ (for related results see [11, 22]).

THEOREM 3.3. *For any $\hat{\lambda} \in L^2(\Gamma_f)$ the solutions y_γ of the regularized problems (P_γ) converge to the solution $\bar{y}$ of the original problem (P) strongly in Y as $\gamma \rightarrow \infty$. Furthermore, the solutions λ_γ of the dual problems (P_γ^*) converge to the solution $\bar{\lambda}$ of (P^*) weakly in $L^2(\Gamma_f)$.*

Proof. Recall the complementarity conditions and the definition of the active and inactive sets (2.5) for the original and (3.3) for the regularized problem. Note that for all $\gamma > 0$ we have $|\lambda_\gamma| \leq g$ a.e. on Γ_f . We now choose an arbitrary sequence γ_n such that $\gamma_n \rightarrow \infty$ for $n \rightarrow \infty$. From the weak compactness of the unit sphere in a Hilbert space, we can infer the existence of $\lambda^* \in L^2(\Gamma_f)$ and a subsequence $\lambda_{\gamma_{n_k}}$ in $L^2(\Gamma_f)$ such that

$$\lambda_{\gamma_{n_k}} \rightharpoonup \lambda^* \text{ weakly in } L^2(\Gamma_f).$$

Since closed convex sets in Hilbert spaces are weakly closed, we have $|\lambda^*| \leq g$ a.e. in $L^2(\Gamma_f)$. The weak convergence of $(\lambda_{\gamma_{n_k}})$ in $L^2(\Gamma_f)$ implies $y_{\gamma_{n_k}} \rightharpoonup y^*$ weakly in Y for some $y^* \in Y$ and that the pair (y^*, λ^*) also satisfies $e(y^*, \lambda^*) = 0$. We henceforth drop the subscript n_k with γ_{n_k} . It follows that

$$(3.5) \quad a(y_\gamma - \bar{y}, y_\gamma - \bar{y}) = (\tau_f \bar{y}, \lambda_\gamma - \bar{\lambda})_{L^2(\Gamma_f)} + (\tau_f y_\gamma, \bar{\lambda} - \lambda_\gamma)_{L^2(\Gamma_f)}.$$

We are now going to estimate the above two terms separately. Let us first turn our attention to the term $(\tau_f \bar{y}, \lambda_\gamma - \bar{\lambda})_{L^2(\Gamma_f)}$. We have that

$$\tau_f \bar{y}(\lambda_\gamma - \bar{\lambda}) = \tau_f \bar{y}(\lambda_\gamma - g) \leq 0 \text{ a.e. on } \mathcal{A}_+,$$

since $\tau_f \bar{y} \geq 0$ and $\lambda_\gamma \leq g$. Similarly we find that $\tau_f \bar{y}(\lambda_\gamma - \bar{\lambda}) \leq 0$ on $\mathcal{A}_-$ utilizing $\tau_f \bar{y} \leq 0$ and $\lambda_\gamma \geq -g$. Finally, on $\mathcal{I}$ we have $\tau_f \bar{y} = 0$ which yields, since $\Gamma_f = \mathcal{A}_- \cup \mathcal{A}_+ \cup \mathcal{I}$, that

$$(3.6) \quad (\tau_f \bar{y}, \lambda_\gamma - \bar{\lambda})_{L^2(\Gamma_f)} \leq 0.$$

Next we consider $\tau_f y_\gamma(\bar{\lambda} - \lambda_\gamma)$ on the sets $\mathcal{A}_{\gamma,-}$, $\mathcal{A}_{\gamma,+}$, and $\mathcal{I}_\gamma$, which also form a disjoint splitting of Γ_f . On $\mathcal{A}_{\gamma,-}$ the variable λ_γ is equal to $-g$ and $\gamma \tau_f y_\gamma + \hat{\lambda} \leq -g$ holds. This implies

$$(3.7) \quad \tau_f y_\gamma(\bar{\lambda} - \lambda_\gamma) = \tau_f y_\gamma(\bar{\lambda} + g) \leq \gamma^{-1}(-g - \hat{\lambda})(\bar{\lambda} + g) \text{ a.e. on } \mathcal{A}_{\gamma,-}.$$

By a similar calculation one finds

$$(3.8) \quad \tau_f y_\gamma(\bar{\lambda} - \lambda_\gamma) \leq \gamma^{-1}(g - \hat{\lambda})(\bar{\lambda} - g) \text{ a.e. on } \mathcal{A}_{\gamma,+}.$$

On $\mathcal{I}_\gamma$ we have $\lambda_\gamma = \gamma \tau_f y_\gamma + \hat{\lambda}$ and thus $|\gamma \tau_f y_\gamma + \hat{\lambda}| < g$, which shows that a.e.

$$(3.9) \quad \begin{aligned} \tau_f y_\gamma(\bar{\lambda} - \lambda_\gamma) &= \tau_f y_\gamma(\bar{\lambda} - \gamma \tau_f y_\gamma - \hat{\lambda}) = -\gamma |\tau_f y_\gamma|^2 + \tau_f y_\gamma(\bar{\lambda} - \hat{\lambda}) \\ &\leq -\gamma |\tau_f y_\gamma|^2 + |\tau_f y_\gamma| |\bar{\lambda} - \hat{\lambda}| \leq -\gamma |\tau_f y_\gamma|^2 + \gamma^{-1}(g + |\hat{\lambda}|) |\bar{\lambda} - \hat{\lambda}|. \end{aligned}$$

Hence, using (3.7), (3.8), and (3.9), one gets

$$(3.10) \quad (\tau_f y_\gamma, \bar{\lambda} - \lambda_\gamma)_{L^2(\Gamma_f)} \leq \gamma^{-1}(g + |\hat{\lambda}|, |\bar{\lambda}| + |\hat{\lambda}| + g)_{L^2(\Gamma_f)}.$$

Using (3.5), (3.6), (3.10), and the coercivity (with constant $\nu > 0$) of $a(\cdot, \cdot)$ on Y , we can estimate

$$(3.11) \quad \begin{aligned} 0 &\leq \limsup_{\gamma \rightarrow \infty} \nu \|y_\gamma - \bar{y}\|_{H^1(\Omega)}^2 \leq \limsup_{\gamma \rightarrow \infty} a(y_\gamma - \bar{y}, y_\gamma - \bar{y}) \\ &\leq \lim_{\gamma \rightarrow \infty} (\tau_f y_\gamma, \bar{\lambda} - \lambda_\gamma)_{L^2(\Gamma_f)} \leq \lim_{\gamma \rightarrow \infty} \gamma^{-1} (g + |\hat{\lambda}|, |\bar{\lambda}| + |\hat{\lambda}| + g)_{L^2(\Gamma_f)} = 0. \end{aligned}$$

It follows that $y_\gamma \rightarrow \bar{y}$ strongly in Y and hence $y^* = \bar{y}$. Passing to the limit in $e(y_\gamma, \lambda_\gamma) = 0$ and using that weak limits are unique imply $\lambda^* = \bar{\lambda}$. Thus we have proved that every sequence γ_n with $\gamma_n \rightarrow \infty$ for $n \rightarrow \infty$ contains a subsequence γ_{n_k} such that $\lambda_{\gamma_{n_k}} \rightharpoonup \bar{\lambda}$ in $L^2(\Gamma_f)$ and $y_{\gamma_{n_k}} \rightarrow \bar{y}$ in Y . Since $(\bar{y}, \bar{\lambda})$ is the unique solution to (3.2a)–(3.2c), the whole family $\{(y_\gamma, \lambda_\gamma)\}$ converges in the sense given in the statement of the theorem. $\square$

As a corollary to the proof of Theorem 3.3 one obtains a convergence rate of y_γ to $\bar{y}$.

COROLLARY 3.4. *Let y_γ and $\bar{y}$ be solutions of (P_γ) and (P) , respectively. Then there exists a $C > 0$ independent of γ such that*

$$(3.12) \quad \|y_\gamma - \bar{y}\|_{H^1(\Omega)} \leq \frac{C}{\sqrt{\gamma}}.$$

Proof. The inequality follows from (3.11) and the coercivity of $a(\cdot, \cdot)$ on Y . $\square$

4. Algorithms for the solution of (P_γ^*) and (P^*) . In this section we present iterative algorithms to solve (P_γ^*) and (P^*) . To simplify notation we drop the subscript γ for the iterates $(y_\gamma^k, \lambda_\gamma^k, \xi_\gamma^k)$ of the algorithms. The solution variables of the regularized problem are still denoted by $(y_\gamma, \lambda_\gamma, \xi_\gamma)$.

4.1. Primal-dual active set algorithm for (P_γ^*) . The primal-dual active set algorithm (PDAS) is related to the algorithms in [1, 2, 18, 17, 22] in the context of constrained optimal control and obstacle problems. It is an iterative algorithm which uses the current variables λ^k, ξ^k for (P_γ^*) to predict new active sets $\mathcal{A}_-^{k+1}, \mathcal{A}_+^{k+1}$ for the constrained optimization problem (P_γ^*) , whereas this prediction is motivated from expressing the complementarity condition in the form (3.2c). On these active sets the variable λ^{k+1} is fixed. Thus in each iteration step the method requires solving the equality constrained problem

$$(4.1) \quad \sup_{\lambda \in L^2(\Gamma_f)} J_\gamma^*(\lambda) \quad \text{s.t.} \quad \lambda = -g \text{ on } \mathcal{A}_-^{k+1}, \lambda = g \text{ on } \mathcal{A}_+^{k+1},$$

which admits a unique solution. Note that, compared to inequality constrained optimization, equality constrained problems are significantly easier to handle, both theoretically and numerically. The algorithm is specified next.

Algorithm 1: (PDAS)

1. Choose $y^0 \in \{y \in Y : \frac{\partial y}{\partial n}|_{\Gamma_f} \in L^2(\Gamma_f)\}$, $\sigma > 0$ and set $\lambda^0 := -\frac{\partial y^0}{\partial n}|_{\Gamma_f}$, $\xi^0 := \tau_f y^0 + \gamma^{-1}(\hat{\lambda} - \lambda^0)$, $k := 0$.
2. Determine

$$\begin{aligned} \mathcal{A}_-^{k+1} &= \{x \in \Gamma_f : \xi^k + \sigma(\lambda^k + g) < 0\}, \\ \mathcal{A}_+^{k+1} &= \{x \in \Gamma_f : \xi^k + \sigma(\lambda^k - g) > 0\}, \\ \mathcal{I}^{k+1} &= \Gamma_f \setminus (\mathcal{A}_-^{k+1} \cup \mathcal{A}_+^{k+1}). \end{aligned}$$

3. If $k \geq 1$, $\mathcal{A}_-^{k+1} = \mathcal{A}_-^k$, and $\mathcal{A}_+^{k+1} = \mathcal{A}_+^k$ stop, else

4. Solve problem (4.1) for λ^{k+1} on $\mathcal{I}^{k+1}$ and the corresponding $y^{k+1} \in Y$ and update

$$\xi^{k+1} = \begin{cases} \tau_f y^{k+1} + \gamma^{-1}(\hat{\lambda} + g) & \text{on } \mathcal{A}_-^{k+1}, \\ \tau_f y^{k+1} + \gamma^{-1}(\hat{\lambda} - g) & \text{on } \mathcal{A}_+^{k+1}, \\ 0 & \text{on } \mathcal{I}^{k+1}, \end{cases}$$

$k := k + 1$ and goto step 2.

Note that ξ^{k+1} is the Lagrange multiplier for the equality constraints in (4.1). The justification of the stopping criterion in step 3 of (PDAS) is given in the following lemma (see also [2]). The proof relies on the fact that, if the active sets coincide for two consecutive iterations, the quantities ξ and λ satisfy the sign-structure as required by the complementarity conditions (3.3).

LEMMA 4.1. *If Algorithm (PDAS) stops, the last iterate is the solution to system (3.2a)–(3.2c).*

We now discuss the influence of the parameter σ on the iteration sequence for $k \geq 1$. On $\mathcal{I}^k$ we have that $\xi^k = 0$ and thus σ has no influence when determining the new active and inactive sets. On $\mathcal{A}_-^k$ we have $\lambda^k = -g$ and distinguish two cases: The set where $\xi^k < 0$ belongs to $\mathcal{A}_-^{k+1}$ for the next iteration independently from σ . In case $\xi^k > 0$ we have $\xi^k + \sigma(\lambda^k - g) = \xi^k - 2\sigma g$. The set where $\xi^k - 2\sigma g \leq 0$ moves to $\mathcal{I}^{k+1}$ if $\xi^k - 2\sigma g > 0$ to $\mathcal{A}_+^{k+1}$ for the next iteration. Hence, only in this case σ influences the sequence of iterates. Smaller values for σ make it more likely that points belong to $\mathcal{A}_-^k \cap \mathcal{A}_+^{k+1}$. A similar observation as for $\mathcal{A}_-^k$ holds true for $\mathcal{A}_+^k$, which shows that with $\sigma > 0$ one can control the probability that points are shifted from one active set to the other within one iteration. We also remark that, if for some $\sigma := \sigma_1 > 0$ one has $\mathcal{A}_-^k \cap \mathcal{A}_+^{k+1} = \mathcal{A}_+^k \cap \mathcal{A}_-^{k+1} = \emptyset$, then for every $\sigma \geq \sigma_1$ also $\mathcal{A}_-^k \cap \mathcal{A}_+^{k+1} = \mathcal{A}_+^k \cap \mathcal{A}_-^{k+1} = \emptyset$ and the sets $\mathcal{A}_-^{k+1}$, $\mathcal{A}_+^{k+1}$, and $\mathcal{I}^{k+1}$ are the same for all $\sigma \geq \sigma_1$. This observation will be of interest regarding our local convergence analysis of (PDAS).

4.2. Semismooth Newton method for (P_γ^*) . This section applies an infinite-dimensional semismooth Newton method (SS) to (P_γ^*) . We use the differentiability concept as introduced in [16], which we recall for the reader's convenience later in this section.

We start with writing the optimality system (3.2a)–(3.2c) as one nonlinear operator equation utilizing (3.4). For this purpose we denote by $\tilde{y}$ the solution to the problem

$$a(y, v) - (f, v)_{L^2(\Omega)} = 0 \text{ for all } v \in Y.$$

Further we introduce $B^{-1} \in \mathcal{L}(H^{-\frac{1}{2}}, Y)$, the solution mapping for the variational equality

$$a(y, v) - \langle \lambda, \tau_f v \rangle_{H^{-\frac{1}{2}}, H^{\frac{1}{2}}} = 0 \text{ for all } v \in Y$$

for given $\lambda \in H^{-\frac{1}{2}}(\Gamma_f)$. We can now define the Neumann-to-Dirichlet operator

$$(4.2) \quad C := \tau_f B_{|L^2(\Gamma_f)}^{-1} \in \mathcal{L}(L^2(\Gamma_f), L^2(\Gamma_f))$$

and summarize some of its properties in the next lemma.

LEMMA 4.2. *The Neumann-to-Dirichlet operator defined in (4.2) is self-adjoint, positive definite, injective, and compact.*

Proof. Self-adjointness, positive definiteness, and injectivity follow easily from the properties of $a(\cdot, \cdot)$. Sobolev's embedding theorem implies the compactness of C . $\square$

With the help of the operators B^{-1} and C one can write the solution y to $e(y, \lambda) = 0$ in Y^* for given $\lambda \in L^2(\Gamma_f)$ as $y = -B^{-1}\lambda + \tilde{y}$ and $\tau_f y$ as $-C\lambda + \tau_f \tilde{y}$. This allows elimination of the variable $\tau_f y$ in (3.4). We introduce the mapping $\bar{F} : L^2(\Gamma_f) \rightarrow L^2(\Gamma_f)$ by

$$(4.3) \quad \bar{F}(\lambda) = C\lambda - \tau_f \tilde{y} - \gamma^{-1}(\hat{\lambda} - \lambda) + \max(0, -C\lambda + \tau_f \tilde{y} + \gamma^{-1}(\hat{\lambda} - g)) \\ + \min(0, -C\lambda + \tau_f \tilde{y} + \gamma^{-1}(\hat{\lambda} + g)).$$

Note that $\bar{F}(\lambda) = 0$ characterizes λ as the solution of (P_γ^*) . In what follows we utilize for $S \subset \Gamma_f$ the extension-by-zero operator $E_S : L^2(S) \rightarrow L^2(\Gamma_f)$ defined by

$$(4.4) \quad E_S(g)(x) := \begin{cases} g(x) & \text{if } x \in S, \\ 0 & \text{else.} \end{cases}$$

Its adjoint operator $E_S^* : L^2(\Gamma_f) \rightarrow L^2(S)$ is the restriction operator onto S . Writing the optimality system as done in (4.3) suggests applying a semismooth Newton method to solve $\bar{F}(\lambda) = 0$. We briefly summarize those facts on semismooth Newton methods which are relevant for the following results. Let X, Y , and Z be Banach spaces and $F : D \subset X \rightarrow Z$ be a nonlinear mapping with open domain D .

DEFINITION 4.3. *The mapping $F : D \subset X \rightarrow Z$ is called Newton differentiable on the open subset $U \subset D$ if there exists a mapping $G : U \rightarrow \mathcal{L}(X, Z)$ such that*

$$(4.5) \quad \lim_{h \rightarrow 0} \frac{1}{\|h\|} \|F(x+h) - F(x) - G(x+h)h\| = 0$$

for every $x \in U$.

The mapping G in the above definition is referred to as *generalized derivative*. The following convergence result for a generalized Newton method holds (see [4]).

THEOREM 4.4. *Suppose that $x^* \in D$ is a solution to $F(x) = 0$ and that F is Newton differentiable in an open neighborhood U containing x^* and that $\{\|G(x)^{-1}\| : x \in U\}$ is bounded. Then the Newton-iteration*

$$(4.6) \quad x^{k+1} = x^k - G(x^k)^{-1}F(x^k)$$

converges superlinearly to x^* provided that $\|x^0 - x^*\|$ is sufficiently small.

To apply a Newton iteration to the mapping $\bar{F}$, we need to consider Newton differentiability of the max- and min-operator. For this purpose let X denote a function space of real-valued functions on some $\Omega \subset \mathbb{R}^n$, and further $\max(0, y)$ and $\min(0, y)$ the pointwise max- and min-operations, respectively. We now introduce candidates for the generalized derivatives

$$(4.7) \quad G_{\max}(y)(x) = \begin{cases} 1 & \text{if } y(x) \geq 0, \\ 0 & \text{if } y(x) < 0, \end{cases} \quad G_{\min}(y)(x) = \begin{cases} 1 & \text{if } y(x) \leq 0, \\ 0 & \text{if } y(x) > 0. \end{cases}$$

Then we have the following result (see [16]).

THEOREM 4.5. *The mappings $\max(0, \cdot) : L^q(\Omega) \rightarrow L^p(\Omega)$ and $\min(0, \cdot) : L^q(\Omega) \rightarrow L^p(\Omega)$ with $1 \leq p < q < \infty$ are Newton differentiable on $L^q(\Omega)$ with generalized derivatives $G_{\max}$ and $G_{\min}$, respectively.*

Note that Theorem 4.5 requires a norm gap (i.e., $p < q$) to hold true. In [16] it is shown that the functions defined in (4.7) cannot serve as generalized gradients if $p \geq q$. We now quote a chain rule for Newton differentiability (for a proof see [22]).

THEOREM 4.6. *Let $F_2 : Y \rightarrow X$ be an affine mapping with $F_2 y = By + b$, $B \in \mathcal{L}(Y, X)$, $b \in X$, and assume that $F_1 : D \subset X \rightarrow Z$ is Newton differentiable on the open subset $U \subset D$ with generalized derivative G . If $F_2^{-1}(U)$ is nonempty, then $F = F_1 \circ F_2$ is Newton differentiable on $F_2^{-1}(U)$ with generalized derivative given by $G(By + b)B \in \mathcal{L}(Y, Z)$ for $y \in F_2^{-1}(U)$.*

We can now apply the above results to the mapping $\bar{F}$. Observe that $Rg(C) \subset H^{\frac{1}{2}}(\Gamma_f)$ and that

$$(4.8) \quad H^{\frac{1}{2}}(\Gamma_f) \hookrightarrow L^q(\Gamma_f) \text{ for } \begin{cases} q = \frac{2(n-1)}{n-2} & \text{if } n \geq 3, \\ q < \infty & \text{if } n = 2, \end{cases}$$

where $n \geq 2$ denotes the dimension of Ω . Note that $q > 2$ for all $n \geq 2$. From Theorems 4.5 and 4.6 it follows that $\bar{F}$ is Newton differentiable on $L^2(\Gamma_f)$. A generalized derivative of $\bar{F}$ is given by

$$(4.9) \quad G_{\bar{F}}(\lambda)(\delta) = \left(E_T E_T^* C + \frac{1}{\gamma} \right) \delta$$

with the following definition for $\mathcal{A}_-$, $\mathcal{A}_+$, $\mathcal{I}$:

$$\begin{aligned} \mathcal{A}_- &= \{x \in \Gamma_f : -A\lambda + \tau_f \tilde{y} + \gamma^{-1}(\hat{\lambda} - g) \geq 0\}, \\ \mathcal{A}_+ &= \{x \in \Gamma_f : -A\lambda + \tau_f \tilde{y} + \gamma^{-1}(\hat{\lambda} + g) \leq 0\}, \\ \mathcal{I} &= \Gamma_f \setminus (\mathcal{A}_- \cup \mathcal{A}_+). \end{aligned}$$

Calculating (4.6) explicitly results in a semismooth Newton-iteration step for the solution of $\bar{F}(\lambda) = 0$ that is equal to one iteration step of (PDAS) with $\sigma = \gamma^{-1}$.

An analogous result for unilaterally constrained optimal control problems was established in [16]. Note that the norm gap required for Newton differentiability of the max- and min-function results from directly exploiting the smoothing property of the operator C . This has become possible since we chose $\sigma := \gamma^{-1}$ in (3.2c) which allowed elimination of the explicit appearance of λ in the max- and min-function. Taking advantage of this fact, the above semismooth Newton method does not require a smoothing step, as do the semismooth Newton methods in [33].

We now investigate whether (PDAS) with arbitrary $\sigma > 0$ can also be interpreted as a Newton method. We introduce $F : Y \times L^2(\Gamma_f) \times L^2(\Gamma_f) \rightarrow Y^* \times L^2(\Gamma_f) \times L^2(\Gamma_f)$ by

$$(4.10) \quad F(y, \lambda, \xi) := \begin{pmatrix} e(y, \lambda) \\ \tau_f y + \gamma^{-1}(\hat{\lambda} - \lambda) - \xi \\ \xi - \max(0, \xi + \sigma(\lambda - g)) - \min(0, \xi + \sigma(\lambda + g)) \end{pmatrix}$$

and observe that $F(y, \lambda, \xi) = 0$ characterizes y and λ as solutions to (P_γ) and (P_γ^*) , respectively. Applying now the Newton iteration (4.6) with the generalized derivative of the max- and min-function as given in (4.7) to the mapping F results in Algorithm (PDAS) which can be seen similarly as for (SS). In section 5.2 it is shown that for certain problems, (PDAS) converges locally superlinearly without the necessity of a smoothing step as used in [33] to get local superlinear convergence of semismooth Newton methods.

We conclude this section with some remarks on the development of semismooth Newton methods. The application of generalized Newton methods for semismooth problems in finite dimensions has a rather long history (see, e.g., [10, 31, 30] and the references given there). Recently, in [4, 16, 27, 33] concepts for generalized derivatives in infinite dimensions were introduced. Our work uses the notion of *slant differentiability in a neighborhood* as proposed in [16], which is a slight adaptation of the terminology introduced in [4], where the term *slant differentiability at a point* is also introduced. A similar concept is proposed in [27], where the name “Newton map” is coined. Applications of such pointwise approaches to Newton’s method, however, presuppose knowledge of the solution. The differentiability concept in [16] coincides with a specific application of the theory developed in [33]; we refer to the discussion on this relationship in [16]. As in [22] and also motivated by [27], we use instead of the notion slant differentiability in a neighborhood the name *Newton differentiability*.

4.3. Augmented Lagrangian methods for (P*). Augmented Lagrangian methods (ALMs) combine ordinary Lagrangian methods and penalty methods without suffering of the disadvantages of these methods. For instance, the augmented Lagrangian method converges without requiring that the penalty parameter tends to infinity. For a detailed discussion of these methods we refer to [3, 21].

To argue the close relation of the regularization for (P*) to augmented Lagrangians recall that (3.2b), (3.2c) can equivalently be expressed as (3.4) and, after multiplication with γ , as

$$(4.11) \quad \lambda_\gamma = \gamma \tau_f y_\gamma + \hat{\lambda} - \max(0, \gamma \tau_f y_\gamma + \hat{\lambda} - g) - \min(0, \gamma \tau_f y_\gamma + \hat{\lambda} + g).$$

The augmented Lagrangian method is an iterative algorithm for the calculation of λ in (P*).

It sets $\hat{\lambda} := \lambda^l$ in (4.11) and determines λ^{l+1} from the solution of (4.11), (3.2a). Note that the augmented Lagrangian method can be seen as an implicit version of Uzawa’s algorithm (see [21]). The whole method is specified next.

Algorithm 3: (ALM)

1. Choose $\gamma > 0$, $\lambda^0 \in L^2(\Gamma_f)$ and set $l := 0$.
2. Solve for $(y^{l+1}, \lambda^{l+1}, \xi^{l+1}) \in Y \times L^2(\Gamma_f) \times L^2(\Gamma_f)$ system (3.2a)–(3.2c) with $\hat{\lambda} := \lambda^l$.
3. Update $l := l + 1$ and goto step 2.

The auxiliary problem in step 2 of (ALM) has exactly the form of our regularized problem and can thus efficiently be solved using (PDAS) or (SS). The question arises concerning the precision to which the system in step 2 is solved. Several strategies are possible, such as solving the system exactly for all l or performing only one iteration step of the semismooth Newton method in each iteration. We tested several strategies and report on them in section 6. Note that in (ALM) the regularization parameter γ plays the role of a penalty parameter, which is not necessarily taken to infinity; nevertheless (ALM) detects the solution of (P*), as will be shown in section 5.4.

5. Convergence analysis. In this section we present local convergence results for (SS) and (PDAS) as well as a conditional global convergence result for (SS) and unconditional global convergence of (ALM).

5.1. Local superlinear convergence of (SS). In this section we give a local convergence result for algorithm (SS) for the solution of the regularized friction problem.

THEOREM 5.1. *If $\|\lambda^0 - \lambda_\gamma\|_{L^2(\Gamma_f)}$ is sufficiently small, then the iterates λ^k of (SS) converge to (λ_γ) superlinearly in $L^2(\Gamma_f)$. Furthermore, the corresponding primal iterates y^k converge superlinearly to y_γ in Y .*

Proof. We have only to show superlinear convergence of λ^k to λ_γ in $L^2(\Gamma_f)$; then superlinear convergence of y^k to y_γ in $Y \subset H^1(\Omega)$ follows since B^{-1} is continuous.

We already argued Newton differentiability of $\bar{F} \in \mathcal{L}(L^2(\Gamma_f), L^2(\Gamma_f))$. To apply Theorem 4.4 it remains to verify that the generalized gradients $G_{\bar{F}} \in \mathcal{L}(L^2(\Gamma_f), L^2(\Gamma_f))$ of $\bar{F}$ have uniformly bounded inverses. Recall the definition of the extension-by-zero operator E_- and its adjoint E_-^* as given in (4.4). Let $(h_{A_-}, h_{A_+}, h_I) \in L^2(\mathcal{A}_-) \times L^2(\mathcal{A}_+) \times L^2(\mathcal{I})$ and consider the equation

$$(5.1) \quad G_{\bar{F}}(\lambda)(\delta) = G_{\bar{F}}(\lambda)(\delta_{A_-}, \delta_{A_+}, \delta_I) = (h_{A_-}, h_{A_+}, h_I).$$

Recalling the explicit form (4.9) of $G_{\bar{F}}$, we get from (5.1) that $\delta_{A_-} = \gamma h_{A_-}$ and $\delta_{A_+} = \gamma h_{A_+}$ must hold; further

$$\left(\frac{1}{\gamma} + E_I^* C E_I\right) \delta_I = h_I - \gamma E_I^* C E_{A_-} h_{A_-} - \gamma E_I^* C E_{A_+} h_{A_+}.$$

Due to the positivity of C we can define a new scalar product $\langle\langle \cdot, \cdot \rangle\rangle$ on $L^2(\mathcal{I})$ by

$$\langle\langle x, y \rangle\rangle := \left\langle \left(\frac{1}{\gamma} + E_I^* C E_I\right) x, y \right\rangle_{L^2(\mathcal{I})} \quad \text{for } x, y \in L^2(\mathcal{I}).$$

Utilizing the positivity of C we have that the product $\langle\langle \cdot, \cdot \rangle\rangle$ is coercive with constant γ^{-1} independently from $\mathcal{I}$. Applying the Lax–Milgram lemma, one finds not only that (5.1) admits a unique solution δ_I , but also that

$$\|\delta_I\|_{L^2(\mathcal{I})} \leq \gamma \|h_I\|_{L^2(\mathcal{I})} + \gamma^2 \|C\|_{\mathcal{L}(L^2(\Gamma_f))} \{ \|h_{A_-}\|_{L^2(\mathcal{A}_-)} + \|h_{A_+}\|_{L^2(\mathcal{A}_+)} \}.$$

This proves the uniform boundedness of $G_{\bar{F}}(\lambda)^{-1}$ with respect to $\lambda \in L^2(\Gamma_f)$ and ends the proof. $\square$

5.2. Local superlinear convergence of (PDAS). As observed at the end of section 4, algorithm (PDAS) cannot directly be interpreted as a locally superlinear convergent semismooth Newton method if no smoothing steps are used. However, exploiting the role of σ in (PDAS) (see the discussion after Lemma 4.1), it turns out that local superlinear convergence holds for (PDAS) as well, provided the dimension n of Ω is 2.

COROLLARY 5.2. *Assume that $n = 2$ and $\Gamma_0 \subset \Gamma$ is a sufficiently regular subset. If $\|\lambda^0 - \lambda_\gamma\|_{L^2(\Gamma_f)}$ is sufficiently small, the iterates (y^k, λ^k) of (PDAS) with $\sigma \geq \gamma^{-1}$ converge superlinearly in $Y \times L^2(\Gamma_f)$.*

Proof. The idea of this proof is to show that in a neighborhood of the solution λ_γ the iterates λ^k of (PDAS) coincide with λ^k from (SS), which allows application of Theorem 5.1 also for (PDAS).

Step 1. We first consider only (SS) and denote by $\delta > 0$ the convergence radius of this semismooth Newton method. We introduce a δ_0 with $0 < \delta_0 \leq \delta$, which will be further specified below, and choose $\lambda^0 \in L^2(\Gamma_f)$ such that $\|\lambda^0 - \lambda_\gamma\|_{L^2(\Gamma_f)} \leq \delta_0$.

Since $\delta_0 \leq \delta$ the method converges and $\|\lambda^k - \lambda^{k+1}\|_{L^2(\Gamma_f)} \leq 2\delta_0$ for $k \geq 1$. Note that the difference of the corresponding variables $y^k - y^{k+1}$ solves

$$a(y^k - y^{k+1}, v) + (\lambda^k - \lambda^{k+1}, \tau_f v)_{L^2(\Gamma_f)} = 0 \text{ for all } v \in Y.$$

It thus follows from regularity results for mixed elliptic problems [32] that

$$\|y^k - y^{k+1}\|_{C^0(\bar{\Omega})} \leq C\|\lambda^k - \lambda^{k+1}\|_{L^2(\Gamma_f)} \text{ for some } C > 0.$$

For the corresponding traces we have

$$(5.2) \quad \|\tau_f(y^k - y^{k+1})\|_{C^0(\Gamma_f)} \leq C\|\lambda^k - \lambda^{k+1}\|_{L^2(\Gamma_f)} \leq 2C\delta_0.$$

We now show that for δ_0 sufficiently small $\mathcal{A}_-^k \cap \mathcal{A}_+^{k+1} = \mathcal{A}_+^k \cap \mathcal{A}_-^{k+1} = \emptyset$. We prove this claim by contradiction; i.e., we assume that $J = \mathcal{A}_+^k \cap \mathcal{A}_-^{k+1} \neq \emptyset$. Then, almost everywhere on J we have

$$\tau y^{k-1} + \gamma^{-1}(\hat{\lambda} - g) > 0 \text{ and } \tau y^k + \gamma^{-1}(\hat{\lambda} + g) < 0,$$

which implies $\tau(y^{k+1} - y^k) > 2g\gamma^{-1}$. Thus, utilizing (5.2)

$$(5.3) \quad 2g\gamma^{-1} < \|\tau(y^{k+1} - y^k)\|_{C^0(\Gamma_f)} \leq 2C\delta_0.$$

If we choose $\delta_0 \leq \frac{g}{C\gamma}$, relation (5.3) cannot hold true and therefore $J = \emptyset$. An analogous observation holds true for $\mathcal{A}_-^k \cap \mathcal{A}_+^{k+1}$, which shows that

$$(5.4) \quad \mathcal{A}_-^k \cap \mathcal{A}_+^{k+1} = \mathcal{A}_+^k \cap \mathcal{A}_-^{k+1} = \emptyset \text{ if } \delta_0 \leq \frac{g}{C\gamma}.$$

Step 2. Recall that the iterates of (PDAS) with $\sigma = \gamma^{-1}$ coincide with those of (SS). Thus, if $\|\lambda^0 - \lambda_\gamma\|_{L^2(\Gamma_f)} \leq \delta_0$, then $\mathcal{A}_-^k \cap \mathcal{A}_+^{k+1} = \mathcal{A}_+^k \cap \mathcal{A}_-^{k+1} = \emptyset$ for (PDAS) with $\sigma = \gamma^{-1}$. It follows from the discussion after Lemma 4.1 that for the active sets calculated from (PDAS) using $\sigma \geq \gamma^{-1}$ also $\mathcal{A}_-^k \cap \mathcal{A}_+^{k+1} = \mathcal{A}_+^k \cap \mathcal{A}_-^{k+1} = \emptyset$ holds. This shows that (SS) and (PDAS) determine the same iterates for the variable λ_γ provided that $\|\lambda^0 - \lambda_\gamma\|_{L^2(\Gamma_f)} < \delta_0$. Hence, superlinear L^2 -convergence for λ^k determined from (PDAS) holds. For the variables y^k superlinear convergence in Y follows from the continuity of the solution mapping B^{-1} . $\square$

5.3. Conditional global convergence of (SS). Our global convergence result is based on an appropriately defined functional which decays when evaluated along the iterates of the algorithm. A related strategy to prove global convergence (i.e., convergence from arbitrary initialization) is used in [23] in the context of optimal control problems. In what follows we use the notation from (PDAS) with $\sigma := \gamma^{-1}$ for (SS). For $(\lambda, \xi) \in L^2(\Gamma_f) \times L^2(\Gamma_f)$ we define the functional

$$(5.5) \quad M(\lambda, \xi) := \frac{1}{\gamma^2} \int_{\Gamma_f} |(\lambda - g)^+|^2 + |(\lambda + g)^-|^2 dx + \int_{\mathcal{A}_+^*} |\xi^-|^2 dx + \int_{\mathcal{A}_-^*} |\xi^+|^2 dx,$$

where $\mathcal{A}_+^* = \{x \in \Gamma_f : \lambda(x) \geq g\}$ and $\mathcal{A}_-^* = \{x \in \Gamma_f : \lambda(x) \leq -g\}$. By $(\cdot)^+$ and $(\cdot)^-$ we denote the positive and negative part, i.e., $(\cdot)^+ := \max(0, \cdot)$ and $(\cdot)^- := -\min(0, \cdot)$. As a preparatory step for the following estimates we prove a lemma that can easily be verified using the spectral theorem for compact and positive definite operators.

LEMMA 5.3. *Let X be a real Hilbert space with inner product $(\cdot, \cdot)$ and $C \in \mathcal{L}(X)$ be injective, self-adjoint, positive, and compact. Then*

$$(y, y) \leq \|C\|_{\mathcal{L}(X)}(C^{-1}y, y)$$

for all $y \in \text{Rg}(C)$.

Following Lemma 4.2 the Neumann-to-Dirichlet mapping C , as given in (4.2), fulfills the conditions of Lemma 5.3. Utilizing the operator C step 4 of (PDAS) implies

$$(5.6) \quad \begin{aligned} C^{-1}\tau_f y^{k+1} = -\lambda^{k+1} = & - \begin{cases} g & \text{on } \mathcal{A}_+^{k+1}, \\ \gamma\tau_f y^{k+1} + \hat{\lambda} & \text{on } \mathcal{I}^{k+1}, \\ -g & \text{on } \mathcal{A}_-^{k+1}, \end{cases} \\ \tau_f y^{k+1} + \gamma^{-1}(\hat{\lambda} - \lambda^{k+1}) - \xi^{k+1} = & 0. \end{aligned}$$

With the above notation we get

$$(5.7) \quad C^{-1}(\tau_f(y^k - y^{k+1})) = \lambda^{k+1} - \lambda^k = \begin{cases} R_{\mathcal{A}_+}^k & \text{on } \mathcal{A}_+^{k+1}, \\ \gamma(\tau_f(y^{k+1} - y^k)) + R_{\mathcal{I}}^k & \text{on } \mathcal{I}^{k+1}, \\ R_{\mathcal{A}_-}^k & \text{on } \mathcal{A}_-^{k+1}, \end{cases}$$

where

$$\begin{aligned} R_{\mathcal{A}_+}^k &= \begin{cases} 0 & \text{on } \mathcal{A}_+^{k+1} \cap \mathcal{A}_+^k, \\ g - \lambda^k < 0 & \text{on } \mathcal{A}_+^{k+1} \cap \mathcal{I}^k, \\ 2g < \gamma\xi^k & \text{on } \mathcal{A}_+^{k+1} \cap \mathcal{A}_-^k, \end{cases} \\ R_{\mathcal{I}}^k &= \begin{cases} \gamma\tau_f y^k + \hat{\lambda} - g = \gamma\xi^k \leq 0 & \text{on } \mathcal{I}^{k+1} \cap \mathcal{A}_+^k, \\ 0 & \text{on } \mathcal{I}^{k+1} \cap \mathcal{I}^k, \\ \gamma\xi^k \geq 0 & \text{on } \mathcal{I}^{k+1} \cap \mathcal{A}_-^k, \end{cases} \\ R_{\mathcal{A}_-}^k &= \begin{cases} -2g > \gamma\xi^k & \text{on } \mathcal{A}_-^{k+1} \cap \mathcal{A}_+^k, \\ -g - \lambda^k > 0 & \text{on } \mathcal{A}_-^{k+1} \cap \mathcal{I}^k, \\ 0 & \text{on } \mathcal{A}_-^{k+1} \cap \mathcal{A}_-^k. \end{cases} \end{aligned}$$

Let us denote by R^k the function defined on Γ_f , whose restrictions to $\mathcal{A}_-^{k+1}$, $\mathcal{I}^{k+1}$, and $\mathcal{A}_+^{k+1}$ coincide with $R_{\mathcal{A}_-}^k$, $R_{\mathcal{I}}^k$, and $R_{\mathcal{A}_+}^k$, respectively. Note that, from the definition of R^k , we have

$$(5.8) \quad \|R^k\|_{L^2(\Gamma_f)}^2 \leq \gamma^2 M(\lambda^k, \xi^k).$$

To shorten the notation we introduce $\delta_y^k := \tau_f(y^{k+1} - y^k)$. Multiplying (5.7) by $-\delta_y^k$ results in

$$C^{-1}(\delta_y^k)(\delta_y^k) = -R^k \delta_y^k - \chi_{\mathcal{I}^{k+1}} \gamma (\delta_y^k)^2$$

a.e. on Γ_f , where $\chi_{\mathcal{I}^{k+1}}$ denotes the characteristic function for $\mathcal{I}^{k+1}$. Thus,

$$(5.9) \quad (C^{-1}\delta_y^k, \delta_y^k)_{L^2(\Gamma_f)} = \int_{\Gamma_f} -R^k \delta_y^k dx - \gamma \int_{\mathcal{I}^{k+1}} (\delta_y^k)^2 dx \leq \|R^k\|_{L^2(\Gamma_f)} \|\delta_y^k\|_{L^2(\Gamma_f)},$$

where we used the Cauchy-Schwarz inequality and $\int_{\mathcal{I}^{k+1}} (\delta_y^k)^2 dx \geq 0$. Utilizing Lemma 5.3 for the Neumann-to-Dirichlet mapping C and (5.9) yields

$$(5.10) \quad \|\delta_y^k\|_{L^2(\Gamma_f)} \leq \|C\|_{\mathcal{L}(L^2(\Gamma_f))} \|R^k\|_{L^2(\Gamma_f)}.$$

We can now prove the following convergence theorem for (SS), or equivalently for (PDAS), with $\sigma = \gamma^{-1}$.

THEOREM 5.4. *If $\gamma < \|C\|_{\mathcal{L}(L^2(\Gamma_f))}^{-1}$, then*

$$M(\lambda^{k+1}, \xi^{k+1}) < M(\lambda^k, \xi^k)$$

for $k = 0, 1, 2, \dots$ with $(\lambda^k, \xi^k) \neq (\lambda_\gamma, \xi_\gamma)$, where (λ^k, ξ^k) denote the iterates of (PDAS) with $\sigma = \gamma^{-1}$. Moreover, (y^k, λ^k, ξ^k) converges to $(y_\gamma, \lambda_\gamma, \xi_\gamma)$ strongly in $Y \times L^2(\Gamma_f) \times L^2(\Gamma_f)$.

Proof. Recall that from the definition of (PDAS) with $\sigma = \gamma^{-1}$ one gets

$$\begin{aligned} \lambda^{k+1} &= \gamma \tau_f y^{k+1} + \hat{\lambda} && \text{on } \mathcal{I}^{k+1}, \\ \xi^{k+1} &= \tau_f y^{k+1} + \gamma^{-1}(\hat{\lambda} + g) && \text{on } \mathcal{A}_+^{k+1}, \\ \xi^{k+1} &= \tau_f y^{k+1} + \gamma^{-1}(\hat{\lambda} - g) && \text{on } \mathcal{A}_+^{k+1}. \end{aligned}$$

We therefore have

$$\xi^{k+1} = \delta_y^k + \tau_f y^k + \gamma^{-1}(\hat{\lambda} - g) = \delta_y^k + \begin{cases} \xi^k - \gamma^{-1}(g - \lambda^k) > 0 & \text{on } \mathcal{A}_+^{k+1} \cap \mathcal{A}_-^k, \\ \gamma^{-1}(\lambda^k - g) > 0 & \text{on } \mathcal{A}_+^{k+1} \cap \mathcal{I}^k, \\ \xi^k > 0 & \text{on } \mathcal{A}_+^{k+1} \cap \mathcal{A}_+^k. \end{cases}$$

Thus, $|(\xi^{k+1})^-| := |\max(0, -\xi^{k+1})| \leq |\delta_y^k|$ a.e. on $\mathcal{A}_+^{k+1}$. Note that

$$\mathcal{A}_+^{*,k+1} := \{x \in \Gamma_f : \lambda^{k+1}(x) \geq g\} = \mathcal{A}_+^{k+1} \cup \{x \in \mathcal{I}^{k+1} : \lambda^{k+1}(x) \geq g\},$$

which implies, using $\xi^{k+1} = 0$ on $\mathcal{I}^{k+1}$, that

$$(5.11) \quad |(\xi^{k+1})^-| \leq |\delta_y^k| \text{ a.e. on } \mathcal{A}_+^{*,k+1}.$$

Analogously, it follows that

$$(5.12) \quad |(\xi^{k+1})^+| \leq |\delta_y^k| \text{ a.e. on } \mathcal{A}_-^{*,k+1},$$

where $\mathcal{A}_-^{*,k+1} := \{x \in \Gamma_f : \lambda^{k+1}(x) \leq -g\}$. Moreover, on $\mathcal{I}^{k+1}$

$$\lambda^{k+1} - g = \gamma \delta_y^k + \gamma \tau_f y^k + \hat{\lambda} - g = \gamma \delta_y^k + \begin{cases} \gamma(\xi^k + \gamma^{-1}(\lambda^k - g)) \leq 0 & \text{on } \mathcal{I}^{k+1} \cap \mathcal{A}_-^k, \\ \lambda^k - g \leq 0 & \text{on } \mathcal{I}^{k+1} \cap \mathcal{I}^k, \\ \gamma \xi^k \leq 0 & \text{on } \mathcal{I}^{k+1} \cap \mathcal{A}_+^k. \end{cases}$$

The above estimate shows that

$$(5.13) \quad |(\lambda^{k+1} - g)^+| \leq \gamma |\delta_y^k| \text{ a.e. on } \mathcal{I}^{k+1},$$

and analogously one can show

$$(5.14) \quad |(\lambda^{k+1} + g)^-| \leq \gamma |\delta_y^k| \text{ a.e. on } \mathcal{I}^{k+1}.$$

Since on active sets λ^{k+1} is set to either g or $-g$, we have on $\mathcal{A}_-^{k+1} \cup \mathcal{A}_+^{k+1}$ that $(\lambda^{k+1} - g)^+ = (\lambda^{k+1} + g)^- = 0$. Further, at most one of the expressions at a.a. $x \in \Gamma_f$

$$|(\lambda^{k+1} - g)^+|, |(\lambda^{k+1} + g)^-|, |(\xi^{k+1})^-|, |(\xi^{k+1})^+|$$

can be strictly positive, which shows, combining (5.11)–(5.14), that

$$(5.15) \quad M(\lambda^{k+1}, \xi^{k+1}) \leq \|\delta_y^k\|_{L^2(\Gamma_f)}^2.$$

Combining (5.8) and (5.10) with (5.15) shows that

$$(5.16) \quad M(\lambda^{k+1}, \xi^{k+1}) \leq \gamma^2 \|C\|_{\mathcal{L}(L^2(\Gamma_f))}^2 M(\lambda^k, \xi^k).$$

Our assumption on γ implies that

$$(5.17) \quad \|C\|_{\mathcal{L}(L^2(\Gamma_f))}^2 \gamma^2 < 1,$$

which shows that

$$M(\lambda^{k+1}, \xi^{k+1}) < M(\lambda^k, \xi^k)$$

unless $(\lambda^k, \xi^k) = (\lambda_\gamma, \xi_\gamma)$. Combining (5.8), (5.10), (5.15), and (5.16) it follows that

$$(5.18) \quad \begin{aligned} \|\delta_y^k\|_{L^2(\Gamma_f)}^2 &\leq \|C\|_{\mathcal{L}(L^2(\Gamma_f))}^2 \|R^k\|_{L^2(\Gamma_f)}^2 \leq \gamma^2 \|C\|_{\mathcal{L}(L^2(\Gamma_f))}^2 M(\lambda^k, \xi^k) \\ &\leq \gamma^2 \|C\|_{\mathcal{L}(L^2(\Gamma_f))}^2 \|\delta_y^{k-1}\|_{L^2(\Gamma_f)}^2 \leq (\gamma \|C\|_{\mathcal{L}(L^2(\Gamma_f))})^{2(k+1)} M(\lambda^0, \xi^0), \end{aligned}$$

which shows, utilizing (5.17), that $\lim_{k \rightarrow \infty} M(\lambda^k, \xi^k) = \lim_{k \rightarrow \infty} \|R^k\|_{L^2(\Gamma_f)} = 0$. Further, summing up (5.18) over k and utilizing (5.17) results in $\sum_{k=1}^{\infty} \|\delta_y^k\|_{L^2(\Gamma_f)}^2 < \infty$, which shows that there exists a $z \in L^2(\Gamma_f)$ such that $\lim_{k \rightarrow \infty} \tau_j y^k = z$ in $L^2(\Gamma_f)$. Using (5.7) results in

$$\begin{aligned} \|\lambda^{k+1} - \lambda^k\|_{L^2(\Gamma_f)} &\leq \gamma \|\delta_y^k\|_{L^2(\Gamma_f)} + \|R^k\|_{L^2(\Gamma_f)} \\ &\leq (\gamma \|C\|_{\mathcal{L}(L^2(\Gamma_f))})^k (\gamma^2 \|C\|_{\mathcal{L}(L^2(\Gamma_f))} + \gamma) M(\lambda^0, \xi^0)^{\frac{1}{2}}, \end{aligned}$$

and thus there exists $\check{\lambda} \in L^2(\Gamma_f)$ with $\lim_{k \rightarrow \infty} \lambda^k = \check{\lambda}$ in $L^2(\Gamma_f)$. From (5.6) one gets $\lim_{k \rightarrow \infty} \xi^k = \check{\xi} \in L^2(\Gamma_f)$ for $\check{\xi} \in L^2(\Gamma_f)$. Since $0 = \lim_{k \rightarrow \infty} M(\lambda^k, \xi^k) = M(\check{\lambda}, \check{\xi})$, the pair $(\check{\lambda}, \check{\xi})$ satisfies condition (3.2c). From (3.2a) follows the existence of $\check{y} \in Y$ such that $\lim_{k \rightarrow \infty} y^k = \check{y}$ in Y . Note that, since (y^k, λ^k, ξ^k) satisfies (3.2a), (3.2b), this is also the case for $(\check{y}, \check{\lambda}, \check{\xi})$. Hence $(\check{y}, \check{\lambda}, \check{\xi}) = (y_\gamma, \lambda_\gamma, \xi_\gamma)$ due to the uniqueness of a solution to (3.2a)–(3.2c), which ends the proof. $\square$

Note that for unilaterally constrained problems global convergence results can possibly be gained using monotonicity properties of the involved operators (see, e.g., [22]), where the maximum principle for the Laplace operator is used to prove global convergence. For bilaterally constrained problems (such as (P_γ^*)), however, monotonicity of the iterates does not hold, even if the operator satisfies a maximum principle.

5.4. Global convergence of (ALM). The next theorem states global convergence of (ALM) for all $\gamma > 0$ and shows that large γ increases the speed of convergence. In the statement of the next theorem we denote the coercivity constant of $a(\cdot, \cdot)$ on Y by $\nu > 0$.

THEOREM 5.5. *The iterates λ^l of (ALM) and the corresponding variables y^l satisfy*

$$(5.19) \quad \nu \|y^{l+1} - \bar{y}\|_{H^1(\Omega)}^2 + \frac{1}{2\gamma} \|\lambda^{l+1} - \bar{\lambda}\|_{L^2(\Gamma_f)}^2 \leq \frac{1}{2\gamma} \|\lambda^l - \bar{\lambda}\|_{L^2(\Gamma_f)}^2$$

and

$$(5.20) \quad \nu \sum_{k=1}^{\infty} \|y^{k+1} - \bar{y}\|_{H^1(\Omega)}^2 \leq \frac{1}{2\gamma} \|\lambda^0 - \bar{\lambda}\|_{L^2(\Gamma_f)}^2,$$

which implies that $y^l \rightarrow \bar{y}$ strongly in Y and $\lambda^l \rightharpoonup \bar{\lambda}$ weakly in $L^2(\Gamma_f)$.

Proof. From the fact that $\bar{y}$ and $\bar{\lambda}$ are the solutions to (P) and (P*), respectively, it follows that

$$(5.21) \quad a(\bar{y}, y^{l+1} - \bar{y}) - (f, y^{l+1} - \bar{y})_{L^2(\Omega)} + (\bar{\lambda}, \tau_f(y^{l+1} - \bar{y}))_{L^2(\Gamma_f)} = 0,$$

and since y^{l+1} and λ^{l+1} solve (P_γ) and (P_γ^*) with $\hat{\lambda} := \lambda^l$, we infer

$$(5.22) \quad a(y^{l+1}, y^{l+1} - \bar{y}) - (f, y^{l+1} - \bar{y})_{L^2(\Omega)} + (\lambda^{l+1}, \tau_f(y^{l+1} - \bar{y}))_{L^2(\Gamma_f)} = 0.$$

Subtracting (5.21) from (5.22) results in

$$(5.23) \quad a(y^{l+1} - \bar{y}, y^{l+1} - \bar{y}) + (\lambda^{l+1} - \bar{\lambda}, \tau_f(y^{l+1} - \bar{y}))_{L^2(\Gamma_f)} = 0.$$

Note that one can write (4.11) and (2.6) as

$$(5.24) \quad \lambda^{l+1} = P(\gamma \tau_f y^{l+1} + \lambda^l) \quad \text{and} \quad \bar{\lambda} = P(\gamma \tau_f \bar{y} + \bar{\lambda}),$$

where $P : L^2(\Gamma_f) \rightarrow L^2(\Gamma_f)$ denotes the pointwise projection onto the convex set $\{v \in L^2(\Gamma_f) : |v| \leq g \text{ a.e. on } L^2(\Gamma_f)\}$. Thus we get

$$\begin{aligned} (\lambda^{l+1} - \bar{\lambda}, \tau_f(y^{l+1} - \bar{y}))_{L^2(\Gamma_f)} &= \gamma^{-1}(\lambda^{l+1} - \bar{\lambda}, (\gamma \tau_f y^{l+1} + \lambda^l) - (\gamma \tau_f \bar{y} + \bar{\lambda}))_{L^2(\Gamma_f)} \\ &\quad - \gamma^{-1}(\lambda^{l+1} - \bar{\lambda}, \lambda^l - \bar{\lambda})_{L^2(\Gamma_f)} \\ &\geq \gamma^{-1} \|\lambda^{l+1} - \bar{\lambda}\|_{L^2(\Gamma_f)}^2 - \gamma^{-1}(\lambda^{l+1} - \bar{\lambda}, \lambda^l - \bar{\lambda})_{L^2(\Gamma_f)}, \end{aligned}$$

where we used that

$$(\lambda^{l+1} - \bar{\lambda}, (\gamma \tau_f y^{l+1} + \lambda^l - \lambda^{l+1}) - (\gamma \tau_f \bar{y} + \bar{\lambda} - \bar{\lambda}))_{L^2(\Gamma_f)} \geq 0,$$

which holds using (5.24) and since P is a projection onto a convex set. Using (5.23) and the coercivity of $a(\cdot, \cdot)$ on Y , we get

$$\begin{aligned} \nu \|y^{l+1} - \bar{y}\|_{H^1(\Omega)}^2 &\leq a(y^{l+1} - \bar{y}, y^{l+1} - \bar{y}) = -(\lambda^{l+1} - \bar{\lambda}, \tau_f(y^{l+1} - \bar{y}))_{L^2(\Gamma_f)} \\ &\leq -\frac{1}{\gamma} \|\lambda^{l+1} - \bar{\lambda}\|_{L^2(\Gamma_f)}^2 + \frac{1}{\gamma} (\lambda^{l+1} - \bar{\lambda}, \lambda^l - \bar{\lambda})_{L^2(\Gamma_f)} \\ &\leq -\frac{1}{2\gamma} \|\lambda^{l+1} - \bar{\lambda}\|_{L^2(\Gamma_f)}^2 + \frac{1}{2\gamma} \|\lambda^l - \bar{\lambda}\|_{L^2(\Gamma_f)}^2, \end{aligned}$$

which proves (5.19). Summing up (5.19) with respect to l , we obtain (5.20). $\square$

TABLE 6.1
Number of iterations for different values of γ for Example 1.

γ	10^2	10^4	10^6	10^8
No. of iterations	2	3	3	3

6. Numerical tests. In this section we present three test examples for the algorithms proposed in section 5 for the solution of (P_γ^*) and (P^*) . For simplicity we use for all examples the unit square as domain, i.e., $\Omega = (0, 1) \times (0, 1)$ and the bilinear form

$$a(y, z) := (\nabla y, \nabla z)_{L^2(\Omega)} + \mu(y, z)_{L^2(\Omega)} \quad \text{for } y, z \in Y,$$

which is coercive if $\mu > 0$ or $\Gamma_0 \subset \Gamma$ has positive measure. For our calculations we utilize a finite difference discretization with the usual five-point stencil approximation to the Laplace operator. The discretization of the normal derivative is based on one-sided differences and all linear systems are solved exactly. We denote by N the number of gridpoints in one space dimension, which means we work on $N \times N$ -grids. To investigate convergence properties we frequently report on

$$(6.1) \quad d_\lambda^l := \|\bar{\lambda} - \lambda^l\|_{L^2(\Gamma_f)},$$

where $\bar{\lambda} := \lambda_{10^{16}}$ is the solution of (P_γ^*) with $\gamma = 10^{16}$ and λ^l denotes the actual iterate. We compare our results with those obtained using the Uzawa algorithm, which can be interpreted as an explicit form of the augmented Lagrangian method [21]. While (ALM) converges for every $\gamma > 0$ the Uzawa method converges for only $\gamma \in [\alpha_1, \alpha_2]$ with $0 < \alpha_1 < \alpha_2$, where α_1 and α_2 are in general not known [11]. We initialize the Uzawa method with $\lambda^0 := 0$ and report on the number of iterations (one iteration requires one linear solve), where the iteration is stopped if $d_\lambda^l < 10^{-4}$. Unless otherwise specified, we use $\hat{\lambda} = 0$, and $\sigma = 1$ for (PDAS). As initialization for (PDAS), (SS), and (ALM), the solution to (3.2a)–(3.2c) with $\xi^0 = 0$ is used, which corresponds to the solution of (P_γ^*) neglecting the constraints on λ .

6.1. Example 1. This example is taken from [11]. The data are as follows: $\Gamma_f = ([0, 1] \times \{0\}) \cup ([0, 1] \times \{1\})$, $g = 1.5$, $\mu = 0$, and

$$f(x) = \begin{cases} 10 & \text{if } x \in (0, \frac{1}{2}) \times (0, 1), \\ -10 & \text{if } x \in [\frac{1}{2}, 1) \times (0, 1). \end{cases}$$

Choosing $N = 80$ the Uzawa algorithm requires 32 iterations for $\gamma = 10$ and 17 iterations for $\gamma = 20$ and does not converge for $\gamma = 30$. In our tests for (PDAS) and (SS) we vary the value for the regularization parameter γ and investigate the convergence as $\gamma \rightarrow \infty$. Table 6.1 reports on the number of iterations needed by (SS) for various values of γ and $N = 80$. It can be seen that the algorithm requires only very few iterations to find the solution and that increasing the regularization parameter for this example does not increase the number of iterations required to detect the solution (cf. [22], where a different behavior for obstacle problems is observed). We remark that for (SS) no points are shifted from the lower active set to the upper or conversely within one iteration and thus (PDAS) determines the same iterates as (SS) for all $\sigma \geq \gamma^{-1}$.

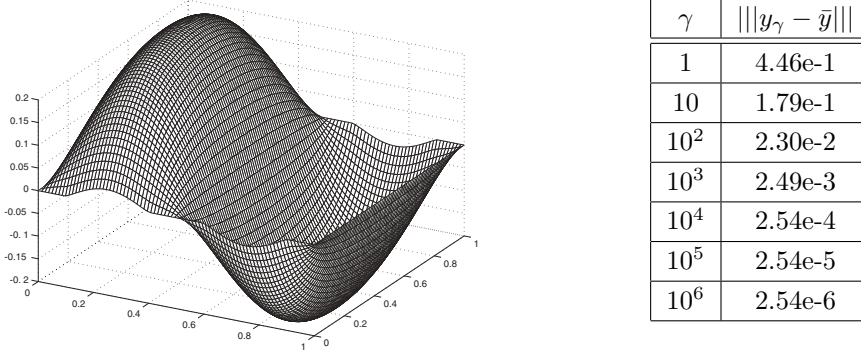


FIG. 6.1. Solution y_γ for $\gamma = 10^8$ (left) and convergence with respect to γ (right) for Example 1.

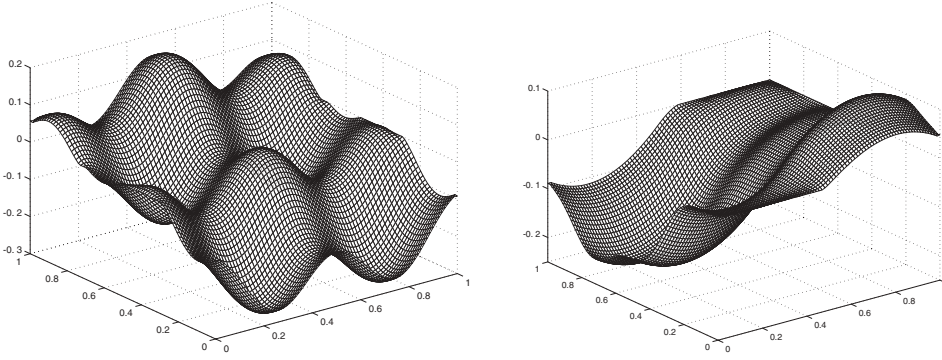


FIG. 6.2. Solution y_γ for Example 2 (left) and Example 3 (right) for $\gamma = 10^{10}$.

The primal solution y_γ for $\gamma = 10^8$ is shown in Figure 6.1 (left). We also investigate the convergence as $\gamma \rightarrow \infty$. The table in Figure 6.1 reports on the value of

$$|||y_\gamma - \bar{y}||| := (a(y_\gamma - \bar{y}, y_\gamma - \bar{y}))^{\frac{1}{2}}$$

for various γ , where we take $\bar{y} := y_{10^{16}}$ as approximation to the solution of the simplified friction problem and use $N = 160$. Note that $||| \cdot |||$ is a norm equivalent to the usual one in $H^1(\Omega)$. The result suggests the convergence rate γ^{-1} , while from Corollary 3.4 we get only the rate $\gamma^{-\frac{1}{2}}$.

6.2. Example 2. This example investigates the behavior of (PDAS), (SS), and (ALM) for a more complicated structure of the active and inactive sets (see the solution y_γ for $\gamma = 10^{10}$ in Figure 6.2 (left)). Uzawa's method faces serious troubles with this example: In all test examples with $\gamma > 0.2$ the method did not converge. For $\gamma = 0.2$ we stopped our test after 400 iterations at $d_\lambda^{400} = 1.03\text{e-}2$. We choose $\Gamma_f = \partial\Omega$, which implies $Y = H^1(\Omega)$; and further $\mu = 0.5$, $g = 0.4$, and the external force $f(x) = 10(\sin 4\pi x + \cos 4\pi x)$.

We investigate superlinear convergence of the iterates, properties of (PDAS) and (SS), and the behavior of algorithm (ALM). Table 6.2 reports on

$$q_\lambda^k := \frac{\|\lambda^k - \lambda_\gamma\|_{L^2(\Gamma_f)}}{\|\lambda^{k-1} - \lambda_\gamma\|_{L^2(\Gamma_f)}}, \quad k = 1, 2, \dots,$$

TABLE 6.2
Values for q_λ^k in Example 2, $\gamma = 50$.

k	1	2	3	4	5
q_λ^k	0.58	0.45	0.25	0.18	0.00

TABLE 6.3
Number of iterations of (SS) and (PDAS) for different values of γ for Example 2, $N = 160$.

γ	3	5	10	50	100	150	160	10^3	10^{10}
$\#iter_{SS}$	3	4	4	5	6	7	div	div	div
$\#iter_{PD}$	3	4	4	5	6	5	5	6	6

TABLE 6.4
Tests for (ALM) with exact solve of the auxiliary problem for Example 2, $N = 160$.

	l	1	2	3	4	5	6	7
$\gamma = 10^2$	$\#iter_{PD}$	6	3	2	1	1	1	1
	d_λ^l	2.55e-2	8.54e-3	3.62e-3	1.62e-3	7.45e-4	3.58e-4	1.77e-4
$\gamma = 10^4$	$\#iter_{PD}$	6	1	1	1	1		
	d_λ^l	4.00e-4	3.06e-6	2.67e-8	2.62e-10	2.84e-12		

for $\gamma = 50$. We observe superlinear convergence of the iterates determined with (SS).

Table 6.3 shows the number of iterations $\#iter_{SS}$ and $\#iter_{PD}$ required by (SS) and (PDAS), respectively, to find the solution for different values of γ . We observe a slight increase in the number of iterations as γ increases. For $\gamma \geq 160$ (SS) does not detect the solution, whereas (PDAS) does. Using (SS) for these examples we can observe the following behavior: Points in Γ_f move from $\mathcal{A}_-^{k-1}$ to $\mathcal{A}_+^k$, and then from $\mathcal{A}_+^k$ back to $\mathcal{A}_-^{k+1}$ for some $k \geq 2$, and due to this scattering the algorithm does not find the solution (cf. the remark on the role of σ in (PDAS) after Lemma 4.1). Algorithm (PDAS) with $\sigma = 1$ does not experience such problems and finds the solution after a few iterations for all tested $\gamma > 0$.

To avoid possible difficulties due to the local convergence of (SS), we test two globalization strategies: First we use a continuation procedure with respect to γ , motivated from the local convergence result for (SS). We solve for $\gamma = 150$ and use the solution as initialization for the algorithm with larger γ . This procedure turns out to be successful for only a moderate increase in γ . Increasing γ moderately, typically only one or two more iterations are needed to find the solution for larger γ . However, this method appears inconvenient and costly. Next we test backtracking with J_γ as merit function to globalize (SS). This strategy works successfully, but in particular for larger γ , several backtracking steps are necessary in each iteration. The resulting stepsize is very small and thus overall up to 50 iterations are needed to find the solution. This behavior becomes more distinct for large γ .

We also apply algorithm (ALM) for the solution of this example. In a first attempt we solve the auxiliary problem in (ALM) exactly using (PDAS), whereas this method is initialized with the solution of the auxiliary problem in the previous iteration step of (ALM). Due to the local superlinear convergence of (PDAS), the auxiliary problem is solved in very few iterations, as can be seen in Table 6.4, where for $\gamma = 10^2$ and $\gamma = 10^4$ we report on the number of iterations $\#iter_{PD}$ required by (PDAS)

TABLE 6.5
Tests for (ALM) with inexact solve of the auxiliary problem for Example 2, $N = 160$.

	l	1	2	3	4	5	6	7
$\gamma = 10^2$	#iter _{PD}	1	2	2	2	1	1	1
	d_λ^l	2.96e-1	5.03e-2	7.60e-3	1.89e-3	9.99e-4	5.36e-4	2.91e-4
$\gamma = 10^4$	#iter _{PD}	1	2	1	2	1	1	1
	d_λ^l	3.81e-1	1.46e-1	3.79e-2	1.87e-4	2.20e-6	2.67e-8	3.28e-10
$\gamma = 10^l$	#iter _{PD}	1	3	2	1	1	1	
	d_λ^l	1.41e-1	2.30e-2	2.16e-3	1.11e-5	1.04e-8	1.14e-12	

TABLE 6.6
Number of iterations for different values of γ (first line) and N (first column) for Example 3.

	5	10	30	50	10^2	10^3	10^5	10^{10}
20	3	3	3	3	3	3	3	3*
40	3	3	4	5	4	4	4	4*
80	3	4	4	5	5	6	6	6*
160	3	3	5	5	6	8	7*	7*
320	3	4	5	6	7	7	8*	8*

in every step l of (ALM). Further we report on d_λ^l as defined in (6.1). We observe a monotone decrease of d_λ^l and a faster convergence in the case that $\gamma = 10^4$. In a second approach we test an inexact version of the augmented Lagrangian method: We stop the (PDAS)-iterations for the auxiliary problem as soon as the initial residual in (3.2c) has been reduced at least by a factor of 2. The results for $\gamma = 10^2, 10^4$ are shown in Table 6.5, where we report on the number of (PDAS)-iterations and the value of d_λ^l . We observe that for $\gamma = 10^2$ the first iterates present a better approximation to $\bar{\lambda}$ than for $\gamma = 10^4$, whereas then the case $\gamma = 10^4$ shows a faster convergence behavior. This leads to the idea of increasing the parameter γ in every step of (ALM). Choosing $\gamma = 10^l$ in the l th (ALM)-iteration, the impressing results shown in the last line of Table 6.5 were obtained.

6.3. Example 3. The solution $y_{10^{10}}$ for this last example is shown in Figure 6.2 (right). Again the Uzawa algorithm only converges for small γ which results in an extremely slow convergence. The data are as follows: $\Gamma_f = \partial\Omega$, $\mu = 0.5$, $g = 0.3$, and

$$f(x) = |3x - 1| + 2\text{sgn}(2y - 1) + 2\text{sgn}(x - 0.75) + 5\sin(6\pi x).$$

We investigate the number of iterations required by (PDAS) and (SS) for various values of γ and N ; further we report on results obtained with (ALM). In a series of test runs (Table 6.6) we investigate the number of iterations for various grids and different values for γ . We observe the low number of iterations for all mesh-sizes and choices for γ . For the calculations in Table 6.6 we use (SS), except for those indicated by *. For these examples with a rather large value of γ , (SS) starts to chatter due to effects described in the previous example. Thus we utilize (PDAS) to solve these problems, which is always successful. Utilizing (PDAS) also for the examples with smaller γ yields the same number of iterations as (SS). Further, note from Table 6.6 that the number of iterations increases very weakly as the mesh becomes finer.

TABLE 6.7
Tests for (ALM) with inexact solve of the auxiliary problem for Example 3, $N = 160$.

	l	1	2	3	4	5	6	7
$\gamma = 10^2$	#iter _{PD}	1	2	1	1	1	1	2
	d_λ^l	1.62e-1	1.99e-2	6.94e-3	1.40e-3	4.67e-4	1.91e-4	9.20e-5
$\gamma = 10^4$	#iter _{PD}	1	3	2	1	1	1	
	d_λ^l	2.14e-1	1.46e-2	4.11e-4	9.06e-7	8.20e-9	8.43e-11	
$\gamma = 10^l$	#iter _{PD}	1	2	2	1	1	1	
	d_λ^l	8.54e-2	2.62e-2	2.40e-3	4.21e-4	9.08e-8	8.23e-12	

We again test (ALM) with (PDAS) for the auxiliary problem. In the case that this inner problem is solved exactly, the overall number of system solves is 16 for $\gamma = 50$ and between 10 and 20 for other tested values of γ , where we used $d_\lambda^l < 10^{-4}$ as stopping criterion and $N = 160$. The results for the case where the auxiliary problem is solved approximately in every (ALM)-iteration are summarized in Table 6.7, where we again report on the number of (PDAS)-iterations and on d_λ^l as defined in (6.1). Again increasing γ as in Example 6.2 turns out to be very efficient.

6.4. Summary of the numerical results. In our numerical testing we observe a remarkable efficiency of algorithms (SS) and (PDAS) for the solution of the regularized simplified friction problem (P_γ^*) and of (ALM) for the solution of (P^*). For moderate values of γ , the iterates for (SS) and (PDAS) coincide and these algorithms converge superlinearly. For large values of γ , (SS) may start to chatter, while (PDAS) always detects the solution. The two tested globalization strategies for (SS) turn out to be successful but inconvenient. The number of iterations of the semismooth Newton methods increases only slightly for finer grid and larger regularization parameter. The efficiency of (SS) and (PDAS) is interesting also with respect to augmented Lagrangian methods since these algorithms present a powerful tool to solve or approximately solve the auxiliary problem in (ALM). Our tests show that, solving the auxiliary problem in (ALM) only approximately using (SS) or (PDAS), the overall number of system solves for (ALM) is rather the same as for the semismooth Newton methods with large γ . However, (ALM) has the advantage that it detects the solution of the dual (P^*) of the original simplified friction problem without requiring that $\gamma \rightarrow \infty$. Finally, we remark the advantage of (PDAS) with large γ that one has a simple stopping criterion at hand that guarantees that the exact solution of (P_γ^*) is detected.

REFERENCES

- [1] M. BERGOUNIOUX, M. HADDOU, M. HINTERMÜLLER, AND K. KUNISCH, *A comparison of a Moreau–Yosida-based active set strategy and interior point methods for constrained optimal control problems*, SIAM J. Optim., 11 (2000), pp. 495–521.
- [2] M. BERGOUNIOUX, K. ITO, AND K. KUNISCH, *Primal-dual strategy for constrained optimal control problems*, SIAM J. Control Optim., 37 (1999), pp. 1176–1194.
- [3] D. P. BERTSEKAS, *Constraint Optimization and Lagrange Multipliers*, Academic Press, Paris, 1982.
- [4] X. CHEN, Z. NASHED, AND L. QI, *Smoothing methods and semismooth methods for nondifferentiable operator equations*, SIAM J. Numer. Anal., 38 (2000), pp. 1200–1216.
- [5] P. W. CHRISTENSEN, A. KLARBRING, J. S. PANG, AND N. STRÖMBERG, *Formulation and comparison of algorithms for frictional contact problems*, Internat. J. Numer. Methods Engrg., 42 (1998), pp. 145–173.

- [6] P. W. CHRISTENSEN AND J. S. PANG, *Frictional contact algorithms based on semismooth Newton methods*, in Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, Appl. Optim. 22, M. Fukushima and L. Qi, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999, pp. 81–116.
- [7] P. W. CHRISTENSEN, *A nonsmooth Newton method for elastoplastic problems*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 1189–1219.
- [8] G. DUVAUT AND J.-L. LIONS, *Inequalities in Mechanics and Physics*, Springer-Verlag, Berlin, 1976.
- [9] I. EKELAND AND R. TÉMAM, *Convex Analysis and Variational Problems*, Classics Appl. Math. 28, SIAM, Philadelphia, 1999.
- [10] F. FACCHINEI AND J.-S. PANG, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Vol. 1 and 2, Springer Ser. Oper. Res., Springer-Verlag, New York, 2003.
- [11] R. GLOWINSKI, J.-L. LIONS, AND R. TRÉMOLIÈRES, *Numerical Analysis of Variational Inequalities*, North-Holland, Amsterdam, 1981.
- [12] R. GLOWINSKI, *Numerical Methods for Nonlinear Variational Problems*, Springer-Verlag, New York, 1984.
- [13] W. HAN, *A regularization procedure for a simplified friction problem*, Math. Comput. Modelling, 15 (1991), pp. 65–70.
- [14] W. HAN, *On the numerical approximation of a frictional contact problem with normal compliance*, Numer. Funct. Anal. Optim., 17 (1996), pp. 307–321.
- [15] J. HASLINGER, Z. DOSTÁL, AND R. KUČERA, *On a splitting type algorithm for the numerical realization of contact problems with Coulomb friction*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 2261–2281.
- [16] M. HINTERMÜLLER, K. ITO, AND K. KUNISCH, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim., 13 (2003), pp. 865–888.
- [17] M. HINTERMÜLLER AND G. STADLER, *A semi-smooth Newton method for constrained linear-quadratic control problems*, ZAMM Z. Angew. Math. Mech., 83 (2003), pp. 219–237.
- [18] M. HINTERMÜLLER, *A primal-dual active set algorithm for bilaterally control constrained optimal control problems*, Quart. Appl. Math., 61 (2003), pp. 131–160.
- [19] I. HLAVÁČEK, J. HASLINGER, J. NEČAS, AND J. LOVÍŠEK, *Solution of Variational Inequalities in Mechanics*, Appl. Math. Sci. 66, Springer-Verlag, New York, 1988.
- [20] H. C. HUANG, W. HAN, AND J. S. ZHOU, *The regularization method for an obstacle problem*, Numer. Math., 69 (1994), pp. 155–166.
- [21] K. ITO AND K. KUNISCH, *Augmented Lagrangian methods for nonsmooth, convex optimization in Hilbert spaces*, Nonlinear Anal., 41 (2000), pp. 591–616.
- [22] K. ITO AND K. KUNISCH, *Semi-smooth Newton methods for variational inequalities of the first kind*, M2AN Math. Model. Numer. Anal., 37 (2003), pp. 41–62.
- [23] K. ITO AND K. KUNISCH, *The primal-dual active set method for nonlinear optimal control problems with bilateral constraints*, SIAM J. Control Optim., 43 (2004), pp. 357–376.
- [24] N. KIKUCHI AND J. T. ODEN, *Contact Problems in Elasticity: A Study of Variational Inequalities and Finite Element Methods*, SIAM Stud. Appl. Math. 8, SIAM, Philadelphia, 1988.
- [25] R. KORNUBER, *Adaptive Monotone Multigrid Methods for Nonlinear Variational Problems*, Adv. Numer. Math., Teubner, Stuttgart, 1997.
- [26] R. KORNUBER, *Monotone multigrid methods for elliptic variational inequalities. II*, Numer. Math., 72 (1996), pp. 481–499.
- [27] B. KUMMER, *Generalized Newton and NCP-methods: Convergence, regularity, actions*, Discuss. Math. Differ. Incl. Control Optim., 20 (2000), pp. 209–244.
- [28] C.-Y. LEE AND J. T. ODEN, *A priori error estimation of hp-finite element approximations of frictional contact problems with normal compliance*, Internat. J. Engrg. Sci., 31 (1993), pp. 927–952.
- [29] A. Y. T. LEUNG, G. CHEN, AND W. CHEN, *Smoothing Newton method for solving two- and three-dimensional frictional contact problems*, Internat. J. Numer. Methods Engrg., 41 (1998), pp. 1001–1027.
- [30] L. QI AND J. SUN, *A nonsmooth version of Newton's method*, Math. Programming, 58 (1993), pp. 353–367.
- [31] L. QI, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res., 18 (1993), pp. 227–244.
- [32] G. SAVARÉ, *Regularity and perturbation results for mixed second order elliptic problems*, Comm. Partial Differential Equations, 22 (1997), pp. 869–899.
- [33] M. ULBRICH, *Semismooth Newton methods for operator equations in function spaces*, SIAM J. Optim., 13 (2003), pp. 805–841.

SUBSET ALGEBRA LIFT OPERATORS FOR 0-1 INTEGER PROGRAMMING*

DANIEL BIENSTOCK[†] AND MARK ZUCKERBERG[†]

Abstract. We extend the Sherali–Adams, Lovász–Schrijver, Balas–Ceria–Cornuéjols, and Lasserre lift-and-project methods for 0-1 optimization by considering liftings to subset algebras. Our methods yield polynomial-time algorithms for solving a relaxation of a set-covering problem at least as strong as that given by the set of all valid inequalities with small coefficients and, more generally, all valid inequalities where the right-hand side is not very large relative to the positive coefficients in the left-hand side. Applied to generalizations of vertex-packing problems, our methods yield, in polynomial time, relaxations that have unbounded rank using, for example, the N_+ operator.

Key words. integer programming, lift-and-project, disjunctive methods

AMS subject classifications. 90C10, 90C27

DOI. 10.1137/S1052623402420346

1. Introduction. Consider a 0-1 integer programming problem

$$\min\{c^T x : x \in \mathcal{F}\},$$

where

$$(1) \quad \mathcal{F} = \{x \in \{0, 1\}^n : Ax \geq b\}.$$

The procedures in [29], [23], [22], and [3] solve this problem by iteratively strengthening its continuous relaxation, until after at most n iterations, the convex hull of $\mathcal{F}$ is obtained. This bound on the number of iterations is tight ([6], [24]; see also [19] and [15] for related topics). Nevertheless, a question of theoretical and practical interest is whether it is possible to modify the procedures so that the earlier iterations produce stronger relaxations.

As shown in [24], the methods used in [29], [23], [22], and [3] can be viewed as relying on a common paradigm: that of “lifting” a point in $\{0, 1\}^n$ to an appropriate zeta-vector of the subset lattice of $\{1, 2, \dots, n\}$.

In this paper we introduce operators that lift instead to the (much larger) subset algebra of $\{0, 1\}^n$. One example of a result which is derived using our operators is the following.

THEOREM 1.1. *Let $k \geq 1$ be a fixed integer. Consider a set-covering problem*

$$\min\{c^T x : Ax \geq e, x \in \{0, 1\}^n\},$$

where A is 0-1 and e is a vector of 1s. Let V_k denote the set of inequalities $a^T x \geq a_0$ which are valid for $\{x \in \{0, 1\}^n : Ax \geq e\}$ and for which $a_j \in \{0, 1, 2, \dots, k\}$, $0 \leq j \leq n$. Let $\mathcal{V}_k$ denote the set of points in $\{0, 1\}^n$ that satisfy all inequalities in V_k . Then there is an algorithm of complexity polynomial in n for solving

$$\begin{aligned} \min c^T x \\ \text{s.t. } x \in \mathcal{R}_k, \end{aligned}$$

*Received by the editors December 20, 2002; accepted for publication (in revised form) December 4, 2003; published electronically October 1, 2004. This research was partially funded by NSF awards 29132-5538 (CORC) and DMI-0200221.

<http://www.siam.org/journals/siopt/15-1/42034.html>

[†]Columbia University, New York, NY 10027 (dano@ieor.columbia.edu).

where $\mathcal{R}_k$ is a certain polyhedron satisfying

$$\{x \in \{0, 1\}^n : Ax \geq e\} \subseteq \mathcal{R}_k \subseteq \text{conv}(\mathcal{V}_k).$$

In other words, $\mathcal{R}_k$ is a polyhedral relaxation of the set-covering problem, all of whose points satisfy all inequalities in the set $\mathcal{V}_k$.

Theorem 1.1 is a special case of a more general result. Given an inequality $a^T x \geq a_0$ with indices ordered so that $0 < a_1 \leq a_2 \leq \dots \leq a_J$ and $a_j = 0$ for $j > J$, its *pitch* is the minimum t such that $\sum_{j=1}^t a_j \geq a_0$. Then we have the following.

THEOREM 1.2. *Let $k \geq 1$ be a fixed integer. Consider a set-covering problem*

$$\min\{c^T x : Ax \geq e, x \in \{0, 1\}^n\},$$

where A is an $m \times n$, 0-1 matrix and e is the vector of m 1s. Let P_k denote the set of all valid inequalities for $\{x \in \{0, 1\}^n : Ax \geq e\}$ of pitch $\leq k$. Then there exists a positive integer $g(k)$, a polytope $\mathcal{Q}_k \subseteq R^n$, and a polytope $\bar{\mathcal{Q}}_k \subseteq R^{(m+n)^{g(k)}}$ satisfying the following:

- (a) $\{x \in \{0, 1\}^n : Ax \geq e\} \subseteq \mathcal{Q}_k$.
- (b) $a^T x \geq a_0$ for all $x \in \mathcal{Q}_k$ and for all $(a, a_0) \in P_k$.
- (c) $\mathcal{Q}_k$ is the projection to R^n of $\bar{\mathcal{Q}}_k$.
- (d) $\bar{\mathcal{Q}}_k$ can be described by a system of at most $(m+n)^{g(k)}$ linear constraints, with integral coefficients of absolute value at most k . This system can be computed in time polynomial in n and m for fixed k .

Thus, Theorem 1.2 introduces a natural hierarchy among the valid inequalities for set-covering problems and shows that any fixed level in the hierarchy can be satisfied in polynomial time. How strong are the pitch $\leq k$ inequalities? First, note that the hierarchy is “complete” in the sense that any facet-defining inequality has pitch $\leq n$. Furthermore, one can produce examples of set-covering problems with exponentially many facets with coefficients 0, 1, 2 (Balas and Ng have completely characterized the set of facet-defining inequalities with coefficients 0, 1, 2 [5]). Also, there are examples of Chvátal–Gomory inequalities of rank greater than one with coefficients ≤ 3 , and we conjecture that for each fixed integer $k' > 0$ there is a $k = k(k')$ such that there are arbitrarily large examples of set-covering problems, where some of the inequalities of pitch $\leq k$ have (fractional) Gomory rank $\geq k'$.

At the same time, we show that given a set-covering problem with a full-circulant constraint matrix ($\sum_{j \neq i} x_j \geq 1$ for each $1 \leq i \leq n$) the valid inequality $\sum_j x_j \geq 2$ (a constraint of pitch 2) has rank at least $n - 3$ for a lifting operator stronger than the Sherali–Adams and the N_+ procedures combined. In other words (if somewhat incompletely), lifting to a polynomially larger space yields an operator that is exponentially stronger (for a proof see [10]). These results, and others on set-covering problems, are presented in sections 3.2 and 4.5.

The hierarchy of algorithms that we present (for fixed k we call our level k procedure the Σ^k -algorithm, and we define these for $k \geq 2$) do not require positive-semidefiniteness in order to achieve Theorem 1.2. On the other hand, one of the key results in [23] is that positive-semidefiniteness allows the N_+ operator to guarantee that classical inequalities, for example clique inequalities in vertex-packing problems, are achieved with rank 1. In section 4.6 we consider set-packing problems where, given a graph G and a family of pairwise disjoint index sets S_i , $i \in V(G)$, for every edge $\{i, j\}$, we have the constraint

$$(2) \quad \sum_{k \in S_i} x_k + \sum_{k \in S_j} x_k \leq |S_i| + |S_j| - 1.$$

When $|S_i| = 1$ for each i , we obtain precisely the vertex-packing polyhedron of G . Thus, one can think of generalizations of several classical valid inequalities (clique, odd-hole, odd-antihole, odd-wheel). For example, when G contains a clique with vertex set K the resulting *set-clique* constraint is

$$\sum_{i \in K} \sum_{k \in S_i} x_k \leq \sum_{i \in K} |S_i| - |K| + 1.$$

One can similarly define set-odd-hole, set-odd-antihole, and set-odd-wheel inequalities. In section 4.6 we present the following result.

THEOREM 1.3. *Suppose we have a formulation that contains inequalities (2) corresponding to a clique K (with pairwise disjoint sets S_i). Then the Σ^2 -algorithm generates a vector that satisfies the set-clique constraint. On the other hand, set-clique inequalities have unbounded N_+ - (and Sherali–Adams) rank. Similarly, the Σ^2 -algorithm (resp., the Σ^3 -algorithm) generates a vector that satisfies all set-odd-hole (resp., all set-odd-antihole and set-odd-wheel) constraints.*

Finally, section 4.7 presents further comparisons between our algorithms and the N_+ procedure. The development of our algorithms starts in section 3.

2. Background. Let $\mathcal{F} \subseteq \{0, 1\}^n$ be as before. Lovász and Schrijver [23] and Sherali and Adams [29] introduced the following general paradigm for the problem of separating over $\text{conv}(\mathcal{F})$.

Let $N \gg n$, and suppose we have a function that maps (“lifts”) each $v \in \mathcal{F}$ into $z = z(v) \in \{0, 1\}^N$ with $z_j = v_j$, $1 \leq j \leq n$. Let $\hat{\mathcal{F}} \subseteq \{0, 1\}^N$ denote the image of $\mathcal{F}$ under this mapping. Then given $x \in R_+^n$, the question of whether $x \in \text{conv}(\mathcal{F})$ is equivalent to answering the question of whether there exists $y \in \text{conv}(\hat{\mathcal{F}})$ such that $y_i = x_i$, $1 \leq i \leq n$.

This second membership question may be easier to answer than the original one because the vectors $z(v)$ reveal information about $\mathcal{F}$ in a more explicit way. As pointed out in [23], this basic idea was implicit in earlier work on specific combinatorial problems; see [7], [9], [8], and others. Furthermore, as it turns out, the much earlier work of Balas on disjunctive programming [1], [2] provides a common underlying viewpoint for much of this work.

Writing $E_n = \{1, 2, \dots, n\}$, the concrete application of this idea in [23] and [29] is as follows. We map each $v \in \{0, 1\}^n$ into $\hat{v} \in \{0, 1\}^{2^n}$, where

- (i) the entries of $\hat{v}$ are indexed by subsets of E_n , and
- (ii) for $S \subseteq \{1, \dots, n\}$, $\hat{v}_S = 1$ iff $v_j = 1$ for all $j \in S$.

Clearly, for $1 \leq j \leq n$, $\hat{v}_{\{j\}} = v_j$, and consequently $\mathcal{F}$ is mapped (lifted) into a set $\hat{\mathcal{F}} \subseteq \{0, 1\}^{2^n}$ whose projection to R^n is precisely $\mathcal{F}$. As a result, the separation problem for $\text{conv}(\mathcal{F})$ can be cast as a separation problem over $\text{conv}(\hat{\mathcal{F}})$.

The lifting can be seen to map each $v \in \{0, 1\}^n$ into a distinct column of the zeta matrix $\mathcal{Z}$ of the subset lattice L of E_n (see [28]). Using this observation, one can obtain a complete polyhedral characterization of $\text{conv}(\hat{\mathcal{F}})$ —however, this is not a computationally practicable approach, as it involves an exponential number of variables.

Instead, [29] (and, implicitly, [23]) considers restrictions of the above lifting to “small” subsets: for every integer $t > 1$, one can restrict the lifting to just those coordinates corresponding to the cardinality- t subsets. For fixed t , this gives a polynomially large formulation, thereby producing a hierarchy of relaxations.

A critical ingredient in deriving these formulations is the technique of *symmetrization*. If $\hat{v}$ is the lifting of our 0-1 vector v to the appropriate space of variables, then

the matrix $\hat{v}\hat{v}^T$ is 0-1, symmetric, and positive-semidefinite. Furthermore, each column of the matrix corresponding to a “1” entry in $\hat{v}$ consists of a copy of $\hat{v}$, and every other column has all entries equal to 0. These conditions can be used to derive constraints—in fact, in [23] and [29] one lifts a vector $x \in R^n$ to a matrix M on which appropriate constraints are imposed.

For lack of space, we cannot provide here a comprehensive review of the work in [23] and [29]. The reader is instead referred to [10]. The symmetrization technique will also be used in our algorithms described below. See also [24] for an in-depth comparison of the various lift operators, including Lasserre’s [22].

3. New results. The lifting operators we will describe below map any point of $\mathcal{F}$ to a zeta-vector of the subset algebra of $\{0, 1\}^n$, viewed as a lattice. A completely formal definition of a subset algebra is beyond the scope of this paper (see [14]), but the following should suffice.

Suppose S is a finite set and $\{A_j, 1 \leq j \leq k\}$ is a collection of subsets of S . Let $\bar{A}_j = S - A_j$ for each j .

DEFINITION 3.1. *The subset algebra $\Sigma = \Sigma(A_1, \dots, A_k)$ generated by $A_1, \dots, A_k$ is the set of all subsets of S that can be obtained from set-theoretic expressions involving the $A_1, \dots, A_k$ and $\bar{A}_1, \dots, \bar{A}_k$.*

Thus, expressions of the form $A_1 \cup (A_2 \cap \bar{A}_3)$, etc. are in $\Sigma(A_1, \dots, A_k)$. Also, note that $\Sigma(A_1, \dots, A_k)$ can be viewed as contained in 2^S .

DEFINITION 3.2. *Let $J \subseteq \{1, 2, \dots, k\}$. The subset*

$$\left(\bigcap_{j \in J} A_j \right) \cap \left(\bigcap_{j \notin J} \bar{A}_j \right)$$

is called an atom (also called a Boolean function or a complete product).

It can be shown that every element of Σ can be written as a (finite) union of atoms. Thus, Σ has at most 2^{2^k} distinct elements.

EXAMPLE 3.3. *Consider $S = \{0, 1\}^n$. For each $1 \leq j \leq n$, let $H_j \doteq \{x \in \{0, 1\}^n : x_j = 1\}$. Then the subset algebra generated by the H_j is exactly the set of all subsets of $\{0, 1\}^n$. Each atom corresponds to a distinct point in $\{0, 1\}^n$, and there are 2^{2^n} elements in the algebra.*

Σ is also a lattice. In our case, we will use the reverse of the inclusion order; i.e., we will declare $b \leq a$ when $a \subseteq b$. That we have a lattice follows since Σ is closed under unions and intersections, and hence $\vee$ and $\wedge$ are well defined. For convenience, we will also denote this lattice by Σ . To avoid confusion with the standard subset lattice, we will use ξ to denote a zeta-vector of this lattice. Thus, if $p \in \Sigma$, the ξ -vector for p is the vector $\xi \in \{0, 1\}^\Sigma$ defined by

$$(3) \quad \xi_q^p = \begin{cases} 1 & \text{if } p \subseteq q, \\ 0 & \text{otherwise,} \end{cases}$$

where we use “ $\subseteq$ ” whenever the underlying subsets of S satisfy the relation. In our development of lift operators, we will work with (slight) generalizations of this type of lattice, as indicated next.

As a final note, when Σ has a member for each subset of S , then, as a lattice, Σ is isomorphic to the subset lattice of S (with the order relationship reversed). Even in such a case, however, it is useful to view the lattice as generated by the A_j .

3.1. Lifting to an algebra. We now return to the feasible region (1) of a 0-1 integer program, repeated here for convenience:

$$\mathcal{F} = \{x \in \{0, 1\}^n : Ax \geq b\}.$$

As before, we will use the following notation. For $1 \leq j \leq n$, denote $H_j \doteq \{x \in \{0, 1\}^n : x_j = 1\}$, and

$$(4) \quad Y_j = \mathcal{F} \cap H_j, \quad N_j = \bar{Y}_j = \mathcal{F} - Y_j.$$

In other words, Y_j (resp., N_j) is the subset of $\mathcal{F}$ with $x_j = 1$ (resp., $x_j = 0$). Let Σ denote the subset algebra $\Sigma(Y_1, \dots, Y_n)$. Finally, if $X \in R^\Sigma$, its α -entry (for $\alpha \in \Sigma$) is indicated by $X[\alpha]$, and for a matrix $B \in R^{\Sigma \times \Sigma}$, its α, β -entry is indicated by $B[\alpha, \beta]$.

Our lifting will map points in $\mathcal{F}$ to zeta-vectors arising from Σ . Formally, this is done as follows. Given $x \in \mathcal{F}$, denote by $\alpha(x) \in \Sigma$ the atom

$$(5) \quad \left(\bigcap_{j: x_j=1} Y_j \right) \cap \left(\bigcap_{j: x_j=0} N_j \right).$$

Note that the set-theoretic value of $\alpha(x)$ is precisely x ; i.e., $\alpha(x)$ belongs in the equivalence class defined by x . Our lifting maps

$$(6) \quad x \rightarrow \xi^{\alpha(x)}.$$

An important point to notice is that only a (very small) subset of all zeta-vectors are images under this lifting—in this, our approach differs from those outlined in section 1. Furthermore, even though the dimension of the image space is very large (doubly exponential in n), in our algorithms we will only consider polynomially many coordinates. Also note that the lifting (6) could be applied to any point $x \in \{0, 1\}^n$, not just $x \in \mathcal{F}$. However, below we will impose additional conditions on the lifting, conditions that are not guaranteed to be satisfied by points $x \in \{0, 1\}^n \setminus \mathcal{F}$, and therefore, in general, the lifting will not even exist unless $x \in \mathcal{F}$. This is fine for our purposes: we are interested in generating a relaxation that is as tight as possible. Similarly, we could have defined the target of our lifting to be the subset algebra of $\mathcal{F}$, not $\{0, 1\}^n$. Here the difference is completely semantic: if $x \in \mathcal{F}$, then, when $S \subseteq \{0, 1\}^n \setminus \mathcal{F}$, we have $\xi^{\alpha(x)}[S] = 0$, and when $S \cap \mathcal{F} \neq \emptyset$, we have $\xi^{\alpha(x)}[S] = \xi^{\alpha(x)}[S \cap \mathcal{F}]$.

The following properties of the lifting are easy to verify:

- (i) $\xi^{\alpha(x)}_\emptyset = 0$.
- (ii) $\xi^{\alpha(x)}_{\mathcal{F}} = 1$.
- (iii) For $1 \leq j \leq n$,

$$(7) \quad \xi^{\alpha(x)}[Y_j] = x_j \quad \text{and} \quad \xi^{\alpha(x)}[N_j] = 1 - x_j.$$

- (iv) More generally, suppose we have a collection $\{\beta^i \in \Sigma : i \in I\}$, corresponding to pairwise disjoint subsets of $\mathcal{F}$; i.e., each β^i corresponds to a $\mathcal{J}^i \subseteq \mathcal{F}$ such that $\mathcal{J}^i \cap \mathcal{J}^j = \emptyset$ for all distinct $i, j \in I$. Then

$$(8) \quad \sum_{i \in I} \xi^{\alpha(x)}[\beta^i] \leq 1,$$

with equality when the $\mathcal{J}^i$ form a partition of $\mathcal{F}$.

Property (iv) follows because the point x cannot belong to more than one $\mathcal{J}^i$.

One can also prove results analogous to those outlined in section 1. In particular, define $\hat{\mathcal{F}} \subseteq \{0, 1\}^\Sigma$ to be the image of $\mathcal{F}$ under our lifting. The following are straightforward results that formalize the above.

LEMMA 3.4. *Let $X \in \text{conv}(\hat{\mathcal{F}})$. Suppose we have a collection $\{\beta^i \in \Sigma : i \in I\}$, corresponding to pairwise disjoint subsets of $\mathcal{F}$. Then*

$$(9) \quad \sum_{i \in I} X[\beta^i] \leq 1$$

with equality when $\{\beta^i : i \in I\}$ corresponds to a partition of $\mathcal{F}$. Furthermore,

$$(10) \quad X \geq 0, \quad X[\emptyset] = 0, \quad \text{and} \quad X[\mathcal{F}] = 1.$$

LEMMA 3.5. *Let $x \in R^n$. Then $x \in \text{conv}(\mathcal{F})$ iff there exists a vector $X \in \text{conv}(\hat{\mathcal{F}})$ such that*

$$(11) \quad X[Y_j] = x_j, X[N_j] = 1 - x_j \quad \forall 1 \leq j \leq n.$$

The main question we want to address is what conditions the vector X in Lemma 3.5 can be required to satisfy, in addition to (9)–(11).

To this effect, consider an arbitrary vector $X \in R^\Sigma$ of the form

$$(12) \quad X = \sum_{\alpha \in \Sigma} \lambda_\alpha \xi^\alpha,$$

where $\alpha \in R^\Sigma$ is nonnegative. Define the matrix

$$(13) \quad U^X = \sum_{\alpha \in \Sigma} \lambda_\alpha \xi^\alpha (\xi^\alpha)^T.$$

We have the following basic facts.

LEMMA 3.6. *Suppose λ , X , and U^X are as in (12)–(13). Let $\beta, \gamma \in \Sigma$. Then*

$$(14) \quad U_{\beta, \gamma}^X = \sum_{\alpha \subseteq \beta \cap \gamma} \lambda_\alpha.$$

Proof. Let $\alpha \in \Sigma$. Then the sum in (13) contributes either zero or λ_α to $U_{\beta, \gamma}^X$, and the latter happens when both $\xi^\alpha[\beta] = 1$ and $\xi^\alpha[\gamma] = 1$. By the definition of ξ^α , this is true exactly when $\alpha \subseteq \beta$ and $\alpha \subseteq \gamma$, i.e., $\alpha \subseteq \beta \cap \gamma$, as desired. $\square$

COROLLARY 3.7. *For any $\beta, \gamma \in \Sigma$,*

$$(15) \quad U_{\beta, \gamma}^X = X[\beta \cap \gamma].$$

This equation embodies an extension to the subset algebra lifting of a technique used in [23], [29]. As a consequence of these results, we have the following.

LEMMA 3.8. *Suppose λ , X , and U^X are as in (12)–(13). Then*

- (i) U^X is symmetric and positive-semidefinite;
- (ii) the main diagonal of U^X and its $\mathcal{F}$ -row and $\mathcal{F}$ -column are all equal to X . $\square$

Note that in Lemma 3.8, symmetricity and (ii) are really very weak consequences of (15), which implies that many pairs of entries in U^X are equal. For $\alpha \in \Sigma$, let $e_\alpha \in R^\Sigma$ be the vector with a 1 in position α and zero otherwise.

LEMMA 3.9. Suppose λ , X , and U^X are as in (12)–(13). Consider a vector $\eta \in R^\Sigma$ such that

$$(16) \quad \eta^T \xi^\alpha \geq 0 \quad \forall \alpha \text{ with } \lambda_\alpha > 0.$$

Then

- (i) $\eta^T U^X e_\beta \geq 0$ for all $\beta \in \Sigma$.
- (ii) $\eta^T U^X (e_{\mathcal{F}} - e_\beta) \geq 0$ for all $\beta \in \Sigma$.
- (iii) For every $v \in R^\Sigma$, $v^T U^\kappa v \geq 0$, where $\kappa = U^X \eta$.

Proof. (i) Let $\alpha \in \Sigma$ with $\lambda_\alpha > 0$. Then since $\xi^\alpha (\xi^\alpha)^T e_\beta$ is either identically zero (when $\xi^\alpha[\beta] = 0$) or else it equals ξ^α , we have $\eta^T \xi^\alpha (\xi^\alpha)^T e_\beta \geq 0$, and the result follows.

(ii) Let $\bar{\beta} \in \Sigma$ denote a negation of β (i.e., an element of Σ whose set-theoretic value is that of the complement of β). Then the $\mathcal{F}$ -column of U^X is equal to the sum of the β -column and the $\bar{\beta}$ -column of U^X ; i.e.,

$$(17) \quad U^X e_{\mathcal{F}} = U^X (e_\beta + e_{\bar{\beta}}),$$

because (similarly to the proof of (i))

$$(18) \quad \xi^\alpha (\xi^\alpha)^T e_{\mathcal{F}} = \xi^\alpha (\xi^\alpha)^T (e_\beta + e_{\bar{\beta}}) \quad \forall \alpha \in \Sigma;$$

see (8). The result follows from (17) by applying (i) to $\bar{\beta}$.

(iii) This follows because $\kappa \in R^\Sigma$ satisfies

$$(19) \quad \kappa = \sum_{\lambda_\alpha > 0} \lambda_\alpha \xi^\alpha (\xi^\alpha)^T \eta$$

$$(20) \quad = \sum_{\gamma_\alpha > 0} \gamma_\alpha \xi^\alpha,$$

where $\gamma_\alpha = \lambda_\alpha \eta^T \xi^\alpha \geq 0$. Thus,

$$U^\kappa = \sum_{\gamma_\alpha > 0} \gamma_\alpha \xi^\alpha (\xi^\alpha)^T$$

from which (iii) follows. $\square$

The properties of U^X given in Lemma 3.8 and Lemma 3.9(i) and (ii) parallel the Lovász–Schrijver development of their operators. Property (iii) in Lemma 3.9 is similar to the Sherali–Adams and Lasserre development as shown in [24].

We summarize what we know so far.

LEMMA 3.10. Let $x \in \text{conv}(\mathcal{F})$. Then there is a vector $X \in R^\Sigma$ such that

1. equations (9)–(11) are satisfied;
2. there is a matrix $U^X \in R^{\Sigma \times \Sigma}$ satisfying (15), conditions (i)–(ii) of Lemma 3.8, and conditions (i)–(iii) of Lemma 3.9.

In particular, consider (ii) of Lemma 3.9. One way to use this condition is to start with any inequality $a^T x \geq a_0$ which is valid for $\mathcal{F}$ and from it obtain (as in section 1) an inequality $\hat{a}^T X \geq 0$ which is valid for $\hat{\mathcal{F}}$. (Formally, set $\hat{a}[Y_j] = a_j$ for $1 \leq j \leq n$, $\hat{a}[\mathcal{F}] = -a_0$, and $\hat{a}[\alpha] = 0$ for all other $\alpha \in \Sigma$.)

In addition, there are measure-theoretic valid inequalities that one can use. For example, for $\alpha, \beta \in \Sigma$,

$$(21) \quad X[\alpha \cap \beta] \leq \min(X[\alpha], X[\beta]),$$

$$(22) \quad X[\alpha \cup \beta] = X[\alpha] + X[\beta] - X[\alpha \cap \beta]$$

are valid inequalities. Similar remarks can be made concerning property (iii) in Lemma 3.9.

3.1.1. Variable replication. We need to describe an additional feature that is required only to formally guarantee the validity of our algorithms. The need for this additional machinery is best explained with an example. In applying the lemmas described above, our algorithms will restrict X and U^X to polynomial-sized subsets of Σ . As a simple example, we might generate a symbol σ_1 for the expression $((Y_1 \cap N_2) \cup (N_1 \cap N_2)) \cap N_3$ and use it as one of the variable indices, that is, impose the (valid) constraint

$$X[\sigma_1] = X[Y_1 \cap N_2 \cap N_3] + X[N_1 \cap N_2 \cap N_3].$$

At the same time, we might create a symbol σ_2 for the expression $N_2 \cap N_3$ and use it in constraints, say

$$X[\sigma_2] \leq \min\{X[N_2], X[N_3]\}.$$

The algorithm certainly *would* benefit by imposing the valid requirement

$$(23) \quad X[\sigma_1] = X[\sigma_2]$$

(or, indeed, by using a unique variable)—but the algorithm can only do this if it knows that

$$(Y_1 \cap N_2) \cup (N_1 \cap N_2) = N_2,$$

and the algorithm can make this sort of deduction only by engaging in the symbolic algebra needed to determine this fact. If the algorithm does not impose (23), then we end up using a relaxation, i.e., a weaker formulation than theoretically possible. Sometimes this will be the case with our algorithms (in particular with regard to certain complex set-theoretic expressions) with the result that we end up producing duplicate symbols.

These ideas are formalized as follows.

DEFINITION 3.11. *Let Q be a set. A symbol function is a function $\mathcal{S} : \Sigma \rightarrow 2^Q$ with the following properties:*

- (S.1) *For each $\alpha, \beta \in \Sigma$ with $\alpha \neq \beta$, we have $\mathcal{S}(\alpha) \cap \mathcal{S}(\beta) = \emptyset$.*
- (S.2) *$\mathcal{S}(\emptyset)$ and $\mathcal{S}(\mathcal{F})$ are nonempty.*

For $\alpha \in \Sigma$, the elements of $\mathcal{S}(\alpha)$ are called the symbols associated with α .

Here, the set Q is arbitrary (for example, in a practical implementation we might have $Q = Z_+$). Let $\mathcal{S}(\Sigma)$ denote the union of all the sets $\mathcal{S}(\alpha)$, $\alpha \in \Sigma$. For consistency, we will use the notation $x[i]$ to refer to the i th entry of vector $x \in R^{\mathcal{S}(\Sigma)}$, for $i \in \mathcal{S}(\Sigma)$, and similarly we will use the notation $M[i, j]$ to refer to the i, j entry in a matrix $M \in R^{\mathcal{S}(\Sigma) \times \mathcal{S}(\Sigma)}$.

LEMMA 3.12. *Suppose $\mathcal{S}$ is a symbol function. Let $\mathcal{M} \subseteq \Sigma$ be such that $\emptyset \in \mathcal{M}$, $\mathcal{F} \in \mathcal{M}$, and $\mathcal{S}(\alpha)$ is nonempty for each $\alpha \in \mathcal{M}$. Let $U \in R^{\mathcal{M} \times \mathcal{M}}$. Consider the matrix*

$$\check{U} \in R^{\mathcal{S}(\mathcal{M}) \times \mathcal{S}(\mathcal{M})}$$

defined as follows: for any $\alpha, \beta \in \mathcal{M}$, any $i \in \mathcal{S}(\alpha)$, and $j \in \mathcal{S}(\beta)$, $\check{U}[i, j] = U[\alpha, \beta]$. Then we have the following:

- (a) *If U is symmetric positive-semidefinite, so is $\check{U}$.*
- (b) *Let $\eta \in R^{\mathcal{M}}$ be such that $\eta^T U \geq 0$. Suppose we choose, for each $\alpha \in \mathcal{M}$, a particular element $j_\alpha \in \mathcal{S}(\alpha)$. Then for each column $\check{u}$ of $\check{U}$ we have*

$$\sum_{\alpha \in \mathcal{M}} \eta[\alpha] \check{u}[j_\alpha] \geq 0. \quad \square$$

The proof of this lemma is elementary. The significance of this lemma is that it allows our lifting to create multiple indices that correspond to the same subset of the algebra—though our algorithm may not be aware that these are duplicates of one another. At the same time, when we lift to a matrix we can impose on this matrix all the conditions discussed above (for example, the diagonal is equal to the $\mathcal{F}$ -row and $\mathcal{F}$ -column).

To summarize this section, our algorithms in general refrain from performing the manipulations needed to determine when two set-theoretic expressions are equivalent. This is done to avoid the exponential amount of work that such a certification would sometimes require. On the other hand, some equivalences are easy to check. As we will see, *any* of the expressions considered by our algorithms is explicitly of the form $B_1 \cap B_2 \cap \cdots \cap B_r$, where each B_i is of the form Y_j or N_j (for some j), or is $\mathcal{F}$, or belongs to a polynomial-size class of additional symbols. We will assume that permuting the B_i produces an equivalent expression. This is a requirement that is easy to enforce in polynomial time. Later we will discuss another simplification which is equally easy to enforce.

3.2. Example—a simple algorithm for set-covering. As a prelude to the algorithms we will describe later, here we present a simple algorithm specialized for set-covering problems that achieves provable results, which is a special case of a general algorithm to be described in section 4. In particular, we will show that we obtain a polynomial-time algorithm for optimizing over a relaxation at least as strong as that given by the convex hull of all valid inequalities with coefficient 0, 1, or 2.

Thus, let A denote an $m \times n$ 0-1 matrix, and consider the feasible region $\mathcal{F}$ for a 0-1 set-covering problem,

$$(24) \quad Ax \geq e,$$

$$(25) \quad x \in \{0, 1\}^n,$$

where we assume that no row of A contains another. We denote by $A_i \subseteq \{1, \dots, n\}$ the set of indices of nonzeros in the i th row of A , $1 \leq i \leq m$. After presenting the algorithm (in a somewhat redundant form) we discuss its behavior.

ALGORITHM C.

Step 0. Create the variable $X[\mathcal{F}]$, and for $1 \leq j \leq n$ the variables $X[Y_j]$ and $X[N_j]$, and impose the constraint

$$(26) \quad X[Y_j] + X[N_j] - X[\mathcal{F}] = 0.$$

Step 1. For each $1 \leq i \leq m$ impose the constraint

$$(27) \quad \sum_{j \in A_i} X[Y_j] - X[\mathcal{F}] \geq 0.$$

Step 2. For each unordered pair of indices $i \neq h$, $1 \leq i, h \leq m$, where

$$C^{i,h} \doteq A_i \cap A_h,$$

if $C^{i,h} \neq \emptyset$, we do the following, provided that $C^{i,h}$ has not already been enumerated as the set $C^{\bar{i},\bar{h}}$ for some other pair $\{\bar{i}, \bar{h}\}$.

(2.a) Create the variable $X[\bigcap_{j \in C^{i,h}} N_j]$ and impose the constraints

$$(28) \quad X[N_r] - X\left[\bigcap_{j \in C^{i,h}} N_j\right] \geq 0 \quad \forall r \in C^{i,h}.$$

- (2.b) For each $r \in C^{i,h}$, create the new variable $X[Y_r \cap \bigcap_{j \in C^{i,h-r}} N_j]$ and impose the constraint

$$(29) \quad X[Y_r] - X\left[Y_r \cap \bigcap_{j \in C^{i,h-r}} N_j\right] \geq 0.$$

- (2.c) If $|C^{i,h}| \geq 2$, create the new variable $x[\tau^{i,h}]$ and impose the constraint

$$(30) \quad \sum_{j \in C^{i,h}} X[Y_j] - 2X[\tau^{i,h}] \geq 0.$$

Here, the symbol $\tau^{i,h}$ is a symbol associated with

$$(31) \quad \bigcup_{t \geq 2} \left(\bigcup_{S \subseteq C^{i,h} : |S|=t} \left(\bigcap_{j \in S} Y_j \bigcap_{j \notin S} N_j \right) \right) \in \Sigma.$$

That is to say, $\tau^{i,h}$ represents the union of all expressions involving intersections of Y and N variables indexed by all elements of $C^{i,h}$, where at least two of the variables are Y 's.

- (2.d) Impose the constraint

$$(32) \quad X\left[\bigcap_{j \in C^{i,h}} N_j\right] + \sum_{r \in C^{i,h}} X\left[Y_r \cap \bigcap_{j \in C^{i,h-r}} N_j\right] + X[\tau^{i,h}] - X[\mathcal{F}] = 0,$$

where the $\tau^{i,h}$ term is only used in case $|C^{i,h}| \geq 2$.

Step 3. Let $\mathcal{V}$ denote the set of variable indices we have created so far, and let X denote the vector of variables. Create a matrix U of variables, with rows and columns indexed by $\mathcal{V}$, and impose on the variables in U the following constraints.

- (3.1) U is symmetric, $U[\mathcal{F}, \mathcal{F}] = X[\mathcal{F}]$, and the main diagonal, the $\mathcal{F}$ -row, and the $\mathcal{F}$ -column of U are all equal to X .
- (3.2) For each constraint $\eta^T X \geq 0$ of the form (26), (27), (28), (29), (30), (32), impose the constraints

$$(33) \quad \eta^T U \geq 0.$$

Step 4. Impose

$$(34) \quad 0 \leq U[\alpha, \beta] \leq U[\mathcal{F}, \beta] \quad \forall \alpha, \beta \in \mathcal{V},$$

$$(35) \quad X[\mathcal{F}] = 1.$$

End.

3.2.1. Analysis of Algorithm C. We will first show that Algorithm C provides a valid lifting for $\text{conv}(\mathcal{F})$. In what follows, let $\mathcal{M}$ denote the set of distinct members of Σ , i.e., subsets of $\{0, 1\}^n$, that arise as the set-theoretic value of variable indices $\mathcal{V}$ produced by the algorithm. Thus, $\mathcal{M}$ contains $\emptyset$, $\mathcal{F}$, all the Y_j and N_j , all the (distinct) $\bigcap_{j \in C^{i,h}} N_j$, all the (distinct) $Y_r \cap \bigcap_{j \in C^{i,h-r}} N_j$, and all the (distinct) sets of the form given by the left-hand side of (31). In the language of section 3.1.1, there is an implicit symbol function σ , such that $\sigma(\mathcal{M}) = \mathcal{V}$, that the algorithm has constructed.

THEOREM 3.13. *Suppose $\bar{x} \in \text{conv}(\mathcal{F})$. Then there exists a vector $\check{X}$ satisfying constraints (26), (27), (28), (29), (30), (32), (34), (35), and a matrix $\check{U}$ satisfying the conditions in Step 3, such that $\check{X}[Y_j] = \bar{x}_j$ for all $1 \leq j \leq n$.*

Proof. Since the constraints imposed by Algorithm C define a convex set, it suffices to show the result assuming that $\bar{x} \in \mathcal{F}$.

Thus, let $z \in \{0, 1\}^\Sigma$ be the lifting of $\bar{x}$ that satisfies the conditions in Lemma 3.10, and let U^z denote the corresponding matrix. We will show that z , restricted to $\mathcal{M}$, satisfies (26), (27), (28), (29), (30), (32), (34), (35) (with the variable indices interpreted to obtain their true set-theoretic value) and that U^z , restricted to $\mathcal{M} \times \mathcal{M}$, satisfies the conditions in Step 3. By appealing to Lemma 3.12, we obtain the desired result.

First, by Lemma 3.4, z satisfies (35). Next, consider constraint (32) applied to a given pair i, h . Now $\mathcal{F} \subseteq \{0, 1\}^n$ can be partitioned into $2 + |C^{i,h}|$ sets:

- The set containing those points in $\{0, 1\}^n$ where all coordinates in $C^{i,h}$ equal zero. This is the subset of $\{0, 1\}^n$ corresponding to $\bigcap_{j \in C^{i,h}} N_j$.
- The set containing those points in $\{0, 1\}^n$ where one coordinate $r \in C^{i,h}$ equals one and all other coordinates equal zero. This is the subset corresponding to $Y_r \cap \bigcap_{j \in C^{i,h}-r} N_j$.
- The set containing all remaining points in $\{0, 1\}^n$, i.e., points where at least two coordinates in $C^{i,h}$ equal one. This is the subset corresponding to $\tau^{i,h}$.

Consequently, by Lemma 3.4, z satisfies

$$(36) \quad z \left[\bigcap_{j \in C^{i,h}} N_j \right] + \sum_{r \in C^{i,h}} z \left[Y_r \cap \bigcap_{j \in C^{i,h}-r} N_j \right] + z[\tau^{i,h}] = 1,$$

and we obtain (32), since $z[\mathcal{F}] = 1$. Similarly, it follows that z and U satisfy all the other constraints. $\square$

Furthermore, we have the following.

LEMMA 3.14. *The total number of variables and constraints created by Algorithm C is $O(m^4 n^2)$.* $\square$

In preparation for our main result, we need the following step.

LEMMA 3.15. *Let (X, U) satisfy all the conditions set by Algorithm C. Suppose $1 \leq i, h \leq m$ are distinct rows of A with $C^{i,h} \neq \emptyset$. Let V be the submatrix of U formed by taking the column indexed by $\bigcap_{j \in C^{i,h}} N_j$, all the columns indexed by $Y_r \cap \bigcap_{j \in C^{i,h}-r} N_j$ (for every $r \in C^{i,h}$), and the column indexed $\tau^{i,h}$ (when $|C^{i,h}| \geq 2$).*

Suppose $\eta \in R^\Sigma$ is such that $\eta^T V \geq 0$. Then $\eta^T X \geq 0$.

Proof. By step (3.2), every column of U satisfies (32). But U is symmetric, and therefore every row of U satisfies (32). Therefore, the column of U indexed by $\mathcal{F}$ is obtained by adding together the column indexed by $\bigcap_{j \in C^{i,h}} N_j$, the columns indexed by $Y_r \cap \bigcap_{j \in C^{i,h}-r} N_j$ (for every $r \in C^{i,h}$), and the column indexed by $\tau^{i,h}$ (if $|C^{i,h}| \geq 2$). By assumption, each column v in this sum satisfies $\eta^T v \geq 0$. We conclude that the column indexed by $\mathcal{F}$ satisfies this inequality as well. Now we are done, by step (3.1). $\square$

THEOREM 3.16. *Consider an inequality $a^T x \geq a_0$ which is valid for $\mathcal{F}$ and such that $a_j \in \{0, 1, 2\}$ for $0 \leq j \leq n$. Let (X, U) be a vector and matrix satisfying all the constraints imposed by Algorithm C. Then*

$$(37) \quad \sum_{j=1}^n a_j X[Y_j] \geq a_0.$$

Proof. Assume without loss of generality that $a^T x \geq a_0$ is not dominated by another valid inequality with coefficients in $\{0, 1, 2\}$.

If $a_0 = 1$, it follows that no indices j satisfy $a_j = 2$. Then $a^T x \geq a_0$ is one of the constraints $Ax \geq e$ that define $\mathcal{F}$, and thus (37) follows by (27).

If $a_0 = 0$, then (37) is implied by (34).

We are left with the case $a_0 = 2$. If all nonzero a_j equal 2, then by dividing by 2 we return to the case $a_0 = 1$. Consequently, we can assume that $a^T x \geq a_0$ is of the form $2x(T) + x(S) \geq 2$ for disjoint subsets $T, S \subseteq \{1, \dots, n\}$, where $S \neq \emptyset$. (Here and elsewhere, for a vector v and index set J , we write $v(J) = \sum_{j \in J} v_j$.)

Recall that the support of the k th constraint is denoted by A_k . For each element $j \in S$ it is easy to see that

$$(38) \quad \text{for some } 1 \leq k \leq m, \quad A_k \subseteq T \cup S - j,$$

since otherwise the point $z \in \{0, 1\}^n$ with $z_g = 1$ iff $g \in \{1, 2, \dots, n\} - (T \cup S)$, or $g = j$, is in $\mathcal{F}$, but $a^T z < a_0$. Consequently, since $S \neq \emptyset$, we obtain that one of the original constraints $x(A_i) \geq 1$ has $A_i \subseteq T \cup S$.

Clearly $A_i \cap S \neq \emptyset$ —or else $a^T x \geq 2$ is dominated by $x(A_i) \geq 1$, a contradiction. Hence, if we pick any $j \in A_i \cap S$ and apply (38) to j we find an index $1 \leq h \leq m$ with $A_h \subseteq T \cup S - j$. Necessarily, $i \neq h$.

Suppose first that $A_i \cap A_h = \emptyset$. In that case, $a^T x \geq 2$ is dominated by the sum of $x(A_i) \geq 1$ and $x(A_h) \geq 1$, and by (27) applied to i and h , X satisfies (37).

We will therefore assume $C^{i,h} = A_i \cap A_h \neq \emptyset$. In the remainder of the proof, we will show that if the vector v is the column of U indexed by $\bigcap_{j \in C^{i,h}} N_j$, by $Y_r \cap \bigcap_{j \in C^{i,h} - r} N_j$ (for some $r \in C^{i,h}$), or by $\tau^{i,h}$, then

$$(39) \quad \sum_{j=1}^n a_j v[Y_j] - 2v[\mathcal{F}] \geq 0.$$

By Lemma 3.15, this will complete the proof. Consider first the case of the column v corresponding to $\bigcap_{j \in C^{i,h}} N_j$. By step (3.1) of the algorithm, we have

$$(40) \quad v \left[\bigcap_{j \in C^{i,h}} N_j \right] = v[\mathcal{F}].$$

By constraint (28) imposed by the algorithm,

$$(41) \quad v[N_r] \geq v \left[\bigcap_{j \in C^{i,h}} N_j \right] \quad \forall r \in C^{i,h}.$$

Consequently, (26) and (35) applied to v , together with (40) and (41), imply

$$(42) \quad v[Y_r] = 0 \quad \forall r \in C^{i,h}.$$

However, by step (3.2), v satisfies (27) applied to i . Together with (42) this implies

$$(43) \quad \sum_{j \in A_i - C^{i,h}} v[Y_j] - v[\mathcal{F}] \geq 0.$$

Similarly,

$$(44) \quad \sum_{j \in A_h - C^{i,h}} v[Y_j] - v[\mathcal{F}] \geq 0,$$

and since by construction of $C^{i,h}$ in Step 2 we have $(A_i - C^{i,h}) \cap (A_h - C^{i,h}) = \emptyset$, we conclude that

$$\sum_{j \in A_i \cup A_h} v[Y_j] - 2v[\mathcal{F}] \geq 0,$$

which dominates (39).

Consider now the case that v is the column of U corresponding to some $Y_r \cap \bigcap_{j \in C^{i,h-r}} N_j$. Then, by step (3.1) $v[\mathcal{F}] = v[Y_r \cap \bigcap_{j \in C^{i,h-r}} N_j]$, and by constraint (29) imposed by the algorithm we conclude that

$$(45) \quad v[Y_r] \geq v[\mathcal{F}].$$

If $a_r = 2$, we conclude that v satisfies (39). If, on the other hand, $a_r = 1$, then by (38) there is a constraint $x(A_k) \geq 1$ of the original system with $A_k \subseteq T \cup S - r$. By (27), we conclude that $\sum_{j \in A_k} v[Y_j] - v[\mathcal{F}] \geq 0$. This, together with (45), implies v satisfies (39).

Finally, consider the case when v is the column of U corresponding to $\tau^{i,h}$. By step (3.1), $v[\tau^{i,h}] = v[\mathcal{F}]$, and therefore by constraint (30) $\sum_{j \in C^{i,h}} v[Y_j] - 2v[\mathcal{F}] \geq 0$, which dominates (39). The theorem is proved. $\square$

REMARK 3.17. *Note that Algorithm C does not require positive- semidefiniteness of the matrix U . Nevertheless, even though the formulation generated by the algorithm is of polynomial size, it can imply exponentially many facets.*

To see the last point, consider the following example. Suppose V and W are sets and for each $v \in V$ there is a subset $J_v \subseteq W$ such that the J_v ($v \in V$) are pairwise disjoint. Consider the system $C_{W,V}$ of constraints

$$(46) \quad y_h + \sum_{j \in V-v} x_j \geq 1 \quad \forall v \in V, \quad h \in J_v,$$

$$(47) \quad (y, x) \in \{0, 1\}^W \times \{0, 1\}^V.$$

(When $W = \emptyset$, the constraint matrix defined by these inequalities is a full-circulant matrix.) The total number of constraints of type (46) is at most $|V| \times |W|$. We have the following.

PROPOSITION 3.18. *Suppose that for each $v \in V$ we choose an element $w(v) \in J_v$. Then*

$$\sum_{v \in V} y_{w(v)} + \sum_{v \in V} x_v \geq 2$$

defines a facet of $\text{conv}(C_{W,V})$. $\square$

Thus, we have a family of $\prod_{v \in V} |J_v|$ facets, all with coefficients in $\{0, 1, 2\}$ and therefore implied by Algorithm C.

Suppose we have a pure full-circulant example, i.e., $W = \emptyset$. In [10] we describe a lifting procedure that is stronger than both the N_+ and the Sherali–Adams operators, and yet needs at least $|V| - 3$ rounds to prove the valid inequality

$$(48) \quad \sum_{v \in V} x_v \geq 2.$$

In fact [11] the formulation obtained by running this stronger procedure for any fixed number of rounds t contains points x with $\sum_{v \in V} x_v \leq 1 + O(\frac{t}{|V|})$, i.e., a

violation of (48) by nearly a factor of two. It was already known [6], [19] that the N_+ procedure can be weak on covering-like formulations. What distinguishes the full-circulant example from the prior results is that (i) (48) only has small coefficients, (ii) the iterate obtained by running the N_+ operator for a polynomial amount of time can violate (48) by a large percentage, and (iii) in fact (ii) holds for a much stronger operator.

Balas and Ng [5] have completely characterized the set of facet-defining inequalities for set covering with coefficients 0, 1, 2 as rank-1 Chvátal–Gomory cuts. Also note that Letchford [26] describes a disjunctive procedure (on the original space of variables) that proves (48) in the full-circulant example.

As a final comment on Algorithm C, note that in step (2.c) we use the symbol $\tau^{i,h}$ to represent the set-theoretic expression in (31). This is necessary, as the expression in (31) has exponential length.

4. Main algorithm.

4.1. Literals and obstructions. The algorithm we describe here generalizes the approach presented in the last section. The notion of an *obstruction* plays a central role. First we need some notation.

Notation. Let $1 \leq j \leq n$. In what follows, the notation M_j will be used to denote a symbol that is either Y_j or N_j ; $\bar{Y}_j$ denotes N_j and $\bar{N}_j$ denotes Y_j . The symbols Y_j and N_j will be called *literals*.

DEFINITION 4.1. An obstruction for $a^T x \geq a_0$ is an element $\omega \in \Sigma$ of the form

$$\omega = M_{j_1} \cap M_{j_2} \cap \cdots \cap M_{j_h}$$

(where for each $1 \leq i \leq h$ we have $1 \leq j_i \leq n$ and $a_{j_i} \neq 0$) such that any $x \in \{0, 1\}^n$ satisfying

$$(49) \quad x_{j_i} = \begin{cases} 1 & \text{if } M_{j_i} = Y_{j_i}, \\ 0 & \text{if } M_{j_i} = N_{j_i} \end{cases}$$

for every $1 \leq i \leq h$ satisfies $a^T x < a_0$.

Obstructions are closely related to *covers* of knapsacks; see [27].

DEFINITION 4.2. For $1 \leq j \leq n$, $Y_j \cap N_j$ is called a *trivial obstruction*.

Using Definition 4.1, $Y_j \cap N_j$ is an obstruction to $x_j + (1 - x_j) = 1$.

Examples.

(a) *Set-covering.* Given a constraint $\sum_{j \in S} x_j \geq 1$, its unique obstruction is $\bigcap_{j \in S} N_j$.

(b) *Set-packing.* Given a constraint $\sum_{j \in S} x_j \leq 1$, with $|S| = N$, there are $\binom{N}{2}$ minimal obstructions (here “minimal” refers to the set of literals in the obstruction). Each of them is of the form $Y_{i(1)} \cap Y_{i(2)}$ for some pair $i(1), i(2)$ of distinct indices from S .

(c) *Set-partitioning.* Given a constraint $\sum_{j \in S} x_j = 1$, its set of obstructions are those obtained by viewing the constraint as a set-covering and as a set-packing constraint.

(d) *Multicovering or multipacking.* Consider a constraint which is either of covering type, $\sum_{j \in S} x_j \geq |S| - k$, or of packing type, $\sum_{j \in S} x_j \leq |S| - k$. In the covering case, the minimal obstructions consist of intersections of $k + 1$ literals N_j , and in the packing case, they consist of intersections of $k + 1$ literals Y_j . For k fixed, we can therefore enumerate all minimal obstructions in polynomial time.

(e) *Mixed covering and packing.* Consider a constraint of the form

$$\sum_{j \in I} \alpha_j x_j + \sum_{j \in J} \beta_j (1 - x_j) \geq b,$$

where I and J are disjoint, $\alpha_j > 0$ and $\beta_j > 0$ for each j , and $b \geq 0$. This generalizes the examples in (a)–(d). Clearly, no minimal obstruction contains symbols of the form Y_j , $j \in I$, or N_j , $j \in J$.

DEFINITION 4.3. Let $k \geq 0$. Consider an inequality of the form

$$\sum_{j \in J^+} a_j x_j - \sum_{j \in J^-} a_j x_j \geq b,$$

where $a_j > 0$ for all $j \in J^+ \cup J^-$. An obstruction to this inequality is called k -small if it is of the form

$$\bigcap_{j \in A^+} N_j \cap \bigcap_{j \in A^-} Y_j,$$

where $A^+ \subseteq J^+$ and either $|A^+| \geq |J^+| - k$ or $|A^+| \leq k$; and $A^- \subseteq J^-$ and either $|A^-| \geq |J^-| - k$ or $|A^-| \leq k$.

Note. for any fixed k , we can enumerate all k -small obstructions to any inequality in polynomial time. Also note that all minimal obstructions for set-covering, set-packing, and set-partitioning constraints are k -small for $k \leq 2$.

4.2. Walls and tiers. Now we return to our general lifting procedure. The algorithm we will present is a generalization of Algorithm C given in section 3.2. That algorithm created variables indexed by members of Σ , all of which were either intersections of N_j terms, intersections of N_j terms and one Y_j term, or more complicated expressions: unions of intersections of N_j terms and some number of Y_j terms. The particular expressions actually generated by the algorithm were not the complete set of all possible such expressions—rather, they reflected the structure of the problem.

The algorithm presented below generalizes this approach. Broadly speaking, the algorithm will perform three steps.

- (a) Generate a (polynomial-size) family of set-theoretic indices $\mathcal{V}$. This family will include all Y_j and N_j as well as more complex expressions defined below that we call “walls” and “tiers,” both of which are derived from obstructions to the constraints.
- (b) Write constraints that a lifting of any point $x \in \mathcal{F}$ to $\{0, 1\}^\Sigma$ must satisfy. These constraints include all (homogenized) inequalities present in the continuous relaxation of $\mathcal{F}$ and also set-theoretic constraints with support in $\mathcal{V}$. (The support of a vector v , denoted $\text{suppt}(v)$, is the set of indices j such that $v_j \neq 0$.) In particular, there will be constraints that generalize constraint (32) in Algorithm C.
- (c) Create a matrix of variables $U \in R^{\mathcal{V} \times \mathcal{V}}$. This matrix is designed to capture the implications of Lemma 3.10. So, for example, we can impose on its columns all constraints generated in (b).

The expressions created in (a) are clearly critical for the success of such an approach. We introduce them next.

DEFINITION 4.4. Suppose we are given an expression $\gamma = \bigcap_{i=1}^h M_{j_i}$, where, for $1 \leq i \leq h$, we have $1 \leq j_i \leq n$.

(0) We write $|\gamma| = h$.

(1) For $0 \leq t \leq h$, a negation of γ of order t is an expression of the form $\bigcap_{i=1}^h M'_{j_i}$ such that for exactly t indices j_i we have $M'_{j_i} = \bar{M}_{j_i}$, and for the remaining $h - t$ indices j_i we have $M'_{j_i} = M_{j_i}$.

(2) $0 \leq r \leq h$, $\mathcal{N}(\gamma, r)$ is the set of negations of γ of order r (if $r = 0$, then this is just the set $\{\gamma\}$).

(3) For $0 \leq t < h$, the negation of γ of order greater than t is the expression $\gamma^{>t}$ defined by

$$\gamma^{>t} \doteq \bigcup_{r=t+1}^h \left(\bigcup_{\beta \in \mathcal{N}(\gamma, r)} \beta \right).$$

EXAMPLE 4.5. Consider

$$\gamma = Y_4 \bigcap Y_5 \bigcap Y_6 \bigcap N_7.$$

Then $|\gamma| = 4$, and

$$\gamma' = N_4 \bigcap N_5 \bigcap Y_6 \bigcap N_7$$

is a negation of γ of order 2.

The concept of negation will be of central importance to our algorithms given later. Consider the case of γ_2 above, and let $t = 1$. Then we can partition $\mathcal{F}$ as

$$\begin{aligned} \mathcal{F} = & \gamma \bigcup \left(N_4 \bigcap Y_5 \bigcap Y_6 \bigcap N_7 \right) \bigcup \left(Y_4 \bigcap N_5 \bigcap Y_6 \bigcap N_7 \right) \\ (50) \quad & \left(Y_4 \bigcap Y_5 \bigcap N_6 \bigcap N_7 \right) \bigcup \left(Y_4 \bigcap Y_5 \bigcap Y_6 \bigcap Y_7 \right) \bigcup \gamma^{>1}. \end{aligned}$$

Since this is a partition, if $x \in \mathcal{F}$ is lifted to $X \in R^\Sigma$, we will have

$$\begin{aligned} 1 = & X[\gamma] + X \left[N_4 \bigcap Y_5 \bigcap Y_6 \bigcap N_7 \right] + X \left[Y_4 \bigcap N_5 \bigcap Y_6 \bigcap N_7 \right] \\ (51) \quad & X \left[Y_4 \bigcap Y_5 \bigcap N_6 \bigcap N_7 \right] + X \left[Y_4 \bigcap Y_5 \bigcap Y_6 \bigcap Y_7 \right] + X[\gamma^{>1}]. \end{aligned}$$

This valid constraint satisfied by X can be viewed as a multiway disjunction (see [1]). Our algorithm will systematically generate a (polynomial) number of such disjunctions, which are driven by the set of obstructions to the constraints in the problem. The following definitions lay the groundwork for this.

DEFINITION 4.6.

(a) Consider a set of expressions $E = \{\gamma_1, \gamma_2, \dots, \gamma_t\}$, where each γ_i is an intersection of literals. Then the wall derived from E is the expression $\Omega(E) = \bigcap M_j$ containing every literal M_j occurring in more than one γ_i . If no such symbol exists, we will say that the wall is empty.

(b) A tier is an expression of the form $\gamma = \omega_1 \bigcap \omega_2 \bigcap \dots \bigcap \omega_t$, where each ω_i is a negation of a wall (of some order t or greater than t for some $t \geq 0$).

EXAMPLE 4.7. Consider

$$\begin{aligned} \gamma_1 &= N_1 \bigcap Y_2 \bigcap Y_3, \\ \gamma_2 &= N_1 \bigcap Y_2 \bigcap Y_4 \bigcap N_5, \\ \gamma_3 &= N_1 \bigcap Y_4. \end{aligned}$$

Then the wall derived from $\{\gamma_1, \gamma_2\}$ is $\omega_1 = N_1 \cap Y_2$, and the wall derived from $\{\gamma_2, \gamma_3\}$ is $N_1 \cap Y_4$, of which $\omega_2 = N_1 \cap N_4$ is a negation of it. Thus, $\omega_1 \cap \omega_2 = N_1 \cap Y_2 \cap N_1 \cap Y_4$ is a tier.

Note that the last expression in this example is redundant. To handle this and similar situations we introduce the following formal procedure.

DEFINITION 4.8.

(i) Given an expression $\gamma = \bigcap_{i=1}^h M_{j_i}$, the simplification of γ is the expression of the form $\bigcap M_j$ obtained by taking exactly one copy of each distinct literal M_j appearing in γ .

(ii) Given an expression of the form $\gamma = \beta \cap \omega^{>r}$, where $\beta = \bigcap_{i=1}^h M_{j_i}$, ω is a wall and $r > 0$, and the simplification of γ is the expression $\beta' \cap \omega^{>r}$, where β' is the simplification of β .

(iii) Suppose we have an expression of the form $\beta = \gamma \cap \omega_1^{>r_1} \cap \omega_2^{>r_2}$, where ω_1 and ω_2 are walls, r_1, r_2 are positive, and γ is as in (i). Let γ' be the simplification of γ . Then the simplification of β is the expression $\gamma' \cap \omega_1^{>r_1} \cap \omega_2^{>r_2}$, if $\omega_1 \neq \omega_2$, and it equals $\gamma' \cap \omega_1^{>\max\{r_1, r_2\}}$ otherwise.

(iv) Suppose we have an expression of the form $\beta = \mathcal{F} \cap \gamma$, where γ is of the form considered in (i) or (ii). Then the simplification of β equals the simplification of γ .

COMMENT 4.9. All expressions that the algorithm given below will generate will be of one of the following types:

$\Sigma.1.$ a literal, $\mathcal{F}$, a wall, or a tier;

$\Sigma.2.$ an expression of the form $\beta \cap \gamma$, where β and γ are of type $\Sigma.1$.

Any expression of type $\Sigma.1$ – $\Sigma.2$ is of the form $B_1 \cap B_2 \cap \cdots \cap B_r$, where each B_i is a literal, $\mathcal{F}$, or of the form $\omega^{>r}$. Recall the discussion at the end of section 3.1.1. We will use the assumption that any permutation of the B_i yields an equivalent expression and also apply the simplification operator to enforce equivalences between the symbols generated by our algorithms.

DEFINITION 4.10. Given two expressions α, β that are intersections of literals, we say that β is a *superstring* of α if every literal appearing in the simplification of α also appears in the simplification of β .

4.3. The Σ^k -algorithm ($k \geq 2$). Now we can define our basic general algorithm. This algorithm embodies one particular formalization of the idea of lifting to variables indexed by the subset algebra—several other such formalizations are possible. The primary reason we use the algorithm below is that we can show that it has provably good properties. However, in a practical implementation most likely one would use only some of the algorithmic ideas we present. The algorithm is presented in somewhat redundant form.

Let $k \geq 2$ be a fixed integer, and as before we have a feasible set $\mathcal{F} = \{x \in \{0, 1\}^n : Ax \geq b\}$, where A is $m \times n$. The i th constraint is denoted $a_i^T x \geq b_i$.

1. Generate the symbol $\mathcal{F}$, and for $1 \leq j \leq n$, generate the symbols Y_j and N_j , and impose the constraint

$$(52) \quad X[Y_j] + X[N_j] - X[\mathcal{F}] = 0.$$

For $1 \leq i \leq m$, impose the constraint

$$(53) \quad \sum_{j=1}^n a_{ij} X[Y_j] - b_i X[\mathcal{F}] \geq 0.$$

2. For $1 \leq i \leq m$, enumerate all the k -small obstructions to each of the constraints $a_i^T x \geq b_i$, as well as all trivial obstructions. For each enumerated obstruction γ ,

impose the constraint

$$(54) \quad \sum_{M_j \in \gamma}^n X[\bar{M}_j] - X[\mathcal{F}] \geq 0,$$

where the notation “ $M_j \in \gamma$ ” means that the sum is over all those literals M_j occurring in γ .

3. For every set E of distinct enumerated obstructions with $1 < |E| \leq k$, compute $\Omega(E)$, the wall derived from E . If $\Omega(E) \neq \emptyset$ and $\Omega(E)$ does not contain both terms Y_j and N_j for some j ($1 \leq j \leq n$), then generate a symbol for $\Omega(E)$.

4. For every $1 \leq \mathcal{H} < k$, every tier of the form $\theta = \bigcap_{t=1}^{\mathcal{H}} \omega_t$, where ω_t is an enumerated wall for $1 \leq t \leq \mathcal{H}$, is processed to generate additional symbols and constraints, as follows.

Enumerate every $\mathcal{H}$ -vector of integers $o = (o_1, o_2, \dots, o_{\mathcal{H}})$ such that $0 \leq o_t \leq \min\{|\omega_t|, k\}$ for $1 \leq t \leq \mathcal{H}$ and $\sum_{t=1}^{\mathcal{H}} o_t < k$. Given o , enumerate every $\mathcal{H}$ -tuple $\omega' = (\omega'_1, \omega'_2, \dots, \omega'_{\mathcal{H}})$, where ω'_t is a negation of ω_t of order o_t for $1 \leq t \leq \mathcal{H}$. For each such enumerated pair (o, ω') , perform the following steps.

(i) Create a symbol for

$$\theta_o^\# = \bigcap_{t=1}^{\mathcal{H}} \omega'_t$$

and, for each $1 \leq j \leq n$ and each literal M_j such that M_j is one of the literals in $\theta_o^\#$, impose the constraint

$$(55) \quad X[M_j] - X[\theta_o^\#] \geq 0.$$

Furthermore, if the simplification σ of $\theta_o^\#$ is a superstring of some enumerated obstruction, or if it contains both symbols Y_j and N_j for some $1 \leq j \leq n$, impose

$$(56) \quad X[\theta_o^\#] = 0.$$

Finally, impose

$$(57) \quad X[\theta_o^\#] - \sum_{M_j \in \sigma} X[M_j] + (|\sigma| - 1)X[\mathcal{F}] \geq 0.$$

(ii) If $\sum_{t=1}^{\mathcal{H}} o_t = k - 1$ and $o_{\mathcal{H}} < |\omega_{\mathcal{H}}|$, create a symbol for

$$\theta_o^> = \left(\bigcap_{t=1}^{\mathcal{H}-1} \omega'_t \right) \cap \omega_{\mathcal{H}}^{>o_{\mathcal{H}}},$$

where as before $\omega_i^{>o_i}$ is the negation of ω_i of order greater than o_i . Furthermore, for each $1 \leq j \leq n$ and each literal M_j such that M_j is one of the symbols in any of the ω'_t , for $1 \leq t \leq \mathcal{H} - 1$, impose the constraint

$$(58) \quad X[M_j] - X[\theta_o^>] \geq 0.$$

Moreover, impose the constraint

$$(59) \quad \sum_{M_j \in \omega_{\mathcal{H}}} X[\bar{M}_j] - (1 + o_{\mathcal{H}})X[\theta_o^>] \geq 0.$$

Furthermore, if the simplification of $\bigcap_{t=1}^{\mathcal{H}-1} \omega'_t$ is a superstring of some enumerated obstruction, or if it contains both symbols Y_j and N_j for some $1 \leq j \leq n$, impose

$$(60) \quad X[\theta_o^>] = 0.$$

Strictly speaking, the notation $\theta_o^\#$ and $\theta_o^>$ is not complete, since we need to specify *which* literals are negated, but the correct interpretation will be clear from the context.

5. For each expression $\theta_o^\# = \bigcap_{t=1}^{\mathcal{H}} \omega'_t$ enumerated in step 4 such that $\mathcal{H} < k-1$ and ω'_t is a negation of order o_t of ω_t for $1 \leq t \leq \mathcal{H}$, and for each enumerated wall ω , impose the following constraints:

(a) If $|\omega| + \sum_{t=1}^{\mathcal{H}} o_t < k$, then impose the constraint

$$(61) \quad X[\theta_o^\#] - \sum_{r=0}^{|\omega|} \left(\sum_{\beta \in \mathcal{N}(\omega, r)} X[\theta_o^\# \cap \beta] \right) = 0.$$

(b) If $|\omega| + \sum_{t=1}^{\mathcal{H}} o_t \geq k$, then, setting $R = k-1 - \sum_{t=1}^{\mathcal{H}} o_t$, impose the constraint

$$(62) \quad X[\theta_o^\#] - X[\theta_o^\# \cap \omega^{>R}] - \sum_{r=0}^R \left(\sum_{\beta \in \mathcal{N}(\omega, r)} X[\theta_o^\# \cap \beta] \right) = 0,$$

where, as previously, $\mathcal{N}(\omega, r)$ denotes the set of negations of ω of order r . Similarly, for each enumerated wall ω , if $|\omega| < k$, then impose

$$(63) \quad X[\mathcal{F}] - \sum_{r=0}^{|\omega|} \left(\sum_{\beta \in \mathcal{N}(\omega, r)} X[\beta] \right) = 0,$$

and if $|\omega| \geq k$, then impose

$$(64) \quad X[\mathcal{F}] - X[\omega^{>k-1}] - \sum_{r=0}^{k-1} \left(\sum_{\beta \in \mathcal{N}(\omega, r)} X[\beta] \right) = 0.$$

6. Let $\mathcal{V}$ denote the set of variable indices created so far, and let X denote the vector of variables. Create a matrix U of variables, with rows and columns indexed by $\mathcal{V}$, and impose on the variables in U the following constraints.

(6.1) U is symmetric, $U[\mathcal{F}, \mathcal{F}] = X[\mathcal{F}]$, and the main diagonal, the $\mathcal{F}$ -row and the $\mathcal{F}$ -column of U , are all equal to X .

(6.2) For each constraint $\eta^T X \geq 0$ of the form (53)–(64) (when it applies) impose the constraints

$$(65) \quad \eta^T U \geq 0.$$

(6.3) Suppose $\beta_1, \gamma_1, \beta_2, \gamma_2$ are elements of $\mathcal{V}$ such that $\beta_1 \cap \gamma_1$ and $\beta_2 \cap \gamma_2$ have the same simplification. Then require that $U[\beta_1, \gamma_1] = U[\beta_2, \gamma_2]$. Suppose β, γ are elements of $\mathcal{V}$ such that $\beta \cap \gamma$ is a superstring of some enumerated obstruction or contains both Y_j and N_j for some j . Then impose $U[\beta, \gamma] = 0$.

(6.4) Impose

$$(66) \quad 0 \leq U[\theta, \beta] \leq U[\mathcal{F}, \beta] \quad \forall \theta, \beta \in \mathcal{V},$$

$$(67) \quad X[\mathcal{F}] = 1.$$

7. (Optional) Impose $U \succeq 0$.

8. (Optional) For each constraint $\eta^T X \geq 0$ of the form (53), (54), (57), or (59), let $f^\eta \doteq U\eta$. Form the matrix W_η with rows and columns indexed by $\mathcal{V}$ such that for each pair β, γ of symbols in $\mathcal{V}$, the β, γ -entry of W_η equals $f^\eta[\beta \cap \gamma]$. Impose $W_\eta \succeq 0$.

9. End.

Comments on the algorithm.

1. In step (6.2) some of the constraints imposed on U are of the form “=” rather than “ $\geq$ ” as stated in the algorithm—then in (65) we should use “=”.
2. Note that the simplification step in (6.3) implies that, for example, if β_1 and β_2 are elements of $\mathcal{V}$ with the same simplification, then the columns (and rows) of U corresponding to β_1 and β_2 are identical. In such a case, one could simply keep a single row and column corresponding to the equivalence class of symbols that have the same simplification. To keep the analysis simple, in what follows we will assume that we are using the entire matrix.
3. In steps 5(a), (b), note that the expressions $\theta_o^\# \cap \beta$ are obtained by negating walls in a tier enumerated in step 4. Hence the variables $X[\theta_o^\# \cap \beta]$ will be created by the algorithm and similarly with all other variables in (61)–(64).
4. Step 7 is M_+ -like, using of course a very different set of variables. Similarly, Step 8 is Lasserre-like (refer to Lemma 3.9(iii)).

Notation. In what follows we will use the notation $\text{col}[\theta]$ to mean the column of U corresponding to the symbol $\theta \in \mathcal{V}$. The matrix U will always be clear from the context. Note that $\text{col}[\theta] = Ue_\theta$, where $e_\theta \in R^\mathcal{V}$ has a 1 in position θ and zeros elsewhere, but we prefer the “col” notation.

The following result, which is straightforward, will be used in many of the proofs below, and we record it for reference.

PROPOSITION 4.11. *Let $\mathcal{V}$ and U be the set of symbols and the matrix produced by a run of the Σ^k -algorithm. Suppose there is an equation $\sum_{\theta \in \mathcal{V}} \lambda[\theta]u[\theta] = 0$ which is satisfied by every column u of U , where $\lambda \in R^\mathcal{V}$. Then $\sum_{\theta \in \mathcal{V}} \lambda[\theta]\text{col}[\theta] = \vec{0}$, where $\vec{0} \in R^\mathcal{V}$ is the 0-vector.*

Proof. Since U is symmetric, it follows that $\sum_{\theta \in \mathcal{V}} \lambda[\theta]r[\theta] = 0$ is satisfied by every row r of U , which implies the result. $\square$

4.4. Some basic properties of the Σ^k -algorithm. In this section we prove that the algorithm generates a valid formulation and that it runs in polynomial time for fixed k .

In what follows, let $\mathcal{M}$ denote the set of distinct members of Σ , i.e., subsets of $\{0, 1\}^n$, that arise as the set-theoretic value of variable indices $\mathcal{V}$ produced by the algorithm.

LEMMA 4.12. *Let $\hat{x} \in \mathcal{F}$. Then $\hat{x}$ can be lifted to a vector $\check{X} \in \{0, 1\}^\mathcal{V}$ and matrix $\check{U} \in \{0, 1\}^{\mathcal{V} \times \mathcal{V}}$ that satisfy the constraints imposed by the algorithm.*

Proof. We will show that $\hat{x}$ can be lifted to a vector $\hat{x} \in R^\mathcal{M}$ and a matrix $\mathcal{M}$ which satisfy the desired constraints. By appealing to Lemma 3.12, we will be done.

Let $\alpha(\hat{x}) \in \Sigma$ denote the atom corresponding to $\hat{x}$, i.e.,

$$\alpha(\hat{x}) = \left(\bigcap_{j: \hat{x}_j=1} Y_j \right) \cap \left(\bigcap_{j: \hat{x}_j=0} N_j \right).$$

Define $\hat{X} \in \{0, 1\}^\mathcal{M}$ to be the restriction of $\xi^{\alpha(\hat{x})} \in \{0, 1\}^\Sigma$ to the coordinates in $\mathcal{M}$. Recall here that $\xi^{\alpha(\hat{x})}$ is a 0-1 vector such that for each $\beta \in \Sigma$, $\xi^{\alpha(\hat{x})}[\beta]$ equals one iff

the subset of $\mathcal{F}$ defined by β contains the point $\hat{x}$. Furthermore, let $\hat{U}$ be defined by $\hat{U}[\beta, \gamma] = \hat{X}[\beta \cap \gamma]$.

Let $\check{X}$ denote the restriction of $\hat{X}$ to $\{0, 1\}^\mathcal{V}$ and $\check{U} = \hat{U}$. We claim that $\check{X}$ and $\check{U}$ are as desired.

Clearly (52), (53), and (54) are satisfied. Consider now expressions θ , $\theta_o^\#$ as in step 4 of the algorithm. Since $\theta_o^\#$ is of the form $\bigcap_{j \in J} M_j$ for some set J , clearly (55) is satisfied (that is, the subset of $\mathcal{F}$ defined by $\theta_o^\#$ is contained in the subset defined by M_j for $j \in J$). Similarly, (58) is satisfied. Constraint (57) holds because if $|\sigma|$ terms $\check{X}[M_j]$ equal 1, then so does $\check{X}[\theta_o^\#]$, and hence in any case the left-hand side of (57) is nonnegative. Equations (56) and (58) are clear. Finally, in constructing $\theta_o^>$, at least $1 + o_{\mathcal{H}}$ symbols are negated in $\omega_{\mathcal{H}}$, and consequently (59) holds as well.

Consider now (61)–(64). Suppose that case 5(a) applies. If $\hat{X}[M_j] = 0$, for at least one literal M_j which appears in at least one of the expressions ω'_t , for $1 \leq t \leq \mathcal{H}$, then by construction both terms in (61) are equal to zero and (61) holds. Suppose now that $\hat{X}[\theta_o^\#] = 1$. Let $0 \leq K \leq |\omega|$ be the number of indices j such that either Y_j is a symbol in ω and $\hat{x}_j = 0$, or N_j is a symbol in ω and $\hat{x}_j = 1$. By construction in step 3 of the algorithm, K is well defined; exactly one term $X[\theta_o^\# \cap \beta]$ with $\beta \in \mathcal{N}(\omega, K)$ will equal 1, and all remaining terms $X[\theta_o^\# \cap \beta]$ will be zero, and (61) follows. The proof that (62)–(64) hold, when they apply, is similar.

Finally, by construction, the conditions in (6.1), (6.2), and (6.4) apply; and the fact that the conditions in the optional steps 7 and 8 hold follows from Lemma 3.9. $\square$

COROLLARY 4.13. *Suppose $\tilde{x} \in \text{conv}(\mathcal{F})$. Then there exists a vector $\tilde{X} \in R^\mathcal{V}$ and a matrix $\tilde{U} \in R^{\mathcal{V} \times \mathcal{V}}$ satisfying the conditions imposed by the Σ^k -algorithm such that $\tilde{X}[Y_j] = \tilde{x}_j$ for all $1 \leq j \leq n$.*

Proof. This follows from Lemmas 3.10 and 4.12. $\square$

LEMMA 4.14. *For k fixed, the Σ^k -algorithm generates polynomially many variables and constraints.*

Proof. Suppose that we start with a formulation with n variables. Then there are at most $O(mn^k)$ k -small obstructions. Thus at most $O(m^k n^{k^2})$ walls, and at most $O(m^{k^2-k} n^{k^3-k})$ tiers θ are enumerated. As a result, at most $O(2^{3k} k! m^{k^2-k} n^{k^3})$ symbols $\theta_o^\#$ and $\theta_o^>$ will be created. $\square$

The following result is not needed in what follows, and we state it only for completeness, although its proof is not difficult. Let $\mathcal{R}_k^n$ denote the set of those vectors $x \in [0, 1]^n$ such that x can be lifted to a pair (X, U) which satisfies the Σ^k -algorithm constraints.

LEMMA 4.15. (a) *Let $k \geq 3$. Then $\mathcal{R}_k^n \subseteq \mathcal{R}_{k-1}^n$.* (b) *Suppose that each variable appears in at least one constraint. Then $\mathcal{R}_{n+1}^n = \text{conv}(\mathcal{F})$.* $\square$

COMMENT 4.16. *For a proof, see [30]. In particular, in (b), the U matrix generated by the Σ^k -algorithm ($k \geq 3$) has the property that every column indexed by a literal satisfies the Σ^{k-1} constraints. Furthermore, one can prove general conditions where fewer than n rounds are possible.*

Observation. Some of the expressions produced by the Σ^k -algorithm are of the form $\bigcap_{j=1}^h M_{i_j}$, where the M_{i_j} are literals. These expressions amount to simple conjunctions of variables and their negations, and as such will “eventually” be enumerated by all the other lifting procedures we have discussed, for example the Sherali–Adams operator. The “eventually” here is important, because the quantity of literals in the expression, h , may be very large—in general, our algorithm will enumerate expressions with a number of literals that is unbounded as a function of k . To put it differently,

in polynomial time we enumerate expressions that, using the other operators, are not enumerated until after an exponential amount of work. In addition, the expressions that we do enumerate are obtained from the obstructions to the constraints, i.e., are driven by the structure of the problem. Furthermore, expressions of the form $\gamma^{>t}$ are *never* enumerated by the other algorithms.

The nature of the expressions we enumerate, and of the constraints placed upon them, are concrete algorithmic details that distinguish our procedure from the other lifting procedures. In the following sections we will present specific examples of inequalities that have unbounded rank for (say) the N_+ operator but which the Σ^k -algorithm satisfies for (small) fixed k .

4.5. Applications to set covering. In this section we consider a set-covering problem with feasible region $\mathcal{F} = \{x \in \{0, 1\}^n : Ax \geq e\}$, where the 0-1 matrix A is $m \times n$. We would like to prove, as a generalization of Theorem 3.16, that the Σ^k -algorithm implies all valid inequalities with coefficients $0, 1, 2, \dots, k$. However, a direct inductive proof (on k) does not work, and we need a stronger inductive assumption.

This is provided by the concept of *pitch*: we will prove that the Σ^k -algorithm implies all inequalities of pitch $\leq k$ which are valid for $\mathcal{F}$. For completeness, we redefine pitch here.

DEFINITION 4.17. *Given an inequality $a^T x \geq a_0$ with indices ordered so that $0 < a_1 \leq a_2 \leq \dots \leq a_J$ and $a_j = 0$ for $j > J$, its pitch is the minimum integer $\pi = \pi(a, a_0)$ such that $\sum_{j=1}^{\pi} a_j \geq a_0$.*

In fact, not only is pitch $\leq k$ a right inductive assumption, but it is a more appropriate concept, since it serves to parameterize all valid inequalities, and any valid inequality has pitch $\leq n$. We note the following result, which follows directly from the definition of pitch.

CLAIM 4.18. *Consider a valid inequality $a^T x \geq a_0$. Let $I \subseteq \text{suppt}(a)$ be such that $\sum_{i \in I} a_i < a_0$. Then the pitch of $a^T x \geq a_0$ is at least $|I| + 1$. $\square$*

We denote by A_i the support of the i th row of A . We will assume, without loss of generality, that no A_i contains another.

CLAIM 4.19. *If $\sum_{j \in V} x_j \geq 1$ is valid for $\mathcal{F}$ for some set V , then there is a row h of A with $A_h \subseteq V$. $\square$*

CLAIM 4.20. *Suppose $\alpha^T x \geq \alpha_0$ is a valid inequality for $\mathcal{F}$ with nonnegative coefficients. Consider the set-covering problem with feasible region $\mathcal{G} = \{x \in \{0, 1\}^{\text{suppt}(\alpha)} : \hat{A}x \geq e\}$, where $\hat{A}$ is the submatrix of A obtained by*

- (i) *using the columns in $\text{suppt}(\alpha)$ and*
- (ii) *using those rows i such that $A_i \subseteq \text{suppt}(\alpha)$.*

Let $\bar{\alpha}$ be the restriction of α to $\text{suppt}(\alpha)$. Then $\bar{\alpha}^T x \geq \alpha_0$ is valid for $\mathcal{G}$.

Proof. Assume by contradiction that we can find a point $\hat{x} \in \mathcal{G}$ with $\alpha^T \hat{x} < \alpha_0$. Define $\check{x} \in \{0, 1\}^n$ by setting $\check{x}_j = \hat{x}_j$ if $j \in \text{suppt}(\alpha)$, and $\check{x}_j = 1$ otherwise. Then by construction $\check{x} \in \mathcal{F}$, but $\alpha^T \check{x} < \alpha_0$, a contradiction. $\square$

The following result describes a key structural property of valid inequalities for set-covering problems and will be at the core of our proof of Theorem 1.2.

PROPOSITION 4.21. *Suppose $\alpha^T x \geq \alpha_0$ is a valid inequality for $\mathcal{F}$ and with $\alpha \geq 0$ and $\alpha_0 > 0$. Then there is a subset $\mathcal{C} = \mathcal{C}(\alpha, \alpha_0)$ of the rows of A with $|\mathcal{C}| \leq \pi(\alpha, \alpha_0)$, such that*

$$(68) \quad (i) \ A_i \subseteq \text{suppt}(\alpha) \ \forall \ i \in \mathcal{C}$$

$$(69) \quad (ii) \ \Delta_i \doteq A_i - \bigcup_{r \in \mathcal{C} - i} A_r \neq \emptyset \ \forall \ i \in \mathcal{C}, \text{ and}$$

$$(70) \quad \text{(iii)} \quad \sum_{i \in \mathcal{C}} \min \{ \alpha_j : j \in \Delta_i \} \geq \alpha_0.$$

Proof. The proof of the lemma will be by induction on $\pi = \pi(\alpha, \alpha_0)$. Since $\alpha_0 > 0$, we must have that $\sum_{j \in \text{suppt}(\alpha)} x_j \geq 1$ is valid for $\mathcal{F}$, and thus the case $\pi = 1$ is proved.

In the remainder of the proof, we assume that $A_i \subseteq \text{suppt}(\alpha)$ for every row i ; without loss of generality we can make this assumption by Claim 4.20.

Assume now that $\pi > 1$. Suppose first that some row i of A satisfies $A_i = \text{suppt}(\alpha)$. Since by assumption no row of A contains another, we have a set-covering problem with one constraint, namely $\sum_{j \in \text{suppt}(\alpha)} x_j \geq 1$. Thus, either $\pi = 1$, or $\alpha^T x \geq \alpha_0$ is not valid for $\mathcal{F}$, a contradiction in either case.

Hence, choose any row $i(1)$; $A_{i(1)}$ will then be properly included in $\text{suppt}(\alpha)$. Let $j(1)$ be an index in $A_{i(1)}$ with minimum coefficient $\alpha_{j(1)}$. Assume $\alpha_{j(1)} < \alpha_0$ (or else we are done by setting $\mathcal{C} = \{i(1)\}$).

Let $K = \text{suppt}(\alpha) - A_{i(1)}$. Consider the set-covering problem with feasible region $\mathcal{H} = \{x \in \{0, 1\}^K : \bar{A}x \geq e\}$, where $\bar{A}$ is the submatrix of A , where

1. $\bar{A}$ has column set K ;
2. for any row h of A , with $j(1) \notin A_h$, $\bar{A}$ will have a row with $\bar{A}_h = A_h \cap K$.

We claim that the inequality

$$(71) \quad \sum_{j \in K} \alpha_j x_j \geq \alpha_0 - \alpha_{j(1)}$$

is valid for $\mathcal{H}$. Otherwise we can find $\bar{x} \in \mathcal{H}$ with $\sum_{j \in K} \alpha_j \bar{x}_j < \alpha_0 - \alpha_{j(1)}$. In that case, define $\check{x} \in \{0, 1\}^n$ by setting $\check{x}_j = \bar{x}_j$ if $j \in K$, $\check{x}_{j(1)} = 1$, and $\check{x}_j = 0$ for all $j \in A_{i(1)} - j(1)$ (recall that we are assuming $A_i \subseteq \text{suppt}(\alpha)$ for all i). Clearly $\alpha^T \check{x} < \alpha_0$. However, $\check{x} \in \mathcal{F}$, because no row of A is contained in another (and, in particular, not contained in $A_{i(1)}$). This contradiction shows that (71) is indeed valid for $\mathcal{H}$.

Since $\alpha_{j(1)} > 0$, it follows that the pitch of (71) is less than π . Since $\alpha_{j(1)} < \alpha_0$, the result now follows by induction, as $j(1)$ is not contained in any set $\bar{A}_i$. $\square$

COROLLARY 4.22. *Suppose the inequality $a^T x \geq a_0$ is valid for $\mathcal{F}$ and has pitch ≤ 1 . Then $a^T x \geq a_0$ is dominated by $a_i x \geq 1$ for some row i of A or is dominated by the nonnegativity constraints.* $\square$

In the rest of this section, we will consider a fixed vector $\tilde{X} \in R^\mathcal{V}$ and matrix $\tilde{U} \in R^{\mathcal{V} \times \mathcal{V}}$ that satisfy the constraints imposed by the Σ^k -algorithm. Our goal is to show that Theorem 1.2 holds, i.e., that for every inequality $a^T x \geq a_0$ with $\pi(a, a_0) \leq k$ which is valid for $\mathcal{F}$, $\tilde{X}$ satisfies the (homogenized) inequality

$$(72) \quad \sum_j a_j X[Y_j] - a_0 X[\mathcal{F}] \geq 0,$$

which implies Theorem 1.2 since by construction $\tilde{X}[\mathcal{F}] = 1$.

4.5.1. Brief outline of the proof of Theorem 1.2. We will next outline our strategy toward this goal, which relies on using induction and on the fact that the algorithm enforces constraints (61)–(64).

Suppose that we could express $\tilde{X}$ as a sum of other columns of $\tilde{U}$, each of which satisfies (72). Since (72) is homogeneous, it follows that $\tilde{X}$ satisfies (72) as well. This “decomposition” approach mirrors the strategy we followed in the algorithm given in section 3.2.

To fix ideas, consider an instance of constraint (63) arising from a particular choice of ω . Since this constraint is satisfied by every column of $\tilde{U}$, by Proposition 4.11 we have

$$(73) \quad \tilde{X} = \sum_{r=0}^{|\omega|} \left(\sum_{\beta \in \mathcal{N}(\omega, r)} \text{col}[\beta] \right).$$

Now (73) (or a similar decomposition if (64) applies instead of (63)) holds for every wall ω enumerated in step 5. Consequently, by the discussion in the previous paragraph, our task would be complete if we could select ω so that each column in the sum satisfies (72).

In our proof, the particular ω that will give the desired decomposition will be supplied by Proposition 4.21. In order to show that each resulting β -column of $\tilde{U}$ satisfies (72), we will use a special proof for the case $r = 0$ and for the column corresponding to $\omega^{>k-1}$ if (64) applies. For the cases where $r > 0$ we will use induction.

The induction is applied as follows. Consider the term β corresponding to a particular $r > 0$. Let $\tilde{u} = \text{col}[\beta]$. We want to show that

$$(74) \quad \sum_j a_j \tilde{u}[Y_j] - a_0 \tilde{u}[\mathcal{F}] \geq 0.$$

Since we are dealing with a set-covering problem and $r > 0$, each of the negations of order r yields a Y_j literal appearing in β . As we will argue, for each such j we have $\tilde{u}[Y_j] = \tilde{u}[\mathcal{F}]$. Thus, instead of having to show (74) we will have to show

$$\sum_{j: Y_j \notin \beta} a_j \tilde{u}[Y_j] - \left(a_0 - \sum_{j: Y_j \in \beta} a_j \right) \tilde{u}[\mathcal{F}] \geq 0.$$

It turns out that this is a weaker condition, because, as we will show, the inequality

$$\sum_{j: Y_j \notin \beta} a_j x_j \geq \left(a_0 - \sum_{j: Y_j \in \beta} a_j \right)$$

is valid for $\mathcal{F}$ and has pitch strictly less than the pitch of $a^T x \geq a_0$. Hence, we can apply induction, after another use of Proposition 4.21. Subsequent inductive steps will use (61) and (62) instead of (63) and (64). Finally, for the basis of the induction we have to handle valid inequalities of pitch ≤ 1 —but all such inequalities are satisfied by *all* columns of $\tilde{U}$, as implied by Corollary 4.22.

4.5.2. Formal proof of Theorem 1.2. Now we continue with the formal proof. Note that since we are dealing with a set-covering problem all obstructions will be of the form $\bigcap_{j \in A_h} N_j$ for some h .

The following observation will be of use later.

PROPOSITION 4.23. *Let $\theta = \bigcap_{t=1}^{\mathcal{H}} \omega_t$ be a tier enumerated in step 4 of the algorithm. Consider one of the expressions $\theta_o^\#$ obtained from θ , and let $\tilde{u} = \text{col}[\theta_o^\#]$. Suppose $1 \leq j \leq n$ is such that Y_j appears in $\theta_o^\#$. Then*

$$(75) \quad \tilde{u}[Y_j] = \tilde{u}[\mathcal{F}].$$

Similarly, if $\theta_o^> = (\bigcap_{t=1}^{\mathcal{H}-1} \omega'_t) \cap \omega_{\mathcal{H}}^{>o_{\mathcal{H}}}$ is obtained from θ and Y_j appears in $\bigcap_{t=1}^{\mathcal{H}-1} \omega'_t$, then (75) holds as well.

Proof. This follows from (55), (58), step (6.1), and (66). $\square$

First, we will prove the base for the induction.

LEMMA 4.24. *Suppose $\sum_{j=1}^n \alpha_j x_j \geq \alpha_0$ is an inequality valid for $\mathcal{F}$ of pitch ≤ 1 .*

Let $\tilde{u}$ be any column of $\tilde{U}$. Then $\sum_{j=1}^n \alpha_j \tilde{u}[Y_j] - \alpha_0 \tilde{u}[\mathcal{F}] \geq 0$.

Proof. This follows from Corollary 4.22 and the fact that, due to step 6 of the algorithm, every column of $\tilde{U}$ satisfies each of the set-covering constraints, $Ax \geq e$, homogenized. $\square$

In order to develop the inductive proof outlined in the previous section, we need to parameterize those expressions generated by the algorithm. This will be done through a value that essentially counts (in reverse) the number of negations in an expression. In what follows, we use the following notation. Consider expressions $\theta = \bigcap_{t=1}^{\mathcal{H}} \omega_t$ and $\theta_o^\# = \bigcap_{t=1}^{\mathcal{H}} \omega'_t$ generated in step 4 of the algorithm, where each ω'_t is a negation of order o_t of ω_t .

DEFINITION 4.25. *We will say that $\theta_o^\#$ is positive if $0 < o_t$ for all t , $1 \leq t \leq \mathcal{H}$.*

We let

$$Y(\theta_o^\#) = \{1 \leq j \leq n : Y_j \text{ appears in } \theta_o^\#\}.$$

Note that since we are dealing with a set-covering problem, every $j \in Y(\theta_o^\#)$ appears as N_j in some ω_t and is thus negated in constructing $\theta_o^\#$; i.e., it appears as Y_j in ω'_t . Finally, we write

$$(76) \quad \rho(\theta_o^\#) = k - \sum_{t=1}^{\mathcal{H}} o_t.$$

Strictly speaking, we should write $\rho(\theta_o^\#, \theta)$, but as θ will always be clear from the context we will omit it. Also note that $\rho(\theta_o^\#) < k$ when $\theta_o^\#$ is positive. Furthermore, it is important to notice that the definition of ρ uses $\theta_o^\#$ itself, and not its simplification. We will also write

$$Y(\mathcal{F}) = \emptyset$$

and

$$\rho(\mathcal{F}) = k.$$

The ρ parameter will be used to drive the induction. Next we have the main result, which implies Theorem 1.2.

THEOREM 4.26. *Consider an expression γ so that either $\gamma = \mathcal{F}$, or γ is some positive $\theta_o^\#$ enumerated in step 4. Let*

$$(77) \quad \sum_{j=1}^n \alpha_j x_j \geq \alpha_0$$

be an inequality valid for $\mathcal{F}$ with $\pi(\alpha, \alpha_0) \leq \rho(\gamma)$. Then the column $\tilde{u}$ of $\tilde{U}$ corresponding to γ satisfies

$$(78) \quad \sum_{j=1}^n \alpha_j \tilde{u}[Y_j] - \alpha_0 \tilde{u}[\mathcal{F}] \geq 0.$$

Proof. The proof will be by induction on $\rho(\gamma)$, which, as just discussed, is positive. Lemma 4.24 therefore handles the base of this induction.

By Proposition 4.23, if $j \in Y(\gamma)$, then $\tilde{u}[Y_j] = \tilde{u}[\mathcal{F}]$, and as a result

$$\sum_{j \in Y(\gamma)} \alpha_j \tilde{u}[Y_j] = \left(\sum_{j \in Y(\gamma)} \alpha_j \right) \tilde{u}[\mathcal{F}].$$

Consequently, if we could also prove that the inequality

$$(79) \quad \sum_{j \notin Y(\gamma)} \alpha_j X[Y_j] - \left(\alpha_0 - \sum_{j \in Y(\gamma)} \alpha_j \right) X[\mathcal{F}] \geq 0$$

is satisfied by setting $X = \tilde{u}$, we would complete the proof of the theorem. This is what we will do next.

Rewrite (79) as

$$(80) \quad \sum_{j=1}^n \bar{\alpha}_j X[Y_j] - \bar{\alpha}_0 X[\mathcal{F}] \geq 0,$$

where, for $1 \leq j \leq n$,

$$(81) \quad \bar{\alpha}_j = \begin{cases} \alpha_j & \text{if } j \notin Y(\gamma), \\ 0 & \text{otherwise,} \end{cases}$$

and $\bar{\alpha}_0 = \alpha_0 - \sum_{j \in Y(\gamma)} \alpha_j$. The inequality

$$(82) \quad \bar{\alpha}^T x \geq \bar{\alpha}_0$$

is clearly valid for $\mathcal{F}$ (since (77) is); without loss of generality, $\bar{\alpha}_0 \geq 0$ (else (80) follows trivially), and by the definition of pitch, $\pi \doteq \pi(\bar{\alpha}, \bar{\alpha}_0) \leq \pi(\alpha, \alpha_0)$.

First we claim that, in the case that $\gamma = \theta_o^\#$, without loss of generality γ is a conjunction of strictly fewer than $k - 1$ negated walls. Say $\gamma = \bigcap_{t=1}^{\mathcal{H}} \omega'_t$, where ω'_t is a negation of ω_t of order o_t for $1 \leq t \leq \mathcal{H}$. We want to argue that, without loss of generality, $\mathcal{H} < k - 1$. However, γ is positive, by assumption in this theorem. Thus $o_t > 0$ for $1 \leq t \leq \mathcal{H}$. Then by the definition of ρ , we have that $\rho(\gamma) \leq k - \mathcal{H}$. Hence, if $\mathcal{H} \geq k - 1$ the theorem is proved by Lemma 4.24. In what follows we assume that $\mathcal{H} < k - 1$.

Suppose we apply Proposition 4.21 to (82). Let $\mathcal{C}$ denote the resulting subset of rows of A . Thus we obtain a set of obstructions, one from each row of $\mathcal{C}$; since $|\mathcal{C}| \leq \pi \leq \rho(\gamma) \leq k$, the wall ω derived from this set of obstructions will be enumerated in step 3.

Hence, whether $\gamma = \mathcal{F}$ or $\gamma = \theta_o^\#$, in step 5 the algorithm will impose either

$$(83) \quad X[\gamma] = \sum_{r=0}^{|\omega|} \left(\sum_{\beta \in \mathcal{N}(\omega, r)} X[\gamma \cap \beta] \right)$$

if $|\omega| < \rho(\gamma)$ (cf. (76), and see the rule that determines, for example, that 5(a) applies) or

$$(84) \quad X[\gamma] = X[\gamma \cap \omega^{>R}] + \sum_{r=0}^R \left(\sum_{\beta \in \mathcal{N}(\omega, r)} X[\gamma \cap \beta] \right)$$

otherwise, where $R = \rho(\gamma) - 1$.

Now one of (83) or (84) applies; the algorithm enforces that constraint on all columns of $\tilde{U}$; hence by Proposition 4.11

$$(85) \quad col[\gamma] = \sum_{r=0}^{|\omega|} \left(\sum_{\beta \in \mathcal{N}(\omega, r)} col[\gamma \cap \beta] \right)$$

or

$$(86) \quad col[\gamma] = col[\gamma \cap \omega^{>R}] + \sum_{r=0}^R \left(\sum_{\beta \in \mathcal{N}(\omega, r)} col[\gamma \cap \beta] \right)$$

will hold.

We will next show, using induction, that all columns of the form $col[\gamma \cap \beta]$ arising from values $r > 0$ satisfy (80). Using a special proof, we will show the same fact for the case $r = 0$. Finally, assuming (86) applies, we will show the same for $col[\gamma \cap \omega^{>R}]$. This will complete the proof of the theorem.

Case $r > 0$. Consider a fixed value $r > 0$, and consider a particular $\beta \in \mathcal{N}(\omega, r)$. Thus, β is of the form

$$\bigcap_{j \in S} Y_j \cap \bigcap_{j \in T} N_j,$$

where $|S| = r$, $\omega = \bigcap_{j \in S \cup T} N_j$, and S and T are disjoint. Let $\hat{u} = col[\gamma \cap \beta]$. We want to show that $\hat{u}$ satisfies (80).

By Proposition 4.23, $\hat{u}[Y_j] = \hat{u}[\mathcal{F}]$ for each $j \in S$. Thus, it suffices to prove that $\hat{u}$ satisfies the inequality

$$(87) \quad \sum_{j \notin S} \bar{\alpha}_j X[Y_j] - \left(\bar{\alpha}_0 - \sum_{j \in S} \bar{\alpha}_j \right) X[\mathcal{F}] \geq 0$$

in order to prove that $\hat{u}$ satisfies (80).

But notice that by construction in Proposition 4.21, all rows in the set $\mathcal{C}$ have support contained in $suppt(\bar{\alpha})$. Hence $\bar{\alpha}_j > 0$ for all $j \in S$, and as a result the pitch of (87) is at most the pitch of (82) minus r ; i.e., $\leq \rho(\gamma) - r < \rho(\gamma)$. Consequently, if we can show that

$$\rho(\gamma \cap \beta) \geq \rho(\gamma) - r,$$

then by induction we will have that $\hat{u}$ satisfies (87).

Consider the case $\gamma = \theta_o^\#$. We have that $\gamma \cap \beta$ is obtained from $\theta \cap \omega$ by (i) negating literals in θ to obtain $\theta^\#$, and (ii) since $\gamma \in \mathcal{N}(\omega, r)$, by negating exactly r literals in ω . Thus $\rho(\gamma \cap \beta) = \rho(\gamma) - r$, as desired. In the case $\gamma = \mathcal{F}$, this is also clear.

Case $r = 0$. Let $\hat{u} = col[\gamma \cap \beta] = col[\gamma \cap \omega]$. For convenience, we restate here the properties satisfied by the set of rows $\mathcal{C}$ which we are using in this proof, produced by applying Proposition 4.21 to $\bar{\alpha}^T x \geq \bar{\alpha}_0$:

- (i) $A_i \subseteq suppt(\bar{\alpha}) \quad \forall i \in \mathcal{C}$,
- (ii) $\Delta_i \doteq A_i - \bigcup_{r \in \mathcal{C}-i} A_r \neq \emptyset \quad \forall i \in \mathcal{C}$, and
- (iii) $\sum_{i \in \mathcal{C}} \min \{ \bar{\alpha}_j : j \in \Delta_i \} \geq \bar{\alpha}_0$.

Furthermore, each $i \in \mathcal{C}$ gives rise to one obstruction, and ω is the wall derived from these obstructions. Thus, $\omega = \bigcap_{j \in J} N_j$, where $J \subseteq \{1, 2, \dots, n\}$ is the set of all j appearing in at least two A_i , $i \in \mathcal{C}$. Note that, by (55) and step 6 of the algorithm,

$$\hat{u}[N_j] \geq \hat{u}[\gamma \cap \omega] = \hat{u}[\mathcal{F}] \quad \forall j \in J,$$

and consequently

$$(88) \quad \hat{u}[Y_j] = 0 \quad \forall j \in J.$$

Furthermore, consider any row $i \in \mathcal{C}$. By step (6.2) of the algorithm,

$$\sum_{j=1}^n a_{ij} \hat{u}[Y_j] - \hat{u}[\mathcal{F}] \geq 0.$$

Combining this with (88) we obtain

$$\sum_{j \in \Delta_i} a_{ij} \hat{u}[Y_j] - \hat{u}[\mathcal{F}] \geq 0$$

and consequently

$$(89) \quad \sum_{j \in \Delta_i} \bar{\alpha}_j \hat{u}[Y_j] \geq (\min \{\bar{\alpha}_j : j \in \Delta_i\}) \hat{u}[\mathcal{F}].$$

By construction, the sets Δ_i are pairwise disjoint. So if we sum (89) over all $i \in \mathcal{C}$ we obtain

$$(90) \quad \begin{aligned} \sum_{j=1}^n \bar{\alpha}_j \hat{u}[Y_j] &\geq \left(\sum_{i \in \mathcal{C}} \min \{\bar{\alpha}_j : j \in \Delta_i\} \right) \hat{u}[\mathcal{F}] \\ &\geq \bar{\alpha}_0 \hat{u}[\mathcal{F}], \end{aligned}$$

where the last inequality follows by property (iii) of the set $\mathcal{C}$. This concludes the proof in this case.

Case $\omega^{>R}$. Let $\hat{u} = \text{col}[\gamma \cap \omega^{>R}]$ and write $\omega = \bigcap_{j \in J} N_j$. Recall that $R = \rho(\gamma) - 1$. Now,

$$(91) \quad \begin{aligned} \sum_j \bar{\alpha}_j \hat{u}[Y_j] &= \sum_{j \in \text{suppt}(\bar{\alpha})} \bar{\alpha}_j \hat{u}[Y_j] \\ &\geq \sum_{j \in J} \bar{\alpha}_j \hat{u}[Y_j]. \end{aligned}$$

Furthermore, in step (6.2), the algorithm imposes the constraint (59), applied to $\gamma \cap \omega^{>R}$, i.e., on the column $\hat{u}$. This amounts to imposing

$$(92) \quad \sum_{j \in J} \hat{u}[Y_j] - (1 + R) \hat{u}[\gamma \cap \omega^{>R}] \geq 0.$$

Now $1 + R = \rho(\gamma)$, by the definition of R . Also, by step (6.1), $\hat{u}[\gamma \cap \omega^{>R}] = \hat{u}[\mathcal{F}]$. So (92) is simply

$$(93) \quad \sum_{j \in J} \hat{u}[Y_j] - \rho(\gamma) \hat{u}[\mathcal{F}] \geq 0.$$

Thus, since $J \subseteq \text{suppt}(\bar{\alpha})$, combining (91) and (93), we obtain that

$$\sum_j \bar{\alpha}_j \hat{u}[Y_j]$$

is at least $\hat{u}[\mathcal{F}]$ times the sum of the $\rho(\gamma)$ smallest positive coefficients $\bar{\alpha}_j$. However, we know that $\pi(\bar{\alpha}, \bar{\alpha}_0) \leq \pi(\alpha, \alpha_0)$, which by assumption is at most $\rho(\gamma)$. Thus

$$\sum_j \bar{\alpha}_j \hat{u}[Y_j] \geq \bar{\alpha}_0 \hat{u}[\mathcal{F}],$$

as desired. This completes the proof. $\square$

COMMENT 4.27. *Note that positive-semidefiniteness (steps 7 and 8 of the algorithm) is not used in this theorem.*

4.6. Positive-semidefiniteness and set-packing problems. As discussed previously, positive-semidefiniteness arises as a natural feature in all of the lifting algorithms we described. An important question concerns how much positive-semidefiniteness provably adds to the strength of the algorithms. This question was taken up in [23], with special attention paid to vertex-packing problems. Among many other related results, it was shown in [23] that the N_+ -rank of clique inequalities is 1, whereas their N -rank is in general much higher, thus showing that positive-semidefiniteness can indeed help in a concrete way.

Before describing our results, we note that one can always rewrite a linear inequality in the form

$$\sum_{j \in J^+} a_j x_j + \sum_{j \in J^-} a_j (1 - x_j) \geq b,$$

where all the a_j are nonnegative and $J^+ \cap J^- = \emptyset$. In the context of the Σ -algorithms, this is a natural step since we introduce both variables Y_j and N_j for every j . By reformulating the problem in this manner, one obtains more general results than those we will describe here; see [30]. In particular, one can show how some high-pitch valid inequalities for set-covering problems are enforced by the Σ^k -algorithm for small k (we will touch on this again at the end of this section). However, for the sake of simplicity, in this section we will only explicitly deal with set-packing problems.

Consider the following packing system. Let I be an index set of cardinality at least three, and suppose that for each $i \in I$ we have a set $S_i \subseteq \{1, 2, \dots, n\}$ such that the S_i are pairwise disjoint. Consider the system of constraints

$$(94) \quad x(S_i) + x(S_j) \leq |S_i| + |S_j| - 1 \quad \forall i \neq j,$$

where, as before, we use the notation $v(H) = \sum_{j \in H} v_j$. Thus, if $|S_i| = 1$ for all i , we obtain a vertex-packing system for a t -clique. Write $S = \cup_i S_i$, $N = |S| = \sum_i |S_i|$, and $t = |I|$.

The inequality

$$(95) \quad \sum_{j \in S} x_j \leq N - t + 1$$

is valid: for $x \in \{0, 1\}^S$, if $\sum_{j \in S} x_j = N - t + 2$, then there are at least 2 distinct sets S_i such that $x_j = 1$ for all $j \in S_i$, violating (94). We will refer to (95) as the *set-clique* constraint.

LEMMA 4.28. *Consider a system of inequalities that contains, possibly among others, all constraints (94) for some family $\{S_i : i \in I\}$ of pairwise disjoint sets. Then the vector X produced by the Σ^2 -algorithm, using the positive-semidefiniteness condition 7, satisfies (95).*

Proof. For any $i \in I$, $\bigcap_{j \in S_i} Y_j$ is a tier enumerated by the algorithm and, by condition (57), we have

$$(96) \quad X \left[\bigcap_{j \in S_i} Y_j \right] \geq \sum_{j \in S_i} X[Y_j] - |S_i| + 1.$$

Let $\mu \in R^\mathcal{V}$ be defined by

$$(97) \quad \mu_\beta = \begin{cases} -1 & \text{if } \beta = \bigcap_{j \in S_i} Y_j \text{ for some } i \in I, \\ 1 & \text{if } \beta = \mathcal{F}, \\ 0 & \text{otherwise.} \end{cases}$$

Then $0 \leq \mu^T U \mu = 1 - \sum_{i \in I} X[\bigcap_{j \in S_i} Y_j]$. Together with (96), this implies the result. $\square$

In contrast to this result, we also have [10].

LEMMA 4.29. *Suppose $|S_i| = t$ for every $i \in I$. For $t \geq 3$ the N_+ -rank and the Sherali-Adams rank of the set-clique constraint are both at least $t - 2$.*

Proof. Note that the x vector has $n = t^2$ coordinates if the S_i comprise all the variables. Define

$$(98) \quad T = (t - 2)t^{t-1} + 1, \quad w_1 = \frac{t - 2}{t} T^{-1}, \quad \text{and} \quad w_2 = \frac{2}{(t - 1)t^{t-1}} T^{-1}.$$

Consider the function λ that assigns to each $v \in \{0, 1\}^n$ the value $\lambda(v)$ defined using the following rules.

- (a) Suppose $\sum_{j \in S_i} v_j = t - 1$ for $t - 1$ indices $i \in I$ and $\sum_{j \in S_i} v_j = t$ for the remaining index i . Then $\lambda(v) = w_1$.
- (b) Suppose $\sum_{j \in S_i} v_j = t - 1$ for $t - 2$ indices $i \in I$ and $\sum_{j \in S_i} v_j = t$ for each of the two remaining indices i . Then $\lambda(v) = w_2$.
- (c) In all other cases, we set $\lambda(v) = 0$.

Note that points of type (a) are feasible for system (94), while those of type (b) are not. Also, there are t^t points of type (a), and $\frac{t(t-1)}{2} t^{t-2} = \frac{(t-1)t^{t-1}}{2}$ points of type (b), and thus

$$(99) \quad \sum_{v \in \{0, 1\}^n} \lambda(v) = 1.$$

As a result, λ can be extended to a *measure* on $\{0, 1\}^n$ as follows: for a subset $A \subseteq \{0, 1\}^n$, define

$$\lambda(A) = \sum_{v \in A} \lambda(v).$$

Furthermore, if we define the vector $\bar{x} \in R^n$ by

$$\bar{x}_j = \lambda(\{v \in \{0, 1\}^n : v_j = 1\}) = \sum_{v_j=1} \lambda(v)$$

for $1 \leq j \leq n$, then a quick calculation shows that $\bar{x}$ satisfies all inequalities (94) but violates (95).

Furthermore, $\bar{x}$ can be lifted to a vector $\bar{y}$ indexed by the subset lattice of $\{1, 2, \dots, n\}$. For $p \subseteq \{1, 2, \dots, n\}$, set

$$\bar{y}_p = \sum \{\lambda(v) : v_j = 1 \ \forall j \in p\},$$

and a calculation shows that $\bar{y}$ (and the corresponding matrix $U^{\bar{y}}$) satisfy the requirements for both the rank $(t-3) N_+$ operator and the level $(t-3)$ Sherali–Adams operator. $\square$

This result can be strengthened; see [30]. In particular, when $|S_1| = |S_2| = \lfloor \frac{n}{3} \rfloor$ and $|S_i| = 1$ for all other i , then the N_+ and Sherali–Adams rank of (95) is $\lfloor \frac{n}{3} \rfloor$.

The above results can be generalized in a different way. Let G be an undirected graph. Suppose that to each vertex i we assign a nonempty set of 0-1 variables $x_k, k \in S_i$, where the index sets S_i are pairwise disjoint, and we impose the constraints

$$x(S_i) + x(S_j) \leq |S_i| + |S_j| - 1 \quad \forall \{i, j\} \in E(G).$$

This construction generalizes the standard vertex-packing polyhedron. A question of interest concerns the rank of the corresponding generalizations of classical valid inequalities. As we have shown above, the set-clique inequality has unbounded N_+ -rank but is guaranteed by the Σ^2 -algorithm. In a similar way one obtains, using the obvious notation (see [30] for details) and using the appropriate set-generalization of Lemma 1.5 of [23], the following.

LEMMA 4.30. *The Σ^2 -algorithm (resp., the Σ^3 -algorithm) generates a vector that satisfies all set-odd-hole (resp., all set-odd-antihole and set-odd-wheel) constraints.* $\square$

4.7. Further remarks. Consider the system $\mathcal{F}$ introduced in [13]:

$$(100) \quad \sum_{j \in S} x_j + \sum_{j \notin S} (1 - x_j) \geq \frac{1}{2} \quad \forall S \subseteq \{1, \dots, n\},$$

$$(101) \quad x \in \{0, 1\}^n.$$

This system was analyzed in [13], [6], [19], [24] to show that either n or $n-1$ iterations of various procedures (Sherali–Adams, the N_+ operator) are required to prove that $\mathcal{F}$ is empty.

However, it turns out that running the Σ^3 -algorithm already proves the same result [10].

THEOREM 4.31. *Applying the Σ^3 -algorithm to formulation (100) yields an infeasible linear system.* $\square$

Laurent [24] has conjectured that the Lasserre rank of (100)–(101) is $n-1$. See also [25].

One can also show that the inequality shown in [6] to have N_+ -rank n is implied by the Σ^2 -algorithm.

LEMMA 4.32. *Consider the region $\{x \in \{0, 1\}^n : \sum_j x_j \geq \frac{1}{2}\}$. The Σ^2 -algorithm proves $\sum_j x_j \geq 1$.*

Proof. This follows from constraint (54) in step 2 of the algorithm. $\square$

5. Future work. From the point of view of implementation, a simple idea would be to use the extended formulations for the purposes of separation only. Another idea

would be to apply column generation so as to implement the Σ^k -algorithm: even though (even for small values of k) the number of new variables and constraints may be too large to actually ask a linear programming solver to directly handle them, the number may be small enough that we can generate the formulation, at least in some implicit format. Column generation would then be used to select those variables and constraints that we actually want to incorporate into a working formulation.

Another topic that relates to this work concerns problems that are not, strictly speaking, integer programming formulations in the standard sense but are instead feasibility problems that are given by lists of forbidden configurations, i.e., obstructions. The Σ -algorithms can be directly adapted to handle such problems, and, more generally, to satisfiability problems; see [30]. There is a clear and well-known connection between set theory and logic, and, furthermore, there is a rich literature on the connection between logic and optimization, and on incorporating techniques of logic to integer programming. See, for example, [20]; also see [12]. This topic appears to be underlaid by Balas's work on disjunctive programming, but we have not found in the literature prior work relating logic to, in particular, the combination of the Lovász–Schrijver technique of matrix lifting and the use of variables indexed by the subset algebra.

Finally, a question of interest concerns the relationship between our cuts and Gomory cuts. This question has some substance since it is known that Gomory mixed-integer cuts are equivalent to general disjunctive cuts; see [16]. In [11] we show that the Σ -algorithms can be used to approximate, arbitrarily closely, the Chvátal–Gomory rank- r closure of a set-covering problem for any fixed $r > 0$.

Acknowledgment. We thank Monique Laurent and Bruce Shepherd for many useful comments. This paper was partially written at Bell Laboratories, whose support to us is gratefully acknowledged. We also thank three referees whose suggestions helped us improve the presentation.

REFERENCES

- [1] E. BALAS, *Disjunctive programs: Cutting planes from logical conditions*, in Nonlinear Programming 2, O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds., Academic Press, New York, 1975, pp. 279–312.
- [2] E. BALAS, *Disjunctive Programming*, Ann. Discrete Math., 5 (1979), pp. 3–51.
- [3] E. BALAS, S. CERIA, AND G. CORNUÉJOLS, *A lift-and-project cutting plane algorithm for mixed 0-1 programs*, Math. Program., 58 (1993), pp. 295–324.
- [4] E. BALAS, S. CERIA, AND G. CORNUÉJOLS, *Mixed 0-1 programming by lift-and-project in a branch-and-cut framework*, Management Sci., 42 (1996), pp. 1229–1246.
- [5] E. BALAS AND S. M. NG, *On the set covering polytope: I. All the facets with coefficients in $\{0, 1, 2\}$* , Math. Program., 45 (1989), pp. 1–20.
- [6] W. COOK AND S. DASH, *On the matrix-cut rank of polyhedra*, Math. Oper. Res., 26 (2001), pp. 19–30.
- [7] E. BALAS AND W. PULLEYBLANK, *The perfect matchable subgraph polytope of a bipartite graph*, Networks, 13 (1983), pp. 495–516.
- [8] E. BALAS AND W. PULLEYBLANK, *The perfect matchable subgraph polytope of an arbitrary graph*, Combinatorica, 9 (1989), pp. 321–327.
- [9] F. BARAHONA, *Reducing matching to polynomial size linear programming*, SIAM J. Optim., 3 (1993), pp. 688–695.
- [10] D. BIENSTOCK AND M. ZUCKERBERG, *Subset Algebra Lift Operators for 0-1 Integer Programming*, CORC report TR-2002-01, Computational Optimization Research Center, Columbia University, New York, 2002. Available online at www.corc.ieor.columbia.edu/reports/techreports.html (extended version).

- [11] D. BIENSTOCK AND M. ZUCKERBERG, *Approximate Fixed-Rank Closures of Covering Problems*, CORC report TR-2003-01, Computational Optimization Research Center, Columbia University, New York, 2003. Available online at www.corc.ieor.columbia.edu/reports/techreports.html.
- [12] E. BOROS AND P. L. HAMMER, *Pseudo-Boolean optimization*, Discrete Appl. Math., 123 (2002), pp. 155–225.
- [13] V. CHVÁTAL, W. COOK, AND M. HARTMANN, *On cutting-plane proofs in combinatorial optimization*, Linear Algebra Appl., 114 (1989), pp. 455–499.
- [14] L. COMTET, *Advanced Combinatorics: The Art of Finite and Infinite Expansions*, D. Reidel, Dordrecht, The Netherlands, 1974.
- [15] G. CORNUÉJOLS AND Y. LI, *On the rank of mixed 0,1 polyhedra*, Math. Program., 91 (2002), pp. 391–397.
- [16] G. CORNUÉJOLS AND Y. LI, *Elementary closures for integer programs*, Oper. Res. Lett., 28 (2001), pp. 1–8.
- [17] M. DEZA AND M. LAURENT, *Geometry of Cuts and Metrics*, Springer-Verlag, Berlin, 1997.
- [18] W. FELLER, *An Introduction to Probability Theory and Its Applications*, Wiley, New York, 1966.
- [19] M. X. GOEMANS AND L. TUNÇEL, *When does the positive semidefiniteness constraint help in lifting procedures?*, Math. Oper. Res., 26 (2001), pp. 796–815.
- [20] J. HOOKER, *Logic, optimization and constraint programming*, INFORMS J. Comput., 14 (2002), pp. 295–321.
- [21] U.-U. HAUS, M. KÖPPE, AND R. WEISMANTEL, *The integral basis method for integer programming*, Math. Methods Oper. Res., 3 (2001), pp. 353–361.
- [22] J. B. LASSERRE, *An explicit exact SDP relaxation for nonlinear 0-1 programs*, in Integer Programming and Combinatorial Optimization, Lecture Notes in Comput. Sci. 2081, K. Aardal and A. M. H. Gerards, eds., Springer, Berlin, 2001, pp. 293–303.
- [23] L. LOVÁSZ AND A. SCHRIJVER, *Cones of matrices and set-functions and 0–1 optimization*, SIAM J. Optim., 1 (1991), pp. 166–190.
- [24] M. LAURENT, *A comparison of the Sherali-Adams, Lovász-Schrijver and Lasserre relaxations for 0-1 programming*, Math. Oper. Res., 28 (2003), pp. 470–496.
- [25] M. LAURENT, *Lower bound for the number of iterations in semidefinite relaxations for the cut polytope*, Math. Oper. Res., 28 (2003), pp. 871–883.
- [26] A. N. LETCHFORD, *On disjunctive cuts for combinatorial optimization*, J. Comb. Optim., 5 (2001), pp. 299–315.
- [27] G. NEMHAUSER AND L. A. WOLSEY, *Integer and Combinatorial Optimization*, Wiley, New York, 1988.
- [28] G.-C. ROTA, *On the foundations of combinatorial theory I. Theory of Möbius functions*, Z. Wahrsch. Verw. Gebiete, 2 (1964), pp. 340–368.
- [29] H. D. SHERALI AND W. P. ADAMS, *A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems*, SIAM J. Discrete Math., 3 (1990), pp. 411–430.
- [30] M. ZUCKERBERG, *A Set Theoretic Approach to Lifting Procedures for 0,1 Integer Programming*, Ph.D. Thesis, Columbia University, New York, 2004.

UNIFORM BOUNDEDNESS OF A PRECONDITIONED NORMAL MATRIX USED IN INTERIOR-POINT METHODS*

RENATO D. C. MONTEIRO[†], JEROME W. O'NEAL[†], AND TAKASHI TSUCHIYA[‡]

Abstract. Solving systems of linear equations with “normal” matrices of the form AD^2A^T is a key ingredient in the computation of search directions for interior-point algorithms. In this article, we establish that a well-known basis preconditioner for such systems of linear equations produces scaled matrices with uniformly bounded condition numbers as D varies over the set of all positive diagonal matrices. In particular, we show that when A is the node–arc incidence matrix of a connected directed graph with one of its rows deleted, then the condition number of the corresponding preconditioned normal matrix is bounded above by $m(n - m + 1)$, where m and n are the number of nodes and arcs of the network.

Key words. linear programming, interior-point methods, polynomial bound, network flow problems, condition number, preconditioning, iterative methods for linear equations, normal matrix

AMS subject classifications. 65F35, 90C05, 90C35, 90C51

DOI. 10.1137/S1052623403426398

1. Introduction. Consider the linear programming (LP) problem $\min\{c^T x : Ax = b, x \geq 0\}$, where $A \in \mathbb{R}^{m \times n}$ has full row rank. Interior-point methods for solving this problem require that systems of linear equations of the form $AD^2A^T \Delta y = r$, where D is a positive diagonal matrix, be solved at every iteration. It often occurs that the “normal” matrix AD^2A^T , while positive definite, becomes increasingly ill-conditioned as one approaches optimality. In fact, it has been proven (e.g., see Kovacevic and Asic [3]) that for degenerate LP problems, the condition number of the normal matrix goes to infinity. Because of the ill-conditioned nature of AD^2A^T , many methods for solving the system $AD^2A^T \Delta y = r$ become increasingly unstable. The problem becomes even more serious when conjugate gradient methods are used to solve this linear system. Hence, the development of suitable preconditioners which keep the condition number of the coefficient matrix of the scaled system under control is of paramount importance. We should note, however, that in practice the ill-conditioning of AD^2A^T generally does not cause difficulty when the system $AD^2A^T \Delta y = r$ is solved using a backward-stable direct solver (see, e.g., [12, 13] and references therein).

In this paper, we analyze a preconditioner for the normal matrix AD^2A^T that has been proposed by Resende and Veiga [7] in the context of the minimum cost network flow problem and subsequently by Oliveira and Sorensen [6] for general LP problems. The preconditioning consists of pre- and postmultiplying AD^2A^T by $D_B^{-1}B^{-1}$ and its transpose, respectively, where B is a suitable basis of A and D_B is the corresponding diagonal submatrix of D . Roughly speaking, B is constructed in such a way that

*Received by the editors May 30, 2003; accepted for publication (in revised form) February 3, 2004; published electronically October 14, 2004.

<http://www.siam.org/journals/siopt/15-1/42639.html>

[†]School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332 (monteiro@isye.gatech.edu, joneal@isye.gatech.edu). The first author was supported in part by NSF grants CCR-9902010, CCR-0203113, and INT-9910084 and ONR grant N00014-03-1-0401.

[‡]The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-Ku, Tokyo, 106-8569, Japan (tsuchiya@sun312.ism.ac.jp). This author was supported in part by Japan-U.S. Joint Research Projects of the Japan Society for the Promotion of Science, “Algorithms for Linear Programs over Symmetric Cones.”

columns of A corresponding to larger diagonal elements of D have higher priority to be in B . Our main result is that such a preconditioner yields coefficient matrices with a uniformly bounded condition number regardless of the value of the diagonal elements of D . In the context of interior-point methods, this means that the condition number of the preconditioned normal matrix has a bound that does not depend on the current iterate, regardless of whether it is well centered. We also show that when applied to network flow-based LPs, the condition number of the preconditioned normal matrix is bounded by $m(n - m + 1)$, where m and n are the number of nodes and arcs of the network.

The bound on the condition number of the preconditioned normal matrix developed in this paper is given in terms of a well-known quantity, commonly denoted by $\bar{\chi}_A$, and defined as

$$\bar{\chi}_A \equiv \sup\{\|A^T(ADA^T)^{-1}AD\| : D \in \mathcal{D}_{++}\},$$

where $\mathcal{D}_{++}$ is the set of all positive diagonal matrices. Its finiteness was first established by Dikin [1]. Subsequently, $\bar{\chi}_A$ has been systematically studied by several authors including Stewart [8], Todd [9], and Todd, Tunçel, and Ye [10], and has also played a fundamental role in the analysis of interior-point algorithms (see, for example, Vavasis and Ye [11] and Monteiro and Tsuchiya [4, 5]).

1.1. Notation and terminology. The following notation is used throughout the paper. The superscript T denotes transpose. $\mathbb{R}^p$ denotes the p -dimensional Euclidean space. The set of all $p \times q$ matrices with real entries is denoted by $\mathbb{R}^{p \times q}$. The j th component of a vector w is denoted by w_j . The j th column and (i, j) th entry of a matrix F is denoted by F_j and F_{ij} , respectively. Given an ordered index set $\alpha \subseteq \{1, \dots, p\}$, a vector $w \in \mathbb{R}^p$, and a matrix F with p columns, we denote the subvector $(w_j : j \in \alpha)$ by w_α and the submatrix $[F_j : j \in \alpha]$ by F_α . For a matrix Q , we denote its largest and smallest eigenvalues by $\lambda_{\max}(Q)$ and $\lambda_{\min}(Q)$, respectively. The notation $\|\cdot\|$ denotes the vector Euclidean norm or matrix operator norm, respectively, depending on the context. The Frobenius norm of $Q \in \mathbb{R}^{p \times r}$ is given by $\|Q\|_F \equiv (\sum_{i=1}^p \sum_{j=1}^r Q_{ij}^2)^{1/2}$. For a vector d , $\text{Diag}(d)$ is the diagonal matrix having the elements of d on its diagonal.

2. The preconditioner and main results. In this section, we describe the preconditioner studied in this paper and establish its main properties.

We first describe a procedure which, given an $n \times n$ diagonal matrix $D \in \mathcal{D}_{++}$, finds a suitable basis of a full row rank matrix $A \in \mathbb{R}^{m \times n}$ obtained by giving higher priority to the columns of A with larger corresponding diagonal elements of D . The use of this basis as a way to obtain preconditioners of matrices of the form AD^2A^T was originally proposed by Resende and Veiga [7] in the context of minimum cost network flow problems and was subsequently extended by Oliveira and Sorensen [6] to the context of general LP problems. Note that when this procedure is used in the context of minimum cost network flow interior-point methods, it produces a maximum spanning tree for the network, whose arc weights are given by the diagonal elements of D .

ALGORITHM FOR DETERMINING BASIS B . Let a pair $(A, d) \in \mathbb{R}^{m \times n} \times \mathbb{R}_{++}^n$ be given such that $\text{rank}(A) = m$. Then:

1. Order the elements of d such that $d_1 \geq \dots \geq d_n$; order the columns of A accordingly.
2. Let $\mathcal{B} = \emptyset$ and set $l = 1$.

3. While $|\mathcal{B}| < m$, do
 - (i) If A_l is linearly independent of $\{A_j : j \in \mathcal{B}\}$, set $\mathcal{B} \leftarrow \mathcal{B} \cup \{l\}$.
 - (ii) $l \leftarrow l + 1$.
4. Let $\mathcal{N} = \{1, \dots, n\} \setminus \mathcal{B}$, $B = A_{\mathcal{B}}$, and $N = A_{\mathcal{N}}$.

We will refer to a basis B produced by the above scheme as a *maximum weight basis* associated with the pair $(A, d) \in \mathbb{R}^{m \times n} \times \mathbb{R}_{++}^n$. We begin with a technical, but important, lemma.

LEMMA 2.1. *Let $(A, d) \in \mathbb{R}^{m \times n} \times \mathbb{R}_{++}^n$ be given such that $\text{rank}(A) = m$. Suppose that B is a maximum weight basis associated with the pair (A, d) and define $D \equiv \text{Diag}(d)$ and $D_{\mathcal{B}} \equiv \text{Diag}(d_{\mathcal{B}})$. Then, for every $j = 1, \dots, n$,*

$$(1) \quad d_j \|D_{\mathcal{B}}^{-1} B^{-1} A_j\| \leq \|B^{-1} A_j\|.$$

As a consequence, we have

$$(2) \quad \|D_{\mathcal{B}}^{-1} B^{-1} A D\| \leq \|D_{\mathcal{B}}^{-1} B^{-1} A D\|_F \leq \|B^{-1} A\|_F \leq \sqrt{m} \|B^{-1} A\|.$$

Proof. We first prove (1). For every $j \in \mathcal{B}$, both sides of (1) are the same and hence (1) holds as an equality in this case. Assume now that $j \in \mathcal{N}$. We consider the following two distinct cases: (i) A_j was not considered to enter the basis B in step 3 of the above scheme, and (ii) A_j was a candidate to enter the basis but failed to make it. Consider first case (i). In this case, we have $d_j \leq \min(d_{\mathcal{B}})$ since the d_k 's are arranged in nonincreasing order at step 1 of the above scheme. Thus, we have

$$d_j \|D_{\mathcal{B}}^{-1} B^{-1} A_j\| \leq \frac{d_j}{\min(d_{\mathcal{B}})} \|B^{-1} A_j\| \leq \|B^{-1} A_j\|.$$

Consider now case (ii). Suppose that A_j was a candidate to become the r th column of B . Since it failed to enter B , it must be linearly dependent on the first $r - 1$ columns of B . Hence,

$$B^{-1} A_j = \begin{pmatrix} u \\ 0 \end{pmatrix}$$

for some $u \in \mathbb{R}^{r-1}$. Hence, using the fact that $d_{B_i} \geq d_j$ for every $i = 1, \dots, r - 1$, we conclude that

$$d_j \|D_{\mathcal{B}}^{-1} B^{-1} A_j\| \leq d_j \left(\sum_{i=1}^{r-1} \frac{u_i^2}{d_{B_i}^2} \right)^{1/2} \leq \left(\sum_{i=1}^{r-1} u_i^2 \right)^{1/2} = \|u\| = \|B^{-1} A_j\|.$$

We next prove (2). The first and third inequalities of (2) are well known. The second inequality follows from (1), the identity $\|R\|_F^2 = \sum_{j=1}^n \|R_j\|^2$ for every $R \in \mathbb{R}^{m \times n}$, and the fact that the j th column of $D_{\mathcal{B}}^{-1} B^{-1} A D$ is $d_j D_{\mathcal{B}}^{-1} B^{-1} A_j$. $\square$

Given a pair $(A, d) \in \mathbb{R}^{m \times n} \times \mathbb{R}_{++}^n$ such that $\text{rank}(A) = m$, we next consider a preconditioner for a system of equations of the form $AD^2 A^T p = r$, where $D \equiv \text{Diag}(d)$. Pre- and postmultiplying its coefficient matrix by an invertible matrix $R \in \mathbb{R}^{m \times m}$, we obtain the equivalent system

$$R(AD^2 A^T) R^T \tilde{p} = Rr,$$

where $\tilde{p} = R^{-T} p$. The following results give a suitable choice of R for which the condition number of the coefficient matrix of the above system is uniformly bounded as d varies over $\mathbb{R}_{++}^n$.

LEMMA 2.2. Let $(A, d) \in \mathbb{R}^{m \times n} \times \mathbb{R}_{++}^n$ be given such that $\text{rank}(A) = m$. Suppose that B is a maximum weight basis associated with the pair (A, d) and define $D \equiv \text{Diag}(d)$, $D_{\mathcal{B}} \equiv \text{Diag}(d_{\mathcal{B}})$, and $R \equiv D_{\mathcal{B}}^{-1}B^{-1}$. Then

$$\text{cond}(RAD^2A^TR^T) \leq \|B^{-1}A\|_F^2 \leq m \|B^{-1}A\|^2.$$

Proof. By Lemma 2.1, we have $\|RAD\| \leq \|B^{-1}A\|_F$. Hence,

$$\lambda_{\max}(RAD^2A^TR^T) = \|RAD\|^2 \leq \|B^{-1}A\|_F^2 \leq m \|B^{-1}A\|^2.$$

Moreover, since

$$R(ADA^T)R^T = D_{\mathcal{B}}^{-1}B^{-1}(BD_{\mathcal{B}}^2B^T + ND_{\mathcal{N}}^2N^T)B^{-T}D_{\mathcal{B}}^{-1} = I + WW^T,$$

where $W \equiv D_{\mathcal{B}}^{-1}B^{-1}ND_{\mathcal{N}}$ and $D_{\mathcal{N}} \equiv \text{Diag}(d_{\mathcal{N}})$, we conclude that

$$\lambda_{\min}(RAD^2A^TR^T) \geq 1.$$

Hence, the lemma follows. $\square$

We will refer to the preconditioner R described in Lemma 2.2 as a *maximum weight basis* preconditioner. Finally, as shown by Todd, Tunçel, and Ye [10] and Vavasis and Ye [11], we have

$$\bar{\chi}_A = \max\{\|B^{-1}A\| : B \text{ is a basis of } A\}.$$

Thus we arrive at our main result.

THEOREM 2.3. Let $(A, d) \in \mathbb{R}^{m \times n} \times \mathbb{R}_{++}^n$ be given such that $\text{rank}(A) = m$. Suppose that B is a maximum weight basis associated with the pair (A, d) and define $D \equiv \text{Diag}(d)$, $D_{\mathcal{B}} \equiv \text{Diag}(d_{\mathcal{B}})$, and $R \equiv D_{\mathcal{B}}^{-1}B^{-1}$. Then

$$\text{cond}(RAD^2A^TR^T) \leq m \bar{\chi}_A^2.$$

Another important consequence of Lemma 2.2 is the following result.

THEOREM 2.4. Let $A \in \mathbb{R}^{m \times n}$ denote the node-arc incidence matrix of a connected directed graph with one of its rows deleted. Suppose that B is a maximum weight basis associated with the pair (A, d) for some $d \in \mathbb{R}_{++}^n$. Letting $D \equiv \text{Diag}(d)$, $D_{\mathcal{B}} \equiv \text{Diag}(d_{\mathcal{B}})$, and $R \equiv D_{\mathcal{B}}^{-1}B^{-1}$, we have

$$\text{cond}(RAD^2A^TR^T) \leq m(n - m + 1).$$

Proof. Using the structure of A , it is easy to see that $\|B^{-1}A\|_F^2 \leq m + (n - m)m = m(n - m + 1)$. The result now follows directly from Lemma 2.2. $\square$

3. Concluding remarks. As mentioned earlier, using a maximum weight basis preconditioner in the context of network flow problems yields a maximum spanning tree preconditioner first proposed by Resende and Veiga [7]. In such a case, Judice et al. [2] have attempted to show that the condition number of the preconditioned matrix is bounded. However, their proof is incomplete, in that it deals only with three out of four possible cases, the neglected case being the most difficult and interesting one. Our proof does not attempt to correct theirs; rather, it is based on an entirely different approach. Moreover, our approach also holds for general LP problems in standard form, and shows that the derived bounds on the condition number of the preconditioned normal matrices $R(AD^2A^T)R^T$ hold for any $D \in \mathcal{D}_{++}$. In the context

of interior-point methods, this means that the condition number of the preconditioned normal matrix has a bound that does not depend on the current iterate, regardless of whether it is well centered.

Certain computational issues arise from our analysis. When determining the maximum weight basis B , one must determine whether a set of columns in A is linearly independent. This process tends to be sensitive to roundoff errors. Once the set $\mathcal{B}$ is determined, the matrix R can be computed in a stable fashion, since the condition numbers for all possible bases B are uniformly bounded and multiplication by the diagonal matrix D_B^{-1} can be done in a componentwise manner.

The authors believe that the preconditioner studied in this paper will be computationally effective only for some special types of LP problems. For example, the papers [2, 7] developed effective iterative interior-point methods for solving the minimum cost network flow problem based on maximum spanning tree preconditioners. Another class of LP problems for which iterative interior-point methods based on maximum weight basis preconditioners might be useful are those for which bases of A are sparse but the normal matrices AD^2A^T are dense; this situation generally arises in the context of LP problems for which n is much larger than m . Investigation of the many classes of LP problems which would benefit from the application of such methods is certainly an important area for future research.

Acknowledgments. The authors would like to thank the two anonymous referees and the associate editor, Margaret Wright, for their insightful comments, which have helped to substantially improve the presentation of this paper.

REFERENCES

- [1] I. I. DIKIN AND V. I. ZORKALCEV, *Iterative Solution of Mathematical Programming Problems: Algorithms for the Method of Interior Points*, Nauka, Novosibirsk, USSR, 1980 (in Russian).
- [2] J. J. JUDICE, J. PATRICIO, L. F. PORTUGAL, M. G. C. RESENDE, AND G. VEIGA, *A study of preconditioners for network interior point methods*, Comput. Optim. Appl., 24 (2003), pp. 5–35.
- [3] V. V. KOVACEVIC-VUJCIC AND M. D. ASIC, *Stabilization of interior-point methods for linear programming*, Comput. Optim. Appl., 14 (1999), pp. 331–346.
- [4] R. D. C. MONTEIRO AND T. TSUCHIYA, *A variant of the Vavasis–Ye layered-step interior-point algorithm for linear programming*, SIAM J. Optim., 13 (2003), pp. 1054–1079.
- [5] R. D. C. MONTEIRO AND T. TSUCHIYA, *A new iteration-complexity bound for the MTY predictor-corrector algorithm*, SIAM J. Optim., to appear.
- [6] A. R. L. OLIVEIRA AND D. C. SORENSEN, *Computational Experience with a Preconditioner for Interior Point Methods for Linear Programming*, Technical report, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 1997.
- [7] M. G. C. RESENDE AND G. VEIGA, *An implementation of the dual affine scaling algorithm for minimum-cost flow on bipartite uncappeditated networks*, SIAM J. Optim., 3 (1993), pp. 516–537.
- [8] G. W. STEWART, *On scaled projections and pseudoinverses*, Linear Algebra Appl., 112 (1989), pp. 189–193.
- [9] M. J. TODD, *A Dantzig-Wolfe-like variant of Karmarkar’s interior-point linear programming algorithm*, Oper. Res., 38 (1990), pp. 1006–1018.
- [10] M. J. TODD, L. TUNÇEL, AND Y. YE, *Probabilistic analysis of two complexity measures for linear programming problems*, Math. Program., 90 (2001), pp. 59–69.
- [11] S. A. VAVASIS AND Y. YE, *A primal-dual interior point method whose running time depends only on the constraint matrix*, Math. Program., 74 (1996), pp. 79–120.
- [12] M. H. WRIGHT, *Ill-conditioning and computational error in interior methods for nonlinear programming*, SIAM J. Optim., 9 (1998), pp. 84–111.
- [13] S. J. WRIGHT, *Effects of finite-precision arithmetic on interior-point methods for nonlinear programming*, SIAM J. Optim., 12 (2001), pp. 36–78.

A COMPARATIVE STUDY OF KERNEL FUNCTIONS FOR PRIMAL-DUAL INTERIOR-POINT ALGORITHMS IN LINEAR OPTIMIZATION*

Y. Q. BAI[†], M. EL GHAMI[‡], AND C. ROOS[‡]

Abstract. Recently, so-called self-regular barrier functions for primal-dual interior-point methods (IPMs) for linear optimization were introduced. Each such barrier function is determined by its (univariate) self-regular kernel function. We introduce a new class of kernel functions. The class is defined by some simple conditions on the kernel function and its derivatives. These properties enable us to derive many new and tight estimates that greatly simplify the analysis of IPMs based on these kernel functions. In both the algorithm and its analysis we use a single neighborhood of the central path; the neighborhood naturally depends on the kernel function. An important conclusion is that inverse functions of suitable restrictions of the kernel function and its first derivative more or less determine the behavior of the corresponding IPMs. Based on the new estimates we present a simple and unified computational scheme for the complexity analysis of kernel function in the new class. We apply this scheme to seven specific kernel functions. Some of these functions are self-regular, and others are not. One of the functions differs from the others, and from all self-regular functions, in the sense that its growth term is linear. Iteration bounds for both large- and small-update methods are derived. It is shown that small-update methods based on the new kernel functions all have the same complexity as the classical primal-dual IPM, namely, $O(\sqrt{n} \log \frac{n}{\varepsilon})$. For large-update methods the best obtained bound is $O(\sqrt{n}(\log n) \log \frac{n}{\varepsilon})$, which until now has been the best known bound for such methods.

Key words. linear optimization, interior-point method, primal-dual method, large-update method, polynomial complexity

AMS subject classifications. 90C05, 90C51

DOI. 10.1137/S1052623403423114

1. Introduction. After the groundbreaking paper of Karmarkar [5], linear optimization (LO) became an active area of research. The resulting interior-point methods (IPMs) are now among the most effective methods for solving LO problems. For a survey we refer to recent books on the subject [14, 16, 17]. In this paper we deal with so-called primal-dual IPMs. It is generally agreed that these IPMs are most efficient from a computational point of view (see, e.g. Andersen et al. [1]).

Until now primal-dual IPMs all have used the Newton direction as the search direction; this direction is closely related to the well-known primal-dual logarithmic barrier function. For a discussion of this relation we refer the reader to, e.g., [2, 3, 11]. There is a gap between the practical behavior of these algorithms and the theoretical performance results, especially for so-called large-update methods. If n denotes the number of inequalities in the problem, then the theoretical iteration bound is $O(n \log(n/\varepsilon))$, where ε represents the desired accuracy of the solution. In practice, large-update methods are much more efficient than the so-called small-update methods for which the theoretical iteration bound is only $O(\sqrt{n} \log(n/\varepsilon))$. So the current

*Received by the editors February 14, 2003; accepted for publication (in revised form) February 2, 2004; published electronically October 14, 2004. This research was supported by the Dutch Organization for Scientific Research (NWO grant 613.000.110).

<http://www.siam.org/journals/siopt/15-1/42311.html>

[†]Department of Mathematics, Shanghai University, Shanghai 200436, China (yqbai@staff.shu.edu.cn).

[‡]Faculty of Information Technology and Systems, Delft University of Technology, P.O. Box 5031, 2600 GA Delft, The Netherlands (M.Elghami@ewi.tudelft.nl, C.Roos@ewi.tudelft.nl, <http://www.isa.tudelft.nl/~roos>).

theoretical bounds differ by a factor of $\sqrt{n}$, in favor of the small-update methods. This significant gap between theory and practice has been referred to as the irony of IPMs [13, page 51].

Recently a new class of IPMs was introduced and the aforementioned gap could be narrowed up to a factor of $\log n$. These methods do not use the classical Newton direction. Instead they use a direction that can be characterized as a steepest descent direction (in a scaled space) for a so-called *self-regular* barrier function [11, 12]. Any such barrier function is determined by a simple univariate self-regular function, called its *kernel function*. The prototype self-regular kernel function in [11, 12] is given by

$$(1.1) \quad \Upsilon_{p,q}(t) = \frac{t^{p+1} - 1}{p(p+1)} + \frac{t^{1-q} - 1}{q(q-1)} + \frac{p-q}{pq}(t-1),$$

where $p \geq 1$ and $q > 1$. The parameter p is called the *growth degree*, and q the *barrier degree* of the kernel function. It may be noted that the iteration bounds obtained in [11, 12] depend monotonically on p : the smaller p is, the better is the iteration bound. Hence it is stated in [12, page 63] that in practical implementations small values of p should be used.

In this paper we introduce a new class of kernel functions which are not necessarily self-regular; on the other hand, all self-regular kernel functions $\Upsilon_{p,q}(t)$ with growth degree $p \leq 1$ belong to the new class. We develop some new analytic tools for the analysis of IPMs based on kernel functions from the new class. As a result the analysis in this paper is much simpler than in [11, 12], whereas the iteration bounds are at least as good. In addition, our analysis also applies to (self-regular) functions with growth degree $p \leq 1$. In both the algorithm and its analysis we use a single neighborhood of the central path; the neighborhood naturally depends on the kernel function.

The paper is organized as follows. In section 2.1 we first briefly recall the notion of the central path, which is a basic concept underlying all primal-dual IPMs. In section 2.2 we describe the idea underlying the approach of the paper. A crucial observation is that any function that is strictly convex on the positive orthant and that attains its minimal value in the all-one vector $\mathbf{e}$ determines in a natural way a primal-dual interior-point method, which is formally described in section 2.3.

In section 2.4 we define the notion of a (univariate) kernel function and in section 2.5 we introduce the class of kernel functions considered in this paper. The first property shared by kernel functions in this class is *exponential convexity*. The other properties control the growth and barrier behavior of the kernel function. Section 2.6 introduces seven specific kernel functions belonging to the new class.

In section 3 we study the growth behavior of the new barrier functions. Section 4 is devoted to the analysis of the amount of decrease of the barrier function during an inner iteration. This analysis yields a default step size and, in section 5, it yields the iteration complexity of the algorithm in terms of the kernel function and its first and second derivative. It also turns out that this complexity depends on the inverse function of (a restriction of) the kernel function itself and on the inverse function of (a restriction of) its first derivative. In section 6 we apply our results to seven kernel functions in order to obtain iteration bounds for large-update methods; to make clear that the small-update methods based on the new kernel functions are as good as the classical logarithmic barrier method we also derive iteration bounds for small-update methods. Finally, section 7 contains some concluding remarks and directions for future research.

We use the following notational conventions. If $x, s \in \mathbb{R}^n$, then xs denotes the coordinatewise (or Hadamard) product of the vectors x and s . Furthermore, $\mathbf{e}$ denotes

the all-one vector of length n . The nonnegative orthant and positive orthant are denoted as $\mathbb{R}_+^n$ and $\mathbb{R}_{++}^n$, respectively. Finally, if $z \in \mathbb{R}_+^n$ and $f : \mathbb{R}_+ \rightarrow \mathbb{R}_+$, then $f(z)$ denotes the vector in $\mathbb{R}_+^n$ whose i th component is $f(z_i)$, with $1 \leq i \leq n$. We write $f(x) = O(g(x))$ if $f(x) \leq cg(x)$ for some positive constant c and $f(x) = \Theta(g(x))$ if $c_1g(x) \leq f(x) \leq c_2g(x)$ for positive constants c_1 and c_2 . As in this paragraph, functions like f and g always have a local meaning in this paper.

2. Preliminaries.

2.1. The central path. We deal with the LO problem in standard format

$$(P) \quad \min\{c^T x : Ax = b, x \geq 0\},$$

where $A \in \mathbb{R}^{m \times n}$, $\text{rank}(A) = m$, $b \in \mathbb{R}^m$, and $c \in \mathbb{R}^n$, and its dual problem

$$(D) \quad \max\{b^T y : A^T y + s = c, s \geq 0\}.$$

It is well known that finding an optimal solution of (P) and (D) is equivalent to solving the following system:

$$(2.1) \quad \begin{aligned} Ax &= b, & x &\geq 0, \\ A^T y + s &= c, & s &\geq 0, \\ xs &= 0. \end{aligned}$$

The basic idea of primal-dual IPMs is to replace the third equation in (2.1), the so-called *complementarity condition* for (P) and (D) , by the parameterized equation $xs = \mu e$, with $\mu > 0$. Thus we consider the system

$$(2.2) \quad \begin{aligned} Ax &= b, & x &\geq 0, \\ A^T y + s &= c, & s &\geq 0, \\ xs &= \mu e. \end{aligned}$$

Due to the last equation, any solution (x, y, s) of (2.2) will satisfy $x > 0$ and $s > 0$. So a solution exists only if (P) and (D) satisfy the *interior-point condition* (IPC), i.e., there exists (x^0, s^0, y^0) such that

$$(2.3) \quad Ax^0 = b, \quad x^0 > 0, \quad A^T y^0 + s^0 = c, \quad s^0 > 0.$$

Surprisingly enough, if the IPC is satisfied, then a solution exists, for each $\mu > 0$, and this solution is unique. It is denoted as $(x(\mu), y(\mu), s(\mu))$ and we call $x(\mu)$ the μ -center of (P) and $(y(\mu), s(\mu))$ the μ -center of (D) . The set of μ -centers (with μ running through all positive real numbers) gives a homotopy path, which is called *the central path* of (P) and (D) . The relevance of the central path for LO was recognized first by Sonnevend [15] and Megiddo [6]. If $\mu \rightarrow 0$, then the limit of the central path exists and since the limit points satisfy the complementarity condition, the limit yields optimal solutions for (P) and (D) .

From a theoretical point of view the IPC can be assumed without loss of generality. In fact we may, and will, assume that $x^0 = s^0 = e$. In practice, this can be realized by embedding the given problems (P) and (D) into a homogeneous self-dual problem which has two additional variables and two additional constraints. For this and the other properties mentioned above, see, e.g., [14].

2.2. A wide class of primal-dual path-following methods. IPMs follow the central path approximately. We now describe how this proceeds. We assume that we are given a strictly convex differentiable function $\Psi(v)$, $v \in \mathbb{R}_{++}^n$, such that $\Psi(v)$ is minimal at $v = e$ and $\Psi(e) = 0$. We will argue below that any such function determines in a natural way an IPM.

Without loss of generality we assume that a good approximation of $(x(\mu), y(\mu), s(\mu))$ is known for some positive μ . For example, due to the above assumption we may assume this for $\mu = 1$, with $x(1) = s(1) = e$. We then decrease μ to $\mu := (1 - \theta)\mu$, for some fixed $\theta \in (0, 1)$. At this stage we are given a primal feasible $x > 0$ and a dual feasible $s > 0$ and we aim at finding the new μ -centers $x(\mu)$, $y(\mu)$, and $s(\mu)$, the solution of the system (2.2). Defining

$$(2.4) \quad v := \sqrt{\frac{xs}{\mu}},$$

we observe that (2.2) can be solved by minimizing $\Psi(v)$, subject to the linear constraints in (2.2), since this function has its (unique) minimizer at $v = e$, which is equivalent to $xs = \mu e$.

In this paper we will choose $\Psi(v)$ in such a way that its Hessian at e is a positive multiple of the identity matrix: $\nabla^2 \Psi(e) = \rho I$ for some positive ρ . This implies that the positive multiple

$$(2.5) \quad -\frac{1}{\rho} \nabla \Psi(v)$$

of the steepest descent direction for minimizing Ψ is a good approximation, at least in the neighborhood of e , of the Newton direction and hence this is an excellent search direction for minimizing $\Psi(v)$.

For (small) displacements Δx and Δs , in the x - and s -space, respectively, let Δv be the resulting change in v . Then we have

$$\Delta v = \sqrt{\frac{(x + \Delta x)(s + \Delta s)}{\mu}} - \sqrt{\frac{xs}{\mu}} \approx \frac{x\Delta s + s\Delta x}{2\sqrt{\mu xs}} = \frac{x\Delta s + s\Delta x}{2\mu v}.$$

Hence, since we want this to be about equal to (2.5), we arrive at the equation

$$\frac{x\Delta s + s\Delta x}{2\mu v} = -\frac{1}{\rho} \nabla \Psi(v).$$

Due to the definition of v and up to the positive factor $2/\rho$, this can be rewritten as

$$s\Delta x + x\Delta s = -\mu v \nabla \Psi(v).$$

Thus, adding the conditions for primal feasibility of $x + \Delta x$ and dual feasibility of $s + \Delta s$, neglecting for the moment the nonnegativity constraints, we find the following linear system for the search directions Δx , Δs , and Δy :

$$(2.6) \quad \begin{aligned} A\Delta x &= 0, \\ A^T \Delta y + \Delta s &= 0, \\ s\Delta x + x\Delta s &= -\mu v \nabla \Psi(v). \end{aligned}$$

We proceed by showing that system (2.6) uniquely defines the search directions and that these vanish if and only if $\Psi(v) = 0$, which holds if and only if $v = e$, i.e., if and

only if $(x, y, s) = (x(\mu), y(\mu), s(\mu))$. To this end, using the vector v defined in (2.4), we introduce scaled versions of the displacements Δx and Δs as follows:

$$(2.7) \quad d_x := \frac{v\Delta x}{x}, \quad d_s := \frac{v\Delta s}{s}.$$

Now one easily checks that the system (2.6), which defines the search directions, can be rewritten as

$$(2.8) \quad \begin{aligned} \bar{A}d_x &= 0, \\ \frac{1}{\mu}\bar{A}^T\Delta y + d_s &= 0, \\ d_x + d_s &= -\nabla\Psi(v), \end{aligned}$$

where

$$\bar{A} = AV^{-1}X, \quad V = \text{diag}(v), \quad X = \text{diag}(x).$$

The last equation in (2.8) is called the *scaled centering equation*. It states that the sum of the scaled search directions d_x and d_s is equal to $-\nabla\Psi(v)$, the steepest descent direction of $\Psi(v)$. Note that d_x and d_s are orthogonal vectors, since the vector d_x belongs to the null space and d_s to the row space of the matrix $\bar{A}$. Thus it follows that d_x and d_s form an orthogonal decomposition of the steepest descent direction of the function $\Psi(v)$. Note that since d_x and d_s are orthogonal, we will have

$$d_x = d_s = 0 \Leftrightarrow \nabla\Psi(v) = 0 \Leftrightarrow v = e \Leftrightarrow \Psi(v) = 0,$$

i.e., if and only if $x = x(\mu)$ and $s = s(\mu)$. Hence, if $(x, y, s) \neq (x(\mu), y(\mu), s(\mu))$, then $(\Delta x, \Delta s, \Delta y)$ is nonzero.

By taking a step along the search direction, with the step size α defined by some line search rules, one constructs a new triple (x, y, s) according to

$$(2.9) \quad x_+ = x + \alpha\Delta x, \quad y_+ = y + \alpha\Delta y, \quad s_+ = s + \alpha\Delta s.$$

If necessary, this procedure is repeated until we find iterates (x, y, s) that are “close” enough to $(x(\mu), y(\mu), s(\mu))$. Then μ is again reduced by the factor $1 - \theta$ and we apply the above method targeting the new μ -centers, and so on. This process is repeated until μ is small enough, say, until $n\mu \leq \varepsilon$; at this stage we have found an ε -solution of the problems (P) and (D) .

2.3. A generic primal-dual algorithm. We can now describe the algorithm in a more formal way. The generic form of the algorithm is shown in Figure 1. It is clear from this description that closeness of (x, y, s) to $(x(\mu), y(\mu), s(\mu))$ is measured by the value of $\Psi(v)$, with τ as a threshold value: if $\Psi(v) \leq \tau$, then we start a new *outer iteration* by performing a μ -update; otherwise we enter an *inner iteration* by computing the search directions at the current iterates with respect to the current value of μ and apply (2.9) to get new iterates.

The parameters τ , θ and the step size α should be chosen in such a way that the algorithm is “optimized” in the sense that the number of iterations required by the algorithm is as small as possible. The choice of the *barrier update parameter* θ plays an important role in both theory and practice of IPMs. Usually, if θ is a constant independent of the dimension n of the problem, for instance, $\theta = \frac{1}{2}$, then we call the algorithm

Generic Primal-Dual Algorithm for LO

Input:

A threshold parameter $\tau > 0$;
 an accuracy parameter $\varepsilon > 0$;
 a fixed barrier update parameter $\theta, 0 < \theta < 1$;

begin

$x := \mathbf{e}; s := \mathbf{e}; \mu := 1$;

while $n\mu \geq \varepsilon$ **do**

begin

$\mu := (1 - \theta)\mu$;

while $\Psi(v) > \tau$ **do**

begin

$x := x + \alpha \Delta x$;

$s := s + \alpha \Delta s$;

$y := y + \alpha \Delta y$;

$v := \sqrt{\frac{xs}{\mu}}$;

end

end

end

FIG. 1. *Generic algorithm.*

a *large-update* (or *long-step*) method. If θ depends on the dimension of the problem, such as $\theta = 1/\sqrt{n}$, then the algorithm is called a *small-update* (or *short-step*) method.

The choice of the step size α ($\alpha > 0$) is another crucial issue in the analysis of the algorithm. It has to be made such that the closeness of the iterates to the current μ -center improves by a sufficient amount. In the theoretical analysis the step size α is usually given a value that depends on the closeness of the current iterates to the μ -center.

2.4. Kernel functions. To simplify matters we will restrict ourselves in this paper to the case where $\Psi(v)$ is separable with identical coordinate functions. Thus, letting ψ denote the function on the coordinates, we have

$$(2.10) \quad \Psi(v) = \sum_{i=1}^n \psi(v_i),$$

where $\psi(t) : D \rightarrow \mathbb{R}_+$, with $\mathbb{R}_{++} \subseteq D$, is strictly convex and minimal at $t = 1$, with $\psi(1) = 0$. We call the univariate function $\psi(t)$ the *kernel function* of the barrier function $\Psi(v)$. Obviously, the resulting iteration bound will depend on the kernel function underlying the algorithm, and our main task becomes to find a kernel function that minimizes the iteration bound.

Except for the kernel function considered in [3], all kernel functions considered so far are twice differentiable and go to infinity if either $t \downarrow 0$ or $t \rightarrow \infty$. Thus they satisfy

$$(2.11a) \quad \psi'(1) = \psi(1) = 0,$$

$$(2.11b) \quad \psi''(t) > 0,$$

$$(2.11c) \quad \lim_{t \downarrow 0} \psi(t) = \lim_{t \rightarrow \infty} \psi(t) = \infty.$$

Clearly, (2.11a) and (2.11b) say that $\psi(t)$ is a nonnegative strictly convex function such that $\psi(1) = 0$. Note that this implies that if $\psi(t)$ is twice differentiable, then it is completely determined by its second derivative:

$$(2.12) \quad \psi(t) = \int_1^t \int_1^\xi \psi''(\zeta) d\zeta d\xi.$$

Moreover, (2.11c) expresses that $\psi(t)$ is coercive and has the barrier property. Having such a kernel function $\psi(t)$, its definition is extended to positive n -dimensional vectors v by (2.10), thus yielding the induced (scaled) barrier function $\Psi(v)$. In what follows we assume in this paper that a kernel function satisfies (2.11a)–(2.11c).

As we indicated in the previous section, $\Psi(v)$ not only serves to define a search direction, but also acts as a measure of closeness of the current iterates to the μ -center. In the analysis of the algorithm we also use the *norm-based proximity measure* $\delta(v)$ defined by

$$(2.13) \quad \delta(v) := \frac{1}{2} \|\nabla \Psi(v)\| = \frac{1}{2} \|d_x + d_s\|.$$

Note that since $\Psi(v)$ is strictly convex and minimal at $v = e$ we have

$$\Psi(v) = 0 \Leftrightarrow \delta(v) = 0 \Leftrightarrow v = e.$$

Thus the algorithm considered in this paper uses the barrier function $\Psi(v)$ to measure closeness of the iterates to the μ -center; as will become clear below, in the analysis of the algorithm $\delta(v)$ serves as a second proximity measure. Both measures are naturally determined by the kernel function.

2.5. Further conditions on the kernel function. In this paper we work with five more conditions on the kernel function, namely,

$$(2.14a) \quad t\psi''(t) + \psi'(t) > 0, \quad t < 1,$$

$$(2.14b) \quad t\psi''(t) - \psi'(t) > 0, \quad t > 1,$$

$$(2.14c) \quad \psi'''(t) < 0, \quad t > 0,$$

$$(2.14d) \quad 2\psi''(t)^2 - \psi'(t)\psi'''(t) > 0, \quad t < 1,$$

$$(2.14e) \quad \psi''(t)\psi'(\beta t) - \beta\psi'(t)\psi''(\beta t) > 0, \quad t > 1, \beta > 1.$$

Note that conditions (2.14c) and (2.14d) require that $\psi(t)$ be three times differentiable. Furthermore, condition (2.14a) is obviously satisfied if $t \geq 1$, since then $\psi'(t) \geq 0$ and, similarly, condition (2.14b) is satisfied if $t \leq 1$, since then $\psi'(t) \leq 0$. Also (2.14d) is obviously satisfied if $t \geq 1$ since then $\psi'(t) \geq 0$, whereas $\psi'''(t) < 0$. We conclude that conditions (2.14a) and (2.14d) are conditions on the barrier behavior of $\psi(t)$. On the other hand, condition (2.14b) deals only with $t \geq 1$ and hence concerns the growth behavior of $\psi(t)$. Condition (2.14e) is technically more involved; we will discuss it later.

The next two lemmas make clear that conditions (2.14a) and (2.14b) admit a nice interpretation. The proof of the first lemma can be found in [12]; the proof of the second lemma is of the same spirit and is left to the reader.

LEMMA 2.1 (Lemma 2.1.2 in [12]). *Let $\psi(t)$ be a twice differentiable function for $t > 0$. Then the following three properties are equivalent:*

- (i) $\psi(\sqrt{t_1 t_2}) \leq \frac{1}{2} (\psi(t_1) + \psi(t_2))$ for $t_1, t_2 > 0$;

- (ii) $\psi'(t) + t\psi''(t) \geq 0$, $t > 0$;
- (iii) $\psi(e^\xi)$ is convex.

LEMMA 2.2. *Let $\psi(t)$ be a twice differentiable function for $t > 0$. Then the following three properties are equivalent:*

- (i) $\psi(\sqrt{\frac{t_1^2+t_2^2}{2}}) \leq \frac{1}{2}(\psi(t_1) + \psi(t_2))$;
- (ii) $t\psi''(t) - \psi'(t) \geq 0$, $t > 0$;
- (iii) $\psi(\sqrt{\xi})$ is convex.

Following [8], we call the property described in Lemma 2.1 *exponential convexity*, or *e-convexity* for short. This property has been proven to be very useful in the analysis of primal-dual algorithms based on kernel functions (cf. [10, 11, 12]). We recall from [12, Lemma 2.1.3] that if $\psi(t)$ is *e-convex*, then $\psi(t^\alpha)$ is *e-convex* as well, for every $\alpha \in \mathbb{R}$. Also, for any $\beta_0 \in \mathbb{R}$ the function

$$\beta_0 \log t + \sum_{i=1}^m \beta_i (t^{\alpha_i} - 1)$$

is *e-convex* whenever $\beta_i \geq 0$ and $\alpha_i \in \mathbb{R}$ for $i = 1, \dots, m$ [12, Proposition 2.1.4].

It may be noted that the kernel function $\psi(t)$ is defined to be self-regular in [11, 12] if $\psi(t)$ is *e-convex* and moreover

$$\psi''(t) = \Theta(\Upsilon''_{p,q}(t)),$$

where $\Upsilon_{p,q}(t)$ denotes the prototype self-regular kernel function (1.1). Since

$$\Upsilon''_{p,q}(t) = t^{p-1} + t^{-q-1}, \quad \Upsilon'''_{p,q}(t) = (p-1)t^{p-2} - (q+1)t^{-q-2},$$

the kernel function $\Upsilon_{p,q}(t)$ (and hence each self-regular kernel function) satisfies (2.14c) only if $p \leq 1$.

Remark 2.3. It is worth pointing out that the conditions (2.14a)–(2.14d) are logically independent. The following table shows four kernel functions, and the signs indicate whether a condition is satisfied (+) or not (–).

$\psi(t)$	(2.14a)	(2.14b)	(2.14c)	(2.14d)
$t^2 - t - 1 + e^{1-t}$	–	+	+	+
$t - 1 - \log t$	+	–	+	+
$t^3 + t^{-3} - 2$	+	+	–	+
$8t^2 - 11t + 1 + \frac{2}{\sqrt{t}} - 4 \log t$	+	+	+	–

The last condition, however, is not independent from the other conditions. One has the following lemma.

LEMMA 2.4. *If $\psi(t)$ satisfies (2.14b) and (2.14c), then $\psi(t)$ satisfies (2.14e).*

Proof. For $t > 1$ we consider the function

$$f(\beta) := \psi''(t)\psi'(\beta t) - \beta\psi'(t)\psi''(\beta t), \quad \beta \geq 1.$$

Since $f(1) = 0$, the lemma will follow if $f'(\beta) > 0$ for $\beta > 1$. This can be shown as follows:

$$\begin{aligned} f'(\beta) &= t\psi''(t)\psi''(\beta t) - \psi'(t)\psi''(\beta t) - \beta t\psi'(t)\psi'''(\beta t) \\ &= \psi''(\beta t) [t\psi''(t) - \psi'(t)] - \beta t\psi'(t)\psi'''(\beta t) > 0. \end{aligned}$$

The last inequality follows since $\psi''(\beta t) \geq 0$, $t\psi''(t) - \psi'(t) \geq 0$, by (2.14b), and $-\beta t\psi'(t)\psi'''(\beta t) > 0$, since $t > 1$, which implies $\psi'(t) > 0$, and $\psi'''(\beta t) < 0$, by (2.14c). This proves the lemma. $\square$

The following two lemmas deal with consequences of condition (2.14c).

LEMMA 2.5. *If the kernel function $\psi(t)$ satisfies (2.14c), then*

$$\begin{aligned} \psi(t) &> \frac{1}{2}(t-1)\psi'(t) \quad \text{and} \quad \psi'(t) > (t-1)\psi''(t) \quad \text{if } t > 1, \\ \psi(t) &< \frac{1}{2}(t-1)\psi'(t) \quad \text{and} \quad \psi'(t) > (t-1)\psi''(t) \quad \text{if } t < 1. \end{aligned}$$

Proof. Consider the function $f(t) = 2\psi(t) - (t-1)\psi'(t)$. One has $f(1) = 0$ and $f'(t) = \psi'(t) - (t-1)\psi''(t)$. Hence $f'(1) = 0$ and $f''(t) = -\psi'''(t)$. Using that $\psi'''(t) < 0$ it follows that if $t > 1$, then $f''(t) > 0$, whence $f'(t) > 0$ and $f(t) > 0$, and if $t < 1$, then $f''(t) < 0$, whence $f'(t) > 0$ and $f(t) < 0$. From this the lemma follows. $\square$

LEMMA 2.6. *If the kernel function $\psi(t)$ satisfies (2.14c), then*

$$\begin{aligned} \frac{1}{2}\psi''(t)(t-1)^2 &< \psi(t) < \frac{1}{2}\psi''(1)(t-1)^2 \quad \text{if } t > 1, \\ \frac{1}{2}\psi''(1)(t-1)^2 &< \psi(t) < \frac{1}{2}\psi''(t)(t-1)^2 \quad \text{if } t < 1. \end{aligned}$$

Proof. By using Taylor's theorem and $\psi(1) = \psi'(1) = 0$, we obtain

$$\psi(t) = \frac{1}{2}\psi''(1)(t-1)^2 + \frac{1}{3!}\psi'''(\xi)(\xi-1)^3,$$

where $1 < \xi < t$ if $t > 1$ and $t < \xi < 1$ if $t < 1$. Since $\psi'''(\xi) < 0$ the second inequality for $t > 1$ and the first inequality for $t < 1$ in the lemma follow. The remaining two inequalities are an immediate consequence of Lemma 2.5. $\square$

The theory developed in this paper applies to kernel functions that satisfy conditions (2.14a), (2.14c), (2.14d), and (2.14e). In this paper we call any such function an *eligible kernel function*. Due to Lemma 2.4, any kernel function satisfying the first four conditions, (2.14a)–(2.14d), is eligible.

2.6. Seven kernel functions. By way of example we consider in this paper the seven kernel functions in the second column of Table 2.1. In this section we show that these kernel functions are eligible. For the first six functions this is almost immediate: they satisfy conditions (2.14a)–(2.14d), as we now show. The 5th and 6th columns in Table 2.1 make clear that each of these kernel functions satisfies conditions (2.14a) and (2.14b). From Table 2.2 it follows that they also satisfy conditions (2.14c) and (2.14d). It is obvious from the 2nd column that (2.14c) is satisfied. That (2.14d) also is satisfied is not always immediately clear from the given expressions in the 4th column in Table 2.2; we leave it as an exercise to the reader to verify that these expressions indeed are nonnegative for $t \leq 1$.

The seventh kernel function requires special attention. It is clear from the tables that it satisfies (2.14a), (2.14c), and (2.14d), but it fails to satisfy (2.14b). To prove that this function is eligible it suffices to show that it satisfies (2.14e). This, in fact, is easy. Some straightforward computations yield that for $\psi = \psi_7$ one has

$$\psi''(t)\psi'(\beta t) - \beta\psi'(t)\psi''(\beta t) = \frac{q(1 - \beta^{-q})}{t^{q+1}},$$

TABLE 2.1

The seven kernel functions and the conditions (2.14a) and (2.14b) ($q > 1$).

i	$\psi_i(t)$	$\psi'_i(t)$	$\psi''_i(t)$	$t\psi''_i(t) + \psi'_i(t)$	$t\psi''_i(t) - \psi'_i(t)$
1	$\frac{t^2-1}{2} - \log t$	$t - \frac{1}{t}$	$1 + \frac{1}{t^2}$	$2t$	$\frac{2}{t}$
2	$\frac{1}{2} \left(t - \frac{1}{t}\right)^2$	$t - \frac{1}{t^3}$	$1 + \frac{3}{t^4}$	$2t + \frac{2}{t^3}$	$\frac{4}{t^3}$
3	$\frac{t^2-1}{2} + \frac{t^{1-q}-1}{q-1}$	$t - t^{-q}$	$1 + qt^{-q-1}$	$2t + (q-1)t^{-q}$	$(q+1)t^{-q}$
4	$\frac{t^2-1}{2} + \frac{t^{1-q}-1}{q(q-1)} - \frac{q-1}{q}(t-1)$	$t - 1 - \frac{t^{-q}-1}{q}$	$1 + t^{-q-1}$	$2t + \frac{q-1}{q}(t^{-q}-1)$	$\frac{q-1}{q} + \frac{q+1}{q}t^{-q}$
5	$\frac{t^2-1}{2} + \frac{e^{\frac{1}{t}}-e}{e}$	$t - \frac{e^{\frac{1}{t}}-1}{t^2}$	$1 + \frac{1+2t}{t^4}e^{\frac{1}{t}-1}$	$2t + \frac{1+t}{t^3}e^{\frac{1}{t}-1}$	$\frac{1+3t}{t^3}e^{\frac{1}{t}-1}$
6	$\frac{t^2-1}{2} - \int_1^t e^{\frac{1}{\xi}-1} d\xi$	$t - e^{\frac{1}{t}-1}$	$1 + \frac{e^{\frac{1}{t}-1}}{t^2}$	$2t + \frac{1-t}{t}e^{\frac{1}{t}-1}$	$\frac{1+t}{t}e^{\frac{1}{t}-1}$
7	$t - 1 + \frac{t^{1-q}-1}{q-1}$	$1 - t^{-q}$	qt^{-q-1}	$1 + (q-1)t^{-q}$	$-1 + (q+1)t^{-q}$

TABLE 2.2

The seven kernel functions and the conditions (2.14c) and (2.14d).

i	$\psi_i(t)$	$\psi'''_i(t)$	$2\psi''_i(t)^2 - \psi'_i(t)\psi'''_i(t)$
1	$\frac{t^2-1}{2} - \log t$	$-\frac{2}{t^3}$	$2 + \frac{6}{t^2}$
2	$\frac{1}{2} \left(t - \frac{1}{t}\right)^2$	$-\frac{12}{t^5}$	$2 + \frac{24}{t^4} + \frac{6}{t^8}$
3	$\frac{t^2-1}{2} + \frac{t^{1-q}-1}{q-1}$	$-q(q+1)t^{-q-2}$	$2 \left(1 + \frac{q}{t^{q+1}}\right)^2 + \frac{q(q+1)(t^{q+1}-1)}{t^{2(q+1)}}$
4	$\frac{t^2-1}{2} + \frac{t^{1-q}-1}{q(q-1)} - \frac{q-1}{q}(t-1)$	$-(q+1)t^{-q-2}$	$2 \left(1 + \frac{1}{t^{q+1}}\right)^2 + \frac{(q+1) \left(t-1 + \frac{1-t^{-q}}{q}\right)}{t^{q+2}}$
5	$\frac{t^2-1}{2} + \frac{e^{\frac{1}{t}}-e}{e}$	$-\frac{1+6t+6t^2}{t^6}e^{\frac{1}{t}-1}$	$\frac{2 \left(t^4 + (1+2t)e^{\frac{1}{t}-1}\right)^2 - (1+6t+6t^2) \left(e^{\frac{1}{t}-1} - t^3\right) e^{\frac{1}{t}-1}}{t^8}$
6	$\frac{t^2-1}{2} - \int_1^t e^{\frac{1}{\xi}-1} d\xi$	$-\frac{1+2t}{t^4}e^{\frac{1}{t}-1}$	$2 \left(1 + \frac{e^{\frac{1}{t}-1}}{t^2}\right)^2 + (1+2t) \frac{e^{\frac{1}{t}-1} \left(t - e^{\frac{1}{t}-1}\right)}{t^4}$
7	$t - 1 + \frac{t^{1-q}-1}{q-1}$	$-q(q+1)t^{-q-2}$	$q(q-1 + (q+1)t^q)t^{-2(q+1)}$

which is clearly positive if $t > 1$ and $\beta > 1$.

Thus we have shown that each of the seven kernel functions is eligible. We now briefly discuss these functions. The first function, $\psi_1(t)$, is the kernel function of the classical primal-dual logarithmic barrier function. This can be seen as follows. The barrier function induced by $\psi(t) = \psi_1(t)$ is

$$\Psi(v) = \sum_{i=1}^n \psi(v_i) = \sum_{i=1}^n \left(\frac{v_i^2-1}{2} - \log v_i \right) = \frac{1}{2} \sum_{i=1}^n (v_i^2 - 1 - \log v_i^2).$$

Because of (2.4), by defining $\Phi(x, s, \mu) := \Psi(v)$ the barrier function can be expressed in x and s :

$$\Phi(x, s, \mu) = \frac{1}{2} \sum_{i=1}^n \left(\frac{x_i s_i}{\mu} - 1 - \log \frac{x_i s_i}{\mu} \right) = \frac{1}{2\mu} \left(x^T s - \mu \sum_{i=1}^n \log x_i s_i \right) + \frac{n \log \mu - n}{2}.$$

The expression between the second pair of brackets is the well-known primal-dual

logarithmic barrier function for (P) and (D) . It differs from $\Psi(v)$ only by some additive and multiplicative “constants,” depending on n and μ but not on x and s . Note that for any kernel function, the above definition of $\Phi(x, s, \mu)$ gives rise to a function that goes to infinity if for some i , $x_i s_i$ goes either to infinity or to 0, and hence $\Phi(x, s, \mu)$ is a barrier function for the feasible domain. It may be worth emphasizing that the search direction in this paper, as defined by (2.6), is obtained neither by solving a barrier problem, nor as a primal-dual Newton direction induced by $\Phi(x, s, \mu)$.

The second kernel function has been studied in [9]; one may easily verify that it is a special case of $\psi_3(t)$, by taking $q = 3$. The third one has been studied for general $q > 1$ in [8, 12]. It is in turn a special case of a so-called self-regular kernel function, as studied in [11, 12]. Also note that $\psi_1(t)$ is the limiting value of $\psi_3(t)$ when q approaches 1. The fourth kernel function is the special case of $\Upsilon_{p,q}(t)$ for $p = 1$. The fifth and sixth kernel functions are new, at least to the knowledge of the authors. In each of the first six cases we can write $\psi(t)$ as

$$(2.15) \quad \psi(t) = \frac{t^2 - 1}{2} + \psi_b(t),$$

where $\frac{t^2 - 1}{2}$ is the so-called *growth term* and $\psi_b(t)$ is the *barrier term* of the kernel function. The growth term dominates the behavior of $\psi(t)$ when t goes to infinity, whereas the barrier term dominates its behavior when t approaches zero. Note that in all cases the barrier term is monotonically decreasing in t .

The last kernel function ($\psi_7(t)$, with $q > 1$) differs from all the other kernel functions in that its growth term, i.e., $t - 1$, is linear in t . It was first introduced and analyzed in [4].

3. Growth behavior. Note that at the start of each outer iteration of the algorithm, just before the update of μ with the factor $1 - \theta$, we have $\Psi(v) \leq \tau$. Due to the update of μ the vector v is divided by the factor $\sqrt{1 - \theta}$, with $0 < \theta < 1$, which in general leads to an increase in the value of $\Psi(v)$. Then, during the subsequent inner iterations, $\Psi(v)$ decreases until it passes the threshold τ again. Hence, during the course of the algorithm the largest values of $\Psi(v)$ occur just after the updates of μ . That is why in this section we derive an estimate for the effect of a μ -update on the value of $\Psi(v)$. In other words, with $\beta = \frac{1}{\sqrt{1 - \theta}}$, we want to find an upper bound for $\Psi(\beta v)$ in terms of $\Psi(v)$. We start with the following lemma.

LEMMA 3.1. *Suppose that $\psi(t_1) = \psi(t_2)$, with $t_1 \leq 1 \leq t_2$ and $\beta \geq 1$. Then*

$$\psi(\beta t_1) \leq \psi(\beta t_2).$$

Equality holds if and only if $\beta = 1$ or $t_1 = t_2 = 1$.

Proof. Consider

$$f(\beta) = \psi(\beta t_2) - \psi(\beta t_1).$$

One has $f(1) = 0$ and

$$f'(\beta) = t_2 \psi'(\beta t_2) - t_1 \psi'(\beta t_1).$$

Since $\psi''(t) \geq 0$ for all $t > 0$, $\psi'(t)$ is monotonically increasing. Hence $\psi'(\beta t_2) \geq \psi'(\beta t_1)$. Substitution gives

$$f'(\beta) = t_2 \psi'(\beta t_2) - t_1 \psi'(\beta t_1) \geq t_2 \psi'(\beta t_2) - t_1 \psi'(\beta t_2) = \psi'(\beta t_2) (t_2 - t_1) \geq 0.$$

The last inequality holds since $t_2 \geq t_1$, and $\psi'(t) \geq 0$ for $t \geq 1$. This proves that $f(\beta) \geq 0$ for $\beta \geq 1$, and hence the inequality in the lemma follows. If $\beta = 1$, then we obviously have equality. Otherwise, if $\beta > 1$ and $f(\beta) = 0$, then the mean value theorem implies $f'(\xi) = 0$ for some $\xi \in (1, \beta)$. But this implies $\psi'(\xi t_2) = \psi'(\xi t_1)$. Since $\psi'(t)$ is strictly monotonic, this implies $\xi t_2 = \xi t_1$, whence $t_2 = t_1$. Since also $t_1 \leq 1 \leq t_2$, we obtain $t_2 = t_1 = 1$. This completes the proof. $\square$

We have the following result.

THEOREM 3.2. *Let $\varrho : [0, \infty) \rightarrow [1, \infty)$ be the inverse function of $\psi(t)$ for $t \geq 1$. Then we have for any positive vector v and any $\beta \geq 1$ that*

$$\Psi(\beta v) \leq n\psi\left(\beta\varrho\left(\frac{\Psi(v)}{n}\right)\right).$$

Proof. First we consider the case where $\beta > 1$. We consider the following maximization problem:

$$\max_v \{\Psi(\beta v) : \Psi(v) = z\},$$

where z is any nonnegative number. The first order optimality conditions for this problem are

$$(3.1) \quad \beta\psi'(\beta v_i) = \lambda\psi'(v_i), \quad i = 1, \dots, n,$$

where λ denotes the Lagrange multiplier. Since $\psi'(1) = 0$ and $\beta\psi'(\beta) > 0$, we must have $v_i \neq 1$ for all i . We even may assume that $v_i > 1$ for all i . To see this, let z_i be such that $\psi(v_i) = z_i$. Given z_i , this equation has two solutions: $v_i = v_i^{(1)} < 1$ and $v_i = v_i^{(2)} > 1$. As a consequence of Lemma 3.1 we have $\psi(\beta v_i^{(1)}) \leq \psi(\beta v_i^{(2)})$. Since we are maximizing $\Psi(\beta v)$, it follows that we may assume $v_i = v_i^{(2)} > 1$. Thus we have shown that without loss of generality we may assume that $v_i > 1$ for all i . Note that then (3.1) implies $\beta\psi'(\beta v_i) > 0$ and $\psi'(v_i) > 0$, whence also $\lambda > 0$. Now defining $g(t)$ according to

$$g(t) := \frac{\psi'(t)}{\psi'(\beta t)}, \quad t \geq 1,$$

we deduce from (3.1) that $g(v_i) = \frac{\beta}{\lambda}$ for all i . One has

$$g'(t) = \frac{\psi''(t)\psi'(\beta t) - \beta\psi'(t)\psi''(\beta t)}{(\psi'(\beta t))^2}.$$

At this stage we use that $\psi(t)$ satisfies condition (2.14e). Due to this we have $g'(t) > 0$ for $t > 1$ and $\beta > 1$. So $g(t)$ is strictly monotonically increasing. Hence it follows that all v_i 's are mutually equal. Putting $v_i = t > 1$, for all i , we deduce from $\Psi(v) = z$ that $n\psi(t) = z$. This implies $t = \varrho(\frac{z}{n})$. Hence the maximal value that $\Psi(v)$ can attain is given by

$$\Psi(\beta t) = n\psi(\beta t) = n\psi\left(\beta\varrho\left(\frac{z}{n}\right)\right) = n\psi\left(\beta\varrho\left(\frac{\Psi(v)}{n}\right)\right).$$

This proves the theorem if $\beta > 1$. For the case $\beta = 1$ it suffices to observe that both sides of the inequality in the theorem are continuous in β . $\square$

COROLLARY 3.3. *Using the notation of Theorem 3.2, we have*

$$\Psi(\beta v) \leq \frac{n}{2} \psi''(1) \left(\beta \varrho \left(\frac{\Psi(v)}{n} \right) - 1 \right)^2.$$

Proof. Since $\beta \geq 1$ and $\varrho(\frac{\Psi(v)}{n}) \geq 1$, the corollary follows from Theorem 3.2 by also using Lemma 2.6. $\square$

As a result we have that if $\Psi(v) \leq \tau$ and $\beta = \frac{1}{\sqrt{1-\theta}}$, then

$$(3.2) \quad L_\psi(n, \theta, \tau) := n\psi \left(\frac{\varrho \left(\frac{\tau}{n} \right)}{\sqrt{1-\theta}} \right)$$

is an upper bound for $\Psi(\frac{v}{\sqrt{1-\theta}})$, the value of $\Psi(v)$ after the μ -update.

Remark 3.4. Note that the bound of Theorem 3.2 is sharp: one may easily verify that if $v = \beta e$, with $\beta \geq 1$, then the bound holds with equality.

4. Decrease of the barrier function during an inner iteration.

4.1. Computation of the step size and the decrease. After a damped step we have

$$x_+ = x + \alpha \Delta x, \quad y_+ = y + \alpha \Delta y, \quad s_+ = s + \alpha \Delta s.$$

Hence, with

$$v = \sqrt{\frac{xs}{\mu}}, \quad d_x = \frac{v \Delta x}{x}, \quad d_s = \frac{v \Delta s}{s},$$

we have

$$x_+ = x \left(e + \alpha \frac{\Delta x}{x} \right) = x \left(e + \alpha \frac{d_x}{v} \right) = \frac{x}{v} (v + \alpha d_x),$$

and

$$s_+ = s \left(e + \alpha \frac{\Delta s}{s} \right) = s \left(e + \alpha \frac{d_s}{v} \right) = \frac{s}{v} (v + \alpha d_s).$$

Thus we obtain

$$(4.1) \quad v_+^2 = \frac{x_+ s_+}{\mu} = (v + \alpha d_x) (v + \alpha d_s).$$

Hence, since ψ is e -convex,

$$\Psi(v_+) = \Psi \left(\sqrt{(v + \alpha d_x)(v + \alpha d_s)} \right) \leq \frac{1}{2} (\Psi(v + \alpha d_x) + \Psi(v + \alpha d_s)).$$

Defining

$$f(\alpha) := \Psi(v_+) - \Psi(v),$$

we thus have $f(\alpha) \leq f_1(\alpha)$, where

$$f_1(\alpha) := \frac{1}{2} (\Psi(v + \alpha d_x) + \Psi(v + \alpha d_s)) - \Psi(v).$$

Obviously,

$$f(0) = f_1(0) = 0.$$

Taking the derivative with respect to α , we get

$$f_1'(\alpha) = \frac{1}{2} \sum_{i=1}^n (\psi'(v_i + \alpha d_{x_i}) d_{x_i} + \psi'(v_i + \alpha d_{s_i}) d_{s_i}).$$

This gives, using the third equation in (2.8) and (2.13),

$$(4.2) \quad f_1'(0) = \frac{1}{2} \nabla \Psi(v)^T (d_x + d_s) = -\frac{1}{2} \nabla \Psi(v)^T \nabla \Psi(v) = -2\delta(v)^2.$$

Differentiating once more, we obtain

$$(4.3) \quad f_1''(\alpha) = \frac{1}{2} \sum_{i=1}^n (\psi''(v_i + \alpha d_{x_i}) d_{x_i}^2 + \psi''(v_i + \alpha d_{s_i}) d_{s_i}^2).$$

Below we use the following notation:

$$v_1 := \min(v), \quad \delta := \delta(v),$$

with $\delta(v)$ as defined in (2.13).

LEMMA 4.1. *One has $f_1''(\alpha) \leq 2\delta^2 \psi''(v_1 - 2\alpha\delta)$.*

Proof. Since d_x and d_s are orthogonal, (2.13) implies that $\|(d_x, d_s)\| = 2\delta$. Hence we have $\|d_x\| \leq 2\delta$ and $\|d_s\| \leq 2\delta$. Therefore,

$$v_i + \alpha d_{x_i} \geq v_1 - 2\alpha\delta, \quad v_i + \alpha d_{s_i} \geq v_1 - 2\alpha\delta, \quad 1 \leq i \leq n.$$

Due to (2.14c), $\psi''(t)$ is monotonically decreasing. Therefore, from (4.3) we obtain

$$f_1''(\alpha) \leq \frac{1}{2} \psi''(v_1 - 2\alpha\delta) \sum_{i=1}^n (d_{x_i}^2 + d_{s_i}^2) = 2\delta^2 \psi''(v_1 - 2\alpha\delta).$$

This proves the lemma. $\square$

LEMMA 4.2. *$f_1'(\alpha) \leq 0$ certainly holds if α satisfies the inequality*

$$(4.4) \quad -\psi'(v_1 - 2\alpha\delta) + \psi'(v_1) \leq 2\delta.$$

Proof. We may write, using Lemma 4.1, and also (4.2),

$$\begin{aligned} f_1'(\alpha) &= f_1'(0) + \int_0^\alpha f_1''(\xi) d\xi \\ &\leq -2\delta^2 + 2\delta^2 \int_0^\alpha \psi''(v_1 - 2\xi\delta) d\xi \\ &= -2\delta^2 - \delta \int_0^\alpha \psi''(v_1 - 2\xi\delta) d(v_1 - 2\xi\delta) \\ &= -2\delta^2 - \delta (\psi'(v_1 - 2\alpha\delta) - \psi'(v_1)). \end{aligned}$$

Hence, $f_1'(\alpha) \leq 0$ will certainly hold if α satisfies

$$-\psi'(v_1 - 2\alpha\delta) + \psi'(v_1) \leq 2\delta,$$

which proves the lemma. $\square$

LEMMA 4.3. *Let $\rho : [0, \infty) \rightarrow (0, 1]$ denote the inverse function of the restriction of $-\frac{1}{2}\psi'(t)$ to the interval $(0, 1]$. Then the largest step size α that satisfies (4.4) is given by*

$$(4.5) \quad \bar{\alpha} := \frac{1}{2\delta}(\rho(\delta) - \rho(2\delta)).$$

Proof. We want α such that (4.4) holds, with α as large as possible. Since $\psi''(t)$ is decreasing, the derivative with respect to v_1 of the expression at the left in (4.4) (i.e., $-\psi''(v_1 - 2\alpha\delta) + \psi''(v_1)$) is negative. Hence, fixing δ , the smaller v_1 is, the smaller α will be. One has

$$\delta = \frac{1}{2} \|\nabla \Psi(v)\| \geq \frac{1}{2} |\psi'(v_1)| \geq -\frac{1}{2} \psi'(v_1).$$

Equality holds if and only if v_1 is the only coordinate in v that differs from 1, and $v_1 \leq 1$ (in which case $\psi'(v_1) \leq 0$). Hence, the worst situation for the step size occurs when v_1 satisfies

$$(4.6) \quad -\frac{1}{2} \psi'(v_1) = \delta.$$

The derivative with respect to α of the left expression in (4.4) equals $2\delta\psi''(v_1 - 2\alpha\delta) \geq 0$, and hence the left-hand side is increasing in α . So the largest possible value of α satisfying (4.4) satisfies

$$(4.7) \quad -\frac{1}{2} \psi'(v_1 - 2\alpha\delta) = 2\delta.$$

Due to the definition of ρ , (4.6) and (4.7) can be written as

$$v_1 = \rho(\delta), \quad v_1 - 2\alpha\delta = \rho(2\delta),$$

respectively. This implies

$$\alpha = \frac{1}{2\delta}(v_1 - \rho(2\delta)) = \frac{1}{2\delta}(\rho(\delta) - \rho(2\delta)),$$

proving the lemma. $\square$

LEMMA 4.4. *Let ρ and $\bar{\alpha}$ be as defined in Lemma 4.3. Then*

$$(4.8) \quad \bar{\alpha} \geq \frac{1}{\psi''(\rho(2\delta))}.$$

Proof. By the definition of ρ ,

$$-\psi'(\rho(\delta)) = 2\delta.$$

Taking the derivative with respect to δ , we find

$$-\psi''(\rho(\delta))\rho'(\delta) = 2,$$

which gives

$$(4.9) \quad \rho'(\delta) = -\frac{2}{\psi''(\rho(\delta))} < 0.$$

Hence ρ is monotonically decreasing. An immediate consequence of (4.5) is

$$(4.10) \quad \bar{\alpha} = \frac{1}{2\delta} \int_{2\delta}^{\delta} \rho'(\sigma) d\sigma = \frac{1}{\delta} \int_{\delta}^{2\delta} \frac{d\sigma}{\psi''(\rho(\sigma))},$$

where we also used (4.9). To obtain a lower bound for $\bar{\alpha}$, we want to replace the argument of the last integral by its minimal value. So we want to know when $\psi''(\rho(\sigma))$ is maximal, for $\sigma \in [\delta, 2\delta]$. Due to (2.14c), ψ'' is monotonically decreasing. So $\psi''(\rho(\sigma))$ is maximal for $\sigma \in [\delta, 2\delta]$ when $\rho(\sigma)$ is minimal. Since ρ is monotonically decreasing, this occurs when $\sigma = 2\delta$. Therefore

$$\bar{\alpha} = \frac{1}{\delta} \int_{\delta}^{2\delta} \frac{d\sigma}{\psi''(\rho(\sigma))} \geq \frac{1}{\delta} \frac{\delta}{\psi''(\rho(2\delta))} = \frac{1}{\psi''(\rho(2\delta))},$$

which proves the lemma. $\square$

In what follows we use the notation

$$(4.11) \quad \tilde{\alpha} = \frac{1}{\psi''(\rho(2\delta))}$$

and we will use $\tilde{\alpha}$ as the default step size. By Lemma 4.4 we have $\bar{\alpha} \geq \tilde{\alpha}$.

LEMMA 4.5. *If the step size α is such that $\alpha \leq \bar{\alpha}$, then*

$$(4.12) \quad f(\alpha) \leq -\alpha\delta^2.$$

Proof. Let the univariate function h be such that

$$h(0) = f_1(0) = 0, \quad h'(0) = f_1'(0) = -2\delta^2, \quad h''(\alpha) = 2\delta^2\psi''(v_1 - 2\alpha\delta).$$

Due to Lemma 4.1, $f_1''(\alpha) \leq h''(\alpha)$. As a consequence, $f_1'(\alpha) \leq h'(\alpha)$ and $f_1(\alpha) \leq h(\alpha)$. Taking $\alpha \leq \bar{\alpha}$, with $\bar{\alpha}$ as defined in Lemma 4.3, we have

$$h'(\alpha) = -2\delta^2 + 2\delta^2 \int_0^{\alpha} \psi''(v_1 - 2\xi\delta) d\xi = -2\delta^2 - \delta(\psi'(v_1 - 2\alpha\delta) - \psi'(v_1)) \leq 0.$$

Since $h''(\alpha)$ is increasing in α , using Lemma A.3, we may write

$$f_1(\alpha) \leq h(\alpha) \leq \frac{1}{2}\alpha h'(0) = -\alpha\delta^2.$$

Since $f(\alpha) \leq f_1(\alpha)$, the proof is complete. $\square$

Combining the results of Lemmas 4.4 and 4.5 we obtain the following theorem.

THEOREM 4.6. *With $\tilde{\alpha}$ being the default step size, as given by (4.11), one has*

$$(4.13) \quad f(\tilde{\alpha}) \leq -\frac{\delta^2}{\psi''(\rho(2\delta))}.$$

LEMMA 4.7. *The right-hand side expression in (4.13) is monotonically decreasing in δ .*

Proof. Putting $t = \rho(2\delta)$, which implies $t \leq 1$, and which is equivalent to $4\delta = -\psi'(t)$, t is monotonically decreasing if δ increases. Hence the right-hand expression in (4.13) is monotonically decreasing in δ if and only if the function

$$g(t) := \frac{(\psi'(t))^2}{16\psi''(t)}$$

is monotonically decreasing for $t \leq 1$. Note that $g(1) = 0$ and

$$g'(t) = \frac{2\psi'(t)\psi''(t)^2 - \psi'(t)^2\psi'''(t)}{16\psi''(t)^2}.$$

Hence, since $\psi'(t) < 0$ for $t < 1$, $g(t)$ is monotonically decreasing for $t \leq 1$ if and only if

$$2\psi''(t)^2 - \psi'(t)\psi'''(t) \geq 0, \quad t \leq 1.$$

The last inequality is satisfied, due to condition (2.14d). Hence the lemma is proved. $\square$

We want to express the decrease as a function of $\Psi(v)$. To this end we need a lower bound on $\delta(v)$ in terms of $\Psi(v)$. Such a bound is provided in the following section.

4.2. Bound on $\delta(v)$ in terms of $\Psi(v)$. We need the following lemma.

LEMMA 4.8. *Suppose that $\psi(t_1) = \psi(t_2)$, with $t_1 \leq 1 \leq t_2$. Then $\psi'(t_1) \leq 0$ and $\psi'(t_2) \geq 0$, whereas*

$$-\psi'(t_1) \geq \psi'(t_2).$$

Proof. The lemma is obvious if $t_1 = 1$ or $t_2 = 1$, because then $\psi(t_1) = \psi(t_2) = 0$ implies $t_1 = t_2 = 1$. So we may assume that $t_1 < 1 < t_2$. Since $\psi(t_1) = \psi(t_2)$, Lemma 2.6 implies that

$$\frac{1}{2}(t_1 - 1)^2 \psi''(1) < \psi(t_1) = \psi(t_2) < \frac{1}{2}(t_2 - 1)^2 \psi''(1).$$

Hence, since $\psi''(1) > 0$, it follows that $t_2 - 1 > 1 - t_1$. Using this and Lemma 2.5, while assuming $-\psi'(t_1) < \psi'(t_2)$, we may write

$$\psi(t_2) > \frac{1}{2}(t_2 - 1)\psi'(t_2) > \frac{1}{2}(1 - t_1)\psi'(t_2) > -\frac{1}{2}(1 - t_1)\psi'(t_1) = \frac{1}{2}(t_1 - 1)\psi'(t_1) > \psi(t_1).$$

This contradiction proves the lemma. $\square$

THEOREM 4.9. *One has*

$$\delta(v) \geq \frac{1}{2}\psi'(\varrho(\Psi(v))).$$

Proof. The statement in the lemma is obvious if $v = e$ since then $\delta(v) = \Psi(v) = 0$. Otherwise we have $\delta(v) > 0$ and $\Psi(v) > 0$. To deal with the nontrivial case we consider, for $\omega > 0$, the problem

$$z_\omega = \min_v \left\{ \delta(v)^2 = \frac{1}{4} \sum_{i=1}^n \psi'(v_i)^2 : \Psi(v) = \omega \right\}.$$

The first order optimality condition is

$$\frac{1}{2}\psi'(v_i)\psi''(v_i) = \lambda\psi'(v_i), \quad i = 1, \dots, n,$$

where $\lambda \in \mathbb{R}$. From this we conclude that we have either $\psi'(v_i) = 0$ or $\psi''(v_i) = 2\lambda$ for each i . Since $\psi''(t)$ is monotonically decreasing, this implies that all v_i 's for which

$\psi''(v_i) = 2\lambda$ have the same value. Denoting this value as t , and observing that all other coordinates have value 1 (since $\psi'(v_i) = 0$ for these coordinates), we conclude that, after reordering the coordinates, v has the form

$$v = (\underbrace{t, \dots, t}_{k \text{ times}}, \underbrace{1, \dots, 1}_{n-k \text{ times}}).$$

Now $\Psi(v) = \omega$ implies $k\psi(t) = \omega$. Given k , this uniquely determines $\psi(t)$, whence we have

$$4\delta(v)^2 = k(\psi'(t))^2, \quad \psi(t) = \frac{\omega}{k}.$$

Note that the equation $\psi(t) = \frac{\omega}{k}$ has two solutions, one smaller than 1 and one larger than 1. By Lemma 4.8, the larger value gives the smallest value of $(\psi'(t))^2$. Since we are minimizing $\delta(v)^2$, we conclude that $t > 1$ (since $\omega > 0$). Hence we may write

$$t = \varrho\left(\frac{\omega}{k}\right),$$

where, as before, ϱ denotes the inverse function of $\psi(t)$ for $t \geq 1$. Thus we obtain that

$$(4.14) \quad 4\delta(v)^2 = k(\psi'(t))^2, \quad t = \varrho\left(\frac{\omega}{k}\right).$$

The question now is which value of k minimizes $\delta(v)^2$. To investigate this, we take the derivative with respect to k of (4.14) extended to $k \in \mathbb{R}$. This gives

$$(4.15) \quad \frac{d4\delta(v)^2}{dk} = (\psi'(t))^2 + 2k\psi'(t)\psi''(t)\frac{dt}{dk}.$$

From $\psi(t) = \frac{\omega}{k}$ we derive that

$$\psi'(t)\frac{dt}{dk} = -\frac{\omega}{k^2} = -\frac{\psi(t)}{k},$$

which gives

$$\frac{dt}{dk} = -\frac{\psi(t)}{k\psi'(t)}.$$

Substitution into (4.15) gives

$$\frac{d4\delta(v)^2}{dk} = (\psi'(t))^2 - 2k\psi'(t)\psi''(t)\frac{\psi(t)}{k\psi'(t)} = (\psi'(t))^2 - 2\psi(t)\psi''(t).$$

Defining $f(t) := (\psi'(t))^2 - 2\psi(t)\psi''(t)$, we have $f(1) = 0$ and

$$f'(t) = 2\psi'(t)\psi''(t) - 2\psi'(t)\psi''(t) - 2\psi(t)\psi'''(t) = -2\psi(t)\psi'''(t) > 0.$$

We conclude that $f(t) > 0$ for $t > 1$. Hence $\frac{d\delta(v)^2}{dk} > 0$, so $\delta(v)^2$ increases when k increases. Since we are minimizing $\delta(v)^2$, at optimality we have $k = 1$. Also using that $\psi(t) \geq 0$, we obtain from (4.14) that

$$\min_v \{\delta(v) : \Psi(v) = \omega\} = \frac{1}{2}\psi'(t) = \frac{1}{2}\psi'(\varrho(\omega)) = \frac{1}{2}\psi'(\varrho(\Psi(v))).$$

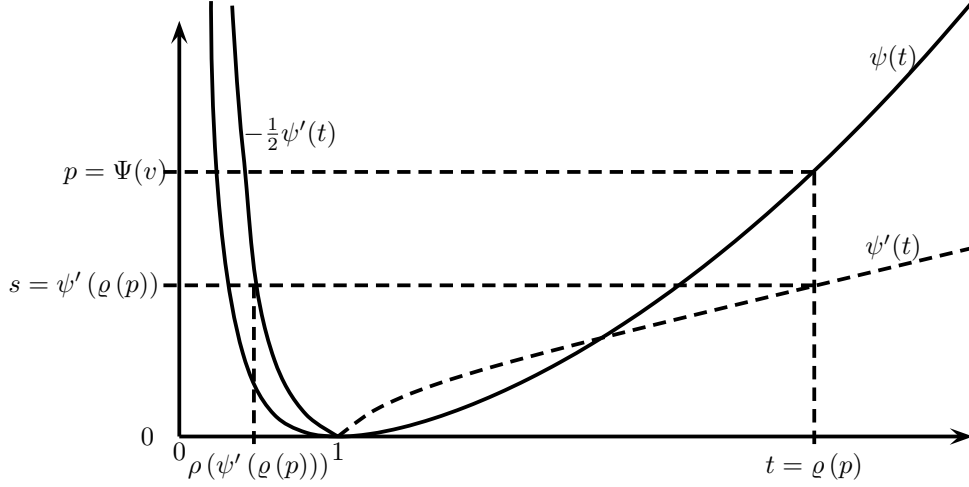


FIG. 2. Graphical illustration of the entities arising in (4.16).

This completes the proof of the theorem. $\square$

Remark 4.10. The bound of Theorem 4.9 is sharp. One may easily verify that if v is such that all coordinates are equal to 1 except one coordinate which is greater than or equal to 1, then the bound holds with equality.

Remark 4.11. It is worth noting that the proof of Theorem 4.9 implies that our kernel functions satisfy the inequality (cf. [12, page 37])

$$\psi'(t)^2 - 2\psi(t)\psi''(t) \geq 0, \quad t \geq 1.$$

Combining the results of Theorems 4.6 and 4.9 we obtain

$$(4.16) \quad f(\tilde{\alpha}) \leq -\frac{(\psi'(\rho(\Psi(v))))^2}{4\psi''(\rho(\psi'(\rho(\Psi(v)))))}.$$

This expresses the decrease in $\Psi(v)$ during an inner iteration completely in ψ , its first and second derivatives, and the inverse functions ρ and ϱ . The entities arising in the right-hand side of (4.16) are depicted in Figure 2.

5. Iteration bounds. After the update of μ to $(1-\theta)\mu$ we have, by Theorem 3.2 and (3.2),

$$(5.1) \quad \Psi(v_+) \leq L_\psi(n, \theta, \tau) = n\psi\left(\frac{\varrho\left(\frac{\tau}{n}\right)}{\sqrt{1-\theta}}\right).$$

We need to count how many inner iterations are required to return to the situation where $\Psi(v) \leq \tau$. We denote the value of $\Psi(v)$ after the μ -update as Ψ_0 , and the subsequent values are denoted as Ψ_k , $k = 1, 2, \dots$. The decrease on each inner iteration is given by (4.16). In what follows we assume that the expression in the right-hand side expression of (4.16) satisfies

$$(5.2) \quad \frac{(\psi'(\varrho(\Psi(v))))^2}{4\psi''(\rho(\psi'(\varrho(\Psi(v)))))} \geq \kappa\Psi(v)^{1-\gamma}$$

for some positive constants κ and γ , with $\gamma \in (0, 1]$. It may be worth noting at this point that property (5.2) has to be checked for each case individually. When dealing with the seven specific examples later on, we will show how appropriate values of κ and γ are obtained in each case.

LEMMA 5.1. *If K denotes the number of inner iterations, we have*

$$K \leq \frac{\Psi_0^\gamma}{\kappa\gamma}.$$

Proof. The definition of K implies $\Psi_{K-1} > \tau$ and $\Psi_K \leq \tau$ and

$$\Psi_{k+1} \leq \Psi_k - \kappa\Psi_k^{1-\gamma}, \quad k = 0, 1, \dots, K-1.$$

Yet we apply Lemma A.2, with $t_k = \Psi_k$. This yields the desired inequality. $\square$

The last lemma provides an estimate for the number of inner iterations in terms of Ψ_0 and the constants κ and γ . Recall that Ψ_0 is bounded above according to (5.1). An upper bound for the total number of iterations is obtained by multiplying (the upper bound for) the number K by the number of barrier parameter updates, which is bounded above by (cf. [14, Lemma II.17, page 116])

$$\frac{1}{\theta} \log \frac{n}{\varepsilon}.$$

Thus we obtain the following upper bound on the total number of iterations:

$$(5.3) \quad \frac{\Psi_0^\gamma}{\theta\kappa\gamma} \log \frac{n}{\varepsilon} \leq \frac{1}{\theta\kappa\gamma} \left(n\psi \left(\frac{\varrho \left(\frac{\tau}{n} \right)}{\sqrt{1-\theta}} \right) \right)^\gamma \log \frac{n}{\varepsilon}.$$

6. Application to the seven kernel functions.

6.1. Introduction. We apply the results of the previous sections to obtain iteration bounds for large- and small-update methods based on the seven kernel functions introduced before. Thus, for each of the kernel functions we will basically do the following.

Step 0: Input a kernel function ψ ; an update parameter θ , $0 < \theta < 1$; a threshold parameter τ ; and an accuracy parameter ε .

Step 1: Solve the equation $-\frac{1}{2}\psi'(t) = s$ to get $\rho(s)$, the inverse function of $-\frac{1}{2}\psi'(t)$, $t \in (0, 1]$. If the equation is hard to solve, derive a lower bound for $\rho(s)$.

Step 2: Calculate the decrease of $\Psi(v)$ in terms of δ for the default step size $\tilde{\alpha}$ from

$$f(\tilde{\alpha}) \leq -\frac{\delta^2}{\psi''(\rho(2\delta))}.$$

Step 3: Solve the equation $\psi(t) = s$ to get $\varrho(s)$, the inverse function of $\psi(t)$, $t \geq 1$. If the equation is hard to solve, derive lower and upper bounds for $\varrho(s)$.

Step 4: Derive a lower bound for δ in terms of $\Psi(v)$ by using

$$\delta(v) \geq \frac{1}{2}\psi'(\varrho(\Psi(v))).$$

Step 5: Using the results of Steps 3 and 4 find positive constants κ and γ , with $\gamma \in (0, 1]$, such that

$$f(\tilde{\alpha}) \leq -\kappa\Psi(v)^{1-\gamma}.$$

Step 6: Calculate the uniform upper bound Ψ_0 for $\Psi(v)$ from

$$\Psi_0 \leq L_\psi(n, \theta, \tau) = n\psi\left(\frac{\varrho\left(\frac{\tau}{n}\right)}{\sqrt{1-\theta}}\right).$$

Step 7: Derive an upper bound for the total number of iterations from

$$\frac{\Psi_0^\gamma}{\theta\kappa\gamma} \log \frac{n}{\epsilon}.$$

Step 8: Set $\tau = O(n)$ and $\theta = \Theta(1)$ so as to calculate an iteration bound for large-update methods, or set $\tau = O(1)$ and $\theta = \Theta(\frac{1}{\sqrt{n}})$ to get an iteration bound for small-update methods.

Using the above scheme, our aim is to compute iteration bounds for large- and small-update methods based on the seven kernel functions. Large-update methods are characterized by $\tau = O(n)$ and $\theta = \Theta(1)$. It may be noted that we could also take smaller values of τ , e.g., $\tau = O(1)$, but one may easily check from the outcome of our analysis that this would not affect the order of magnitude of the bounds. Small-update methods are characterized by $\tau = O(1)$ and $\theta = \Theta(\frac{1}{\sqrt{n}})$.

One more remark is in order. At the start of each inner iteration we have $\Psi(v) \geq \tau$. By Theorem 4.9 this implies that $\delta(v) \geq \frac{1}{2}\psi'(\varrho(\tau))$. We always assume that $\tau \geq 1$, and that τ is large enough to ensure that $\delta(v) \geq 1$ at the start of each inner iteration.

By way of example, we perform in the next section the complete analysis for $\psi(t) = \psi_6(t)$. Later on, we discuss more briefly the analysis for the remaining kernel functions. In the analysis we make use of three more lemmas. The first two lemmas are useful in deriving bounds for the inverse functions ϱ and ρ , in case these functions cannot be computed explicitly. The third lemma sometimes gives a better estimate for Ψ_0 in Step 6. The lemmas apply only to the first six kernel functions.

LEMMA 6.1. *Let $\psi(t) = \psi_i(t)$, with $1 \leq i \leq 6$, and let $\underline{\rho} : [0, \infty) \rightarrow (0, 1]$ be the inverse function of the restriction of $-\psi'_b(t)$ to the interval $(0, 1]$, with $\psi_b(t)$ as defined in (2.15). Then one has*

$$\rho(s) \geq \underline{\rho}(1 + 2s).$$

Proof. Let $t = \rho(s)$. Due to the definition of ρ as the inverse function of $-\frac{1}{2}\psi'(t)$ for $t \leq 1$ this means that

$$-2s = \psi'(t) = t + \psi'_b(t), \quad t \leq 1.$$

Since $t \leq 1$ this implies

$$-\psi'_b(t) = t + 2s \leq 1 + 2s.$$

Since $-\psi'_b(t)$ is monotonically decreasing in all six cases, it follows from this that

$$t = \rho(s) \geq \underline{\rho}(1 + 2s),$$

proving the lemma. $\square$

To illustrate the use of this lemma consider, for example, the case $i = 2$. Then

$$\psi(t) = \frac{1}{2} \left(t - \frac{1}{t} \right)^2 = \frac{t^2 - 1}{2} + \frac{t^{-2} - 1}{2},$$

whence $\psi_b(t) = \frac{t^{-2}-1}{2}$. The inverse function of $-\psi'_b(t) = \frac{1}{t^3}$ is given by $\varrho(s) = \frac{1}{s^{\frac{1}{3}}}$. Hence, by Lemma 6.1,

$$\rho(s) \geq \frac{1}{(1+2s)^{\frac{1}{3}}}.$$

It follows that for the default step size $\tilde{\alpha}$ we have

$$f(\tilde{\alpha}) \leq -\frac{\delta^2}{\psi''(\rho(2\delta))} = -\frac{\delta^2}{1 + \frac{3}{\rho(2\delta)^4}} \leq -\frac{\delta^2}{1 + 3(1+4\delta)^{\frac{4}{3}}} \leq -\frac{\delta^{\frac{2}{3}}}{27}.$$

For the last inequality we used our assumption that $\delta \geq 1$ at the start of each inner iteration.

LEMMA 6.2. *When $\psi(t) = \psi_i(t)$ and $1 \leq i \leq 6$, then*

$$\sqrt{1+2s} \leq \varrho(s) \leq 1 + \sqrt{2s}.$$

Proof. The inverse function of $\psi(t)$ for $t \in [1, \infty)$ is obtained by solving t from the equation $\psi(t) = s$, for $t \geq 1$. In almost all cases it is hard to solve this equation explicitly. However, we can easily find a lower and an upper bound for t , and this suffices for our goal. First one has

$$s = \psi(t) = \frac{t^2 - 1}{2} + \psi_b(t) \leq \frac{t^2 - 1}{2},$$

where $\psi_b(t)$ denotes the barrier term. The inequality is due to the fact that $\psi_b(1) = 0$ and $\psi_b(t)$ is monotonically decreasing. It follows that

$$t = \varrho(s) \geq \sqrt{1+2s}.$$

For the second inequality we use that $\psi''_i(t) \geq 1$ for $1 \leq i \leq 6$, as is clear from Table 2.1. Also using (2.12) we may write

$$s = \psi(t) = \int_1^t \int_1^\xi \psi''(\zeta) d\zeta d\xi \geq \int_1^t \int_1^\xi d\zeta d\xi = \frac{1}{2}(t-1)^2,$$

which implies

$$t = \varrho(s) \leq 1 + \sqrt{2s}.$$

This completes the proof. $\square$

LEMMA 6.3. *Let $1 \leq i \leq 6$. Then one has*

$$L_\psi(n, \theta, \tau) \leq \frac{\psi''(1)}{2} \frac{(\sqrt{2\tau} + \theta\sqrt{n})^2}{1 - \theta}.$$

Hence, if $\tau = O(1)$ and $\theta = \Theta(\frac{1}{\sqrt{n}})$, then $\Psi_0 = O(\psi''(1))$.

Proof. By Lemma 6.2 we have $\varrho(s) \leq 1 + \sqrt{2s}$. Hence, also using (5.1) we have

$$\Psi_0 \leq L_\psi(n, \theta, \tau) = n\psi\left(\frac{\varrho(\frac{\tau}{n})}{\sqrt{1-\theta}}\right) \leq n\psi\left(\frac{1 + \sqrt{\frac{2\tau}{n}}}{\sqrt{1-\theta}}\right).$$

Applying Lemma 2.6 we obtain

$$\Psi_0 \leq \frac{n\psi''(1)}{2} \left(\frac{1 + \sqrt{\frac{2\tau}{n}}}{\sqrt{1-\theta}} - 1 \right)^2 \leq \frac{n\psi''(1)}{2} \left(\frac{\theta + \sqrt{\frac{2\tau}{n}}}{\sqrt{1-\theta}} \right)^2 = \frac{\psi''(1)}{2} \frac{(\sqrt{2\tau} + \theta\sqrt{n})^2}{1-\theta},$$

where we also used

$$(6.1) \quad 1 - \sqrt{1-\theta} = \frac{\theta}{1 + \sqrt{1-\theta}} \leq \theta.$$

This proves the lemma. $\square$

6.2. Example: Analysis of methods based on $\psi_6(t)$. Consider $\psi(t) = \psi_6(t)$:

$$\psi(t) = \frac{t^2 - 1}{2} - \int_1^t e^{\frac{1}{\xi} - 1} d\xi.$$

The inverse function of $-\psi'_b(t) = e^{\frac{1}{t} - 1}$ is given by $\varrho(s) = \frac{1}{1 + \log s}$. Hence, by Lemma 6.1,

$$\rho(s) \geq \frac{1}{1 + \log(1 + 2s)}.$$

It follows that

$$f(\tilde{\alpha}) \leq -\frac{\delta^2}{\psi''(\rho(2\delta))} = -\frac{\delta^2}{1 + \frac{e^{\frac{1}{\rho(2\delta)} - 1}}{\rho(2\delta)^2}} \leq -\frac{\delta^2}{1 + (1 + 4\delta)(1 + \log(1 + 4\delta))^2}.$$

By Lemma 6.2 the inverse function of $\psi(t)$ for $t \in [1, \infty)$ satisfies

$$\sqrt{1 + 2s} \leq \varrho(s) \leq 1 + \sqrt{2s}.$$

Thus we have, omitting the argument v ,

$$\varrho(\Psi(v)) \geq \sqrt{1 + 2\Psi}.$$

Now using that $\delta(v) \geq \frac{1}{2}\psi'(\varrho(\Psi(v)))$, we obtain

$$\delta \geq \frac{1}{2} \left(\sqrt{1 + 2\Psi} - e^{\frac{1}{\sqrt{1 + 2\Psi}} - 1} \right) \geq \frac{1}{2} \left(\sqrt{1 + 2\Psi} - 1 \right) = \frac{\Psi}{1 + \sqrt{1 + 2\Psi}}.$$

Substitution gives, after some elementary reductions, while using $\Psi_0 \geq \Psi \geq \tau \geq 1$,

$$f(\tilde{\alpha}) \leq -\frac{\Psi^{\frac{1}{2}}}{21(1 + \log(1 + \sqrt{\Psi}))^2} \leq -\frac{\Psi^{\frac{1}{2}}}{21(1 + \log(1 + \sqrt{\Psi_0}))^2}.$$

Thus it follows that

$$\Psi_{k+1} \leq \Psi_k - \kappa (\Psi_k)^{1-\gamma}, \quad k = 0, 1, \dots, K-1,$$

with $\kappa = \frac{1}{21(1 + \log(1 + \sqrt{\Psi_0}))^2}$ and $\gamma = \frac{1}{2}$, and where K denotes the number of inner iterations. Hence the number K of inner iterations is bounded above by

$$K \leq \frac{\Psi_0^\gamma}{\kappa\gamma} = 42 \left(1 + \log(1 + \sqrt{\Psi_0}) \right)^2 \Psi_0^{\frac{1}{2}}.$$

TABLE 6.1
Results of Steps 1 and 2.

i	Kernel function $\psi_i(t)$	$\rho(s)$	$f(\tilde{\alpha}) \leq$
1	$\frac{t^2-1}{2} - \log t$	$= \frac{1}{s+\sqrt{1+s^2}}$	$-\frac{1}{19}$
2	$\frac{1}{2} \left(t - \frac{1}{t}\right)^2$	$\geq \frac{1}{(1+2s)^{\frac{1}{3}}}$	$-\frac{\delta^{\frac{2}{3}}}{27}$
3	$\frac{t^2-1}{2} + \frac{t^{1-q}-1}{q-1}$	$\geq \frac{1}{(1+2s)^{\frac{1}{q}}}$	$-\frac{\delta^2}{1+q(1+4\delta)^{\frac{q+1}{q}}}$
4	$\frac{t^2-1}{2} + \frac{t^{1-q}-1}{q(q-1)} - \frac{q-1}{q}(t-1)$	$\geq \frac{1}{(1+2qs)^{\frac{1}{q}}}$	$-\frac{\delta^2}{1+(1+4q\delta)^{\frac{q+1}{q}}}$
5	$\frac{t^2-1}{2} + \frac{e^{\frac{1}{t}}-e}{e}$	$\geq \frac{1}{1+\log(1+2s)}$	$-\frac{\delta^2}{1+15\delta(1+\log(1+4\delta))^2}$
6	$\frac{t^2-1}{2} - \int_1^t e^{\frac{1}{\xi}-1} d\xi$	$\geq \frac{1}{1+\log(1+2s)}$	$-\frac{\delta^2}{1+(1+4\delta)(1+\log(1+4\delta))^2}$
7	$t-1 + \frac{t^{1-q}-1}{q-1}$	$= \frac{1}{(1+2s)^{\frac{1}{q}}}$	$-\frac{\delta^2}{q(4\delta+1)^{\frac{q+1}{q}}}$

We use Lemma 6.3, with $\psi''(1) = 2$, to estimate Ψ_0 . This gives

$$\Psi_0 \leq \frac{(\theta\sqrt{n} + \sqrt{2\tau})^2}{1 - \theta}.$$

Substitution in the expression for K gives

$$K \leq 42 \left(1 + \log \left(1 + \frac{\theta\sqrt{n} + \sqrt{2\tau}}{\sqrt{1-\theta}} \right) \right)^2 \frac{\theta\sqrt{n} + \sqrt{2\tau}}{\sqrt{1-\theta}}.$$

Thus the total number of iterations is bounded above by

$$\frac{K}{\theta} \log \frac{n}{\varepsilon} \leq 42 \left(1 + \log \left(1 + \frac{\theta\sqrt{n} + \sqrt{2\tau}}{\sqrt{1-\theta}} \right) \right)^2 \frac{\theta\sqrt{n} + \sqrt{2\tau}}{\theta\sqrt{1-\theta}} \log \frac{n}{\varepsilon}.$$

For large-update methods (when $\tau = O(n)$ and $\theta = \Theta(1)$) the right-hand side expression becomes $O(\sqrt{n}(\log n)^2 \log \frac{n}{\varepsilon})$, and for small-update methods (when $\tau = O(1)$ and $\theta = \Theta(\frac{1}{\sqrt{n}})$) the right-hand side expression becomes $O(\sqrt{n} \log \frac{n}{\varepsilon})$.

6.3. Analysis of all the examples. In this section we survey the analysis results for all the kernel functions $\psi_i(t)$, $1 \leq i \leq 7$. The analysis for each of the kernel functions proceeds in the same way as in the previous section for $\psi_6(t)$. To save space, we do not present details of the computations for the other kernel functions. We only give the outcome of each step of our computational scheme for each of the kernel functions. For readers who want to see the details of the computations, we refer to an earlier draft of this paper that can be downloaded from <http://www.isa.ewi.tudelft.nl/~roos/wpapers.html>. The outcome of Steps 1–7 is given in Tables 6.1, 6.2, and 6.3.

6.4. Complexity results. To get complexity results we finally have to perform Step 8 in our scheme. Setting $\tau = O(n)$ and $\theta = \Theta(1)$ we obtain the iteration bound for large-update methods, and setting $\tau = O(1)$ and $\theta = \Theta(\frac{1}{\sqrt{n}})$ we obtain the iteration bound for small-update methods. The resulting iteration bounds are summarized in the 3rd and 5th columns of Table 6.4.

TABLE 6.2
Results of Steps 3, 4, and 5.

i	$\varrho(s)$	$\delta(v) \geq$	γ	κ
1	not needed	not needed	1	$\frac{1}{19}$
2	$= \sqrt{\frac{s}{2}} + \sqrt{1 + \frac{s}{2}}$	$\frac{1}{2} \Psi^{\frac{1}{2}}$	$\frac{2}{3}$	0.02333
3	$\sqrt{1+2s} \leq \varrho(s) \leq 1 + \sqrt{2s}$	$\frac{1}{2} \left(\frac{\Psi}{2} \right)^{\frac{1}{2}}$	$\frac{q+1}{2q}$	$\frac{1}{56q}$
4	$\sqrt{1+2s} \leq \varrho(s) \leq 1 + \sqrt{2s}$	$\frac{\Psi}{1+\sqrt{1+2\Psi}}$	$\frac{q+1}{2q}$	$\frac{1}{53q}$
5	$\sqrt{1+2s} \leq \varrho(s) \leq 1 + \sqrt{2s}$	$\frac{\Psi}{1+\sqrt{1+2\Psi}}$	$\frac{1}{2}$	$\frac{1}{44 \left(1 + \log(1 + \sqrt{2\Psi_0}) \right)^2}$
6	$\sqrt{1+2s} \leq \varrho(s) \leq 1 + \sqrt{2s}$	$\frac{\Psi}{1+\sqrt{1+2\Psi}}$	$\frac{1}{2}$	$\frac{1}{21 \left(1 + \log(1 + \sqrt{\Psi_0}) \right)^2}$
7	$1 + s \leq \varrho(s) \leq 1 + \sqrt{s^2 + \frac{q}{q-1}s}$	$\frac{1}{2} \left(1 - \frac{1}{(\Psi+1)q} \right)$	1	$\frac{1}{64q}$

TABLE 6.3
Results of Steps 6 and 7.

i	$\Psi_0 \leq$	$\frac{K}{\theta} \leq$
1	$\frac{(\theta\sqrt{n} + \sqrt{2\tau})^2}{1-\theta}$	$19 \frac{(\theta\sqrt{n} + \sqrt{2\tau})^2}{\theta(1-\theta)}$
2	$\frac{1}{1-\theta} \left(\sqrt{\tau} + \frac{n\theta}{\sqrt{2n}} \right)^2$	$\frac{65}{\theta(1-\theta)^{\frac{2}{3}}} \left(\sqrt{\tau} + \frac{\theta\sqrt{n}}{\sqrt{2}} \right)^{\frac{4}{3}}$
3	$\frac{q+1}{2} \frac{(\theta\sqrt{n} + \sqrt{2\tau})^2}{1-\theta}$	$\frac{112q(q+1)}{\theta} \left(\frac{(\theta\sqrt{n} + \sqrt{2\tau})^2}{2(1-\theta)} \right)^{\frac{q+1}{2q}}$
4	$\frac{(\theta\sqrt{n} + \sqrt{2\tau})^2}{1-\theta}$	$\frac{106q}{\theta} \left(\frac{(\theta\sqrt{n} + \sqrt{2\tau})^2}{1-\theta} \right)^{\frac{q+1}{2q}}$
5	$2 \frac{(\theta\sqrt{n} + \sqrt{2\tau})^2}{1-\theta}$	$88\sqrt{2} \left(1 + \log \left(1 + 2 \frac{\theta\sqrt{n} + \sqrt{2\tau}}{\sqrt{1-\theta}} \right) \right)^2 \frac{\theta\sqrt{n} + \sqrt{2\tau}}{\theta\sqrt{1-\theta}}$
6	$\frac{(\theta\sqrt{n} + \sqrt{2\tau})^2}{1-\theta}$	$42 \left(1 + \log \left(1 + \frac{\theta\sqrt{n} + \sqrt{2\tau}}{\sqrt{1-\theta}} \right) \right)^2 \frac{\theta\sqrt{n} + \sqrt{2\tau}}{\theta\sqrt{1-\theta}}$
7	$\frac{q \left(\theta\sqrt{n} + \sqrt{\frac{\tau^2}{n} + 2\tau} \right)^2}{2(1-\theta)}$	$32 \frac{q^2 \left(\theta\sqrt{n} + \sqrt{\frac{\tau^2}{n} + 2\tau} \right)^2}{\theta(1-\theta)}$

Small-update methods based on the seven kernel functions all have the same complexity as the small-update method based on the logarithmic barrier function, namely, $O(\sqrt{n} \log \frac{n}{\varepsilon})$; as is well known, until now this has been the best iteration bound for methods solving LO problems. It must be noted that where appropriate (i.e., for $i \in \{3, 4, 7\}$) one has to take $q = O(1)$ to achieve this bound; moreover, the derivation of the small-update bound for $i = 7$ is valid only for $q \geq 2$.

Now let us consider the bounds for large-update methods in Table 6.4. It should be mentioned that the large-update bound in this table for $i = 3$ is based on estimates for Ψ_0 and $\frac{K}{\theta}$ that differ from the estimates in Table 6.3; for this case we used that

$$\Psi_0 \leq \frac{\theta n + 2\tau + 2\sqrt{2\tau n}}{2(1-\theta)}, \quad \frac{K}{\theta} \leq \frac{112q}{\theta} \left(\frac{\theta n + 2\tau + 2\sqrt{2\tau n}}{2(1-\theta)} \right)^{\frac{q+1}{2q}}.$$

TABLE 6.4
Complexity results for large- and small-update methods.

i	Kernel function $\psi_i(t)$	Small-update	Ref.	Large-update	Ref.
1	$\frac{t^2-1}{2} - \log t$	$O(\sqrt{n}) \log \frac{n}{\varepsilon}$	e.g., [14]	$O(n) \log \frac{n}{\varepsilon}$	e.g., [14]
2	$\frac{1}{2} \left(t - \frac{1}{t}\right)^2$	$O(\sqrt{n}) \log \frac{n}{\varepsilon}$	[7]	$O\left(n^{\frac{2}{3}}\right) \log \frac{n}{\varepsilon}$	[7]
3	$\frac{t^2-1}{2} + \frac{t^{1-q}-1}{q-1}$	$O(q^2\sqrt{n}) \log \frac{n}{\varepsilon}$	[8]	$O\left(qn^{\frac{q+1}{2q}}\right) \log \frac{n}{\varepsilon}$	[8, 12]
4	$\frac{t^2-1}{2} + \frac{t^{1-q}-1}{q(q-1)} - \frac{q-1}{q}(t-1)$	$O(q\sqrt{n}) \log \frac{n}{\varepsilon}$	[11]	$O\left(qn^{\frac{q+1}{2q}}\right) \log \frac{n}{\varepsilon}$	[11, 12]
5	$\frac{t^2-1}{2} + \frac{e^{\frac{1}{t}-e}}{e}$	$O(\sqrt{n}) \log \frac{n}{\varepsilon}$	new	$O(\sqrt{n} \log^2 n) \log \frac{n}{\varepsilon}$	new
6	$\frac{t^2-1}{2} - \int_1^t e^{\frac{1}{\xi}-1} d\xi$	$O(\sqrt{n}) \log \frac{n}{\varepsilon}$	new	$O(\sqrt{n} \log^2 n) \log \frac{n}{\varepsilon}$	new
7	$t-1 + \frac{t^{1-q}-1}{q-1}$	$O(q^2\sqrt{n}) \log \frac{n}{\varepsilon}$	new	$O(qn) \log \frac{n}{\varepsilon}$	[4]

One may easily verify that the best iteration bound for large-update methods is obtained for $i \in \{3, 4\}$ by taking $q = \frac{1}{2} \log n$. This gives the iteration bound $O(\sqrt{n}(\log n) \log \frac{n}{\varepsilon})$, which is currently the best known bound for large-update methods.

7. Concluding remarks. This paper was inspired by recent work on so-called self-regular barrier functions for primal-dual IPMs for LO [11, 12]. Each such barrier function is determined by its (univariate) self-regular kernel function. We introduced a new class of kernel functions which differs from the class of self-regular kernel functions. The class is defined by some simple conditions on the kernel function which concern the growth and the barrier behavior of the kernel function. These properties enable us to derive many new and tight estimates that greatly simplify the analysis of IPMs based on these kernel functions. It has become clear that inverse functions of suitable restrictions of the kernel function and its first derivative play a key role in the behavior of the corresponding IPMs. Moreover, due to the new estimates the analysis of an IPM based on a given kernel function is greatly simplified and uses tools only from a first year calculus course.

Future research might focus on finding a kernel function for which the complexity of large-update methods is equal to (or even better than) $O(\sqrt{n} \log \frac{n}{\varepsilon})$, or show that such a kernel function does not exist. In this respect it might be of interest to show that the new class of kernel functions contains many functions not considered so far. For example,

$$\psi(t) = \frac{t^2-1}{2} + \frac{6}{\pi} \tan \frac{\pi(1-t)}{4t+2}$$

also is a member of the new class.

Also the extensions to semidefinite optimization and second order cone optimization deserve to be investigated. At present no computational results exist for the methods presented in this paper. This will be another issue for future research.

Appendix. Three technical lemmas. We need three simple technical results. For the sake of completeness we include their (short) proofs. The first lemma is needed only in the proof of the second lemma, which is interesting in itself.

LEMMA A.1 (Lemma 2.1 in [9]). *If $\alpha \in [0, 1]$, then*

$$(A.1) \quad (1+t)^\alpha \leq 1 + \alpha t \quad \forall t \geq -1.$$

Proof. Consider the function $f(t) = (1+t)^\alpha - 1 - \alpha t$ for $t \geq -1$. One has $f'(t) = \alpha(1+t)^{\alpha-1} - \alpha$ and $f''(t) = \alpha(\alpha-1)(1+t)^{\alpha-2}$. Since $f''(t) \leq 0$, $f(t)$ is concave. Since $f'(0) = 0$, the function f is maximal at $t = 0$. Finally, since $f(0) = 0$, the lemma follows. $\square$

LEMMA A.2 (Proposition 2.2 in [9]). *Let $t_0, t_1, \dots, t_K$ be a sequence of positive numbers such that*

$$(A.2) \quad t_{k+1} \leq t_k - \kappa t_k^{1-\gamma}, \quad k = 0, 1, \dots, K-1,$$

where $\kappa > 0$ and $0 < \gamma \leq 1$. Then $K \leq \left\lfloor \frac{t_0^\gamma}{\kappa\gamma} \right\rfloor$.

Proof. Using (A.2), we may write

$$0 < t_{k+1}^\gamma \leq (t_k - \kappa t_k^{1-\gamma})^\gamma = t_k^\gamma (1 - \kappa t_k^{-\gamma})^\gamma \leq t_k^\gamma (1 - \kappa\gamma t_k^{-\gamma}) = t_k^\gamma - \kappa\gamma,$$

where the second inequality follows from (A.1). Hence, for each k , $t_k^\gamma \leq t_0^\gamma - k\gamma\kappa$. Taking $k = K$ we obtain $0 < t_0^\gamma - K\gamma\kappa$, which implies the lemma. $\square$

LEMMA A.3 (Lemma 3.12 in [11]). *Let $h(t)$ be a twice differentiable convex function with $h(0) = 0$, $h'(0) < 0$, and let $h(t)$ attain its (global) minimum at $t^* > 0$. If $h''(t)$ is increasing for $t \in [0, t^*]$, then*

$$h(t) \leq \frac{th'(0)}{2}, \quad 0 \leq t \leq t^*.$$

Proof. Using the hypothesis of the lemma we may write

$$\begin{aligned} h(t) &= \int_0^t h'(\xi) d\xi = h'(0)t + \int_0^t \int_0^\xi h''(\zeta) d\zeta d\xi \leq h'(0)t + \int_0^t \xi h''(\xi) d\xi \\ &= h'(0)t + \int_0^t \xi dh'(\xi) = h'(0)t + (\xi h'(\xi))|_0^t - \int_0^t h'(\xi) d\xi \\ &\leq h'(0)t - \int_0^t dh'(\xi) = h'(0)t - h(t). \end{aligned}$$

This implies the lemma. $\square$

Acknowledgments. The authors kindly acknowledge the help of the associate editor and two anonymous referees in improving the readability of the paper. The new and nice motivation of the search direction presented in section 2.2, as a Newton-like step, was provided by one of the referees, and is especially acknowledged.

REFERENCES

- [1] E.D. ANDERSEN, J. GONDZIO, CS. MÉSZÁROS, AND X. XU, *Implementation of interior-point methods for large scale linear programs*, in Interior Point Methods of Mathematical Programming, T. Terlaky, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996, pp. 189–252.
- [2] Y.Q. BAI, M. EL GHAMI, AND C. ROOS, *A primal-dual interior-point algorithm for linear optimization based on a new proximity function*, Optim. Methods Softw., 17 (2002), pp. 985–1008.

- [3] Y.Q. BAI, M. EL GHAMI, AND C. ROOS, *A new efficient large-update primal-dual interior-point method based on a finite barrier*, SIAM J. Optim., 13 (2003), pp. 766–782.
- [4] Y.Q. BAI AND C. ROOS, *A primal-dual interior-point method based on a new kernel function with linear growth rate*, Proceedings of the 9th Australian Optimisation Day, Perth, Australia, 2002, to appear.
- [5] N.K. KARMAKAR, *A new polynomial-time algorithm for linear programming*, Combinatorica, 4 (1984), pp. 373–395.
- [6] N. MEGIDDO, *Pathways to the optimal set in linear programming*, in Progress in Mathematical Programming: Interior Point and Related Methods, N. Megiddo, ed., Springer-Verlag, New York, 1989, pp. 131–158.
- [7] J. PENG, C. ROOS, AND T. TERLAKY, *New complexity analysis of the primal-dual Newton method for linear optimization*, Ann. Oper. Res., 99 (2000), pp. 23–39.
- [8] J. PENG, C. ROOS, AND T. TERLAKY, *A new and efficient large-update interior-point method for linear optimization*, Vychisl. Tekhnol., 6 (2001), pp. 61–80.
- [9] J. PENG, C. ROOS, AND T. TERLAKY, *A new class of polynomial primal-dual methods for linear and semidefinite optimization*, European J. Oper. Res., 143 (2002), pp. 234–256.
- [10] J. PENG, C. ROOS, AND T. TERLAKY, *Primal-dual interior-point methods for second-order conic optimization based on self-regular proximities*, SIAM J. Optim., 13 (2002), pp. 179–203.
- [11] J. PENG, C. ROOS, AND T. TERLAKY, *Self-regular functions and new search directions for linear and semidefinite optimization*, Math. Program., 93 (2002), pp. 129–171.
- [12] J. PENG, C. ROOS, AND T. TERLAKY, *Self-Regularity: A New Paradigm for Primal-Dual Interior-Point Algorithms*, Princeton University Press, Princeton, NJ, 2002.
- [13] J. RENEGAR, *A Mathematical View of Interior-Point Methods in Convex Optimization*, MPS/SIAM Ser. Optim. 3, SIAM, Philadelphia, 2001.
- [14] C. ROOS, T. TERLAKY, AND J.-PH. VIAL, *Theory and Algorithms for Linear Optimization. An Interior-Point Approach*, John Wiley & Sons, Chichester, UK, 1997.
- [15] G. SONNEVEND, *An “analytic center” for polyhedrons and new classes of global algorithms for linear (smooth, convex) programming*, in System Modelling and Optimization: Proceedings of the 12th IFIP-Conference, Budapest, Hungary, 1985, Lecture Notes in Control and Inform. Sci. 84, A. Prékopa, J. Szelecsán, and B. Strazicky, eds., Springer-Verlag, Berlin, 1986, pp. 866–876.
- [16] S.J. WRIGHT, *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia, 1997.
- [17] Y. YE, *Interior Point Algorithms, Theory and Analysis*, John Wiley & Sons, Chichester, UK, 1997.

A CLASS OF METHODS FOR PROJECTION ON THE INTERSECTION OF SEVERAL ELLIPSOIDS*

ANHUA LIN[†] AND SHIH-PING HAN[‡]

Abstract. The paper is concerned with the basic optimization problem of projecting a point onto the intersection of several ellipsoids. We present a class of methods where the problem is reduced to a sequence of projections onto the intersection of several balls. The subproblems are simpler and more tractable, but the main advantage is that, in so doing, we can avoid solving linear systems completely, and thus the methods are very suitable for large scale problems. The methods have been shown to have nice convergence properties under Slater's constraint qualification and to be very reliable and efficient in our testing.

Key words. projection, ellipsoid, optimization, convex programming

AMS subject classifications. 49M37, 65K05, 90C25, 90C30

DOI. 10.1137/S1052623403422297

1. Introduction. In this paper we present a class of methods for solving the optimization problem of projecting a point onto a set with ellipsoidal constraints:

$$(1) \quad \begin{array}{ll} \text{Minimize} & \|x - a\|^2 \\ \text{subject to} & x \in E_i := \{x | x^T A_i x + 2b_i^T x + \alpha_i \leq 0\}, \\ & i = 1, 2, \dots, m, \end{array}$$

where each A_i is a symmetric positive definite matrix, b_i is an n -vector, α_i is a scalar, and $\|\cdot\|$ is the 2-norm. A main motivation for this research is an attempt to develop general nonlinear programming algorithms based on solving subproblems of this type [6], and to make such an approach viable, we need efficient algorithms for solving this projection problem. In the class of algorithms presented here, the main idea is to replace ellipsoids by balls. More specifically, we solve the problem by iteratively doing projection on the intersection of some balls. There are several advantages to this approach. First, the subproblem is much simpler and can be solved very efficiently. Second, for some algorithms in this class, we can avoid solving linear systems completely; therefore, they are very suitable for large sparse problems.

In section 2 we describe the class of methods and in section 3 we study their convergence properties. Then we discuss how to project a point onto the intersection of several balls in section 4, and in section 5 we describe our numerical experiments. All discussion is limited to the R^n space. For a set S in R^n we use ∂S and $\text{int}(S)$ to denote its boundary and interior, respectively. We also use R_{++} to denote the set of positive real numbers and R_+^m to denote all m -vectors with nonnegative components. Throughout the paper, we assume that $E = \cap_{i=1}^m E_i$ satisfies the Slater condition: E has a nonempty interior, or equivalently in this case, E has more than one point. We also use $\mathcal{A}(x)$ to denote the active index set at x for (1).

*Received by the editors February 6, 2003; accepted for publication (in revised form) January 31, 2004; published electronically October 14, 2004.

<http://www.siam.org/journals/siopt/15-1/42229.html>

[†]Department of Mathematical Sciences, Middle Tennessee State University, 1301 East Main St., Murfreesboro, TN 37132 (alin@mtsu.edu).

[‡]Department of Applied Mathematics and Statistics, The Johns Hopkins University, 3400 N. Charles St., Baltimore, MD 21218-2682 (han@ams.jhu.edu).

2. The class of algorithms. The algorithms we present here are iterative and in each of them we generate a sequence of feasible points $\{z^k\}$ which will be shown to converge to the unique optimal solution. Having a point z^k at the k th iteration, we first solve a subproblem of the form

$$(2) \quad \begin{array}{ll} \text{Minimize} & \|x - a\|^2 \\ \text{subject to} & x \in B_i(z^k), \quad i = 1, 2, \dots, m, \end{array}$$

where each $B_i(z^k)$ is a ball approximation to the ellipsoid E_i at z^k . Let y^k be the solution of this subproblem. Generally, y^k may not be feasible in the original problem, and therefore we apply a feasibility-restoration procedure to y^k to produce the next feasible point z^{k+1} . To give a clear presentation of our methods, we explain how to generate the balls $B_i(z^k)$ and how to carry out the feasibility-restoration procedure in the following subsections.

2.1. Ball approximation. Given a general ellipsoid $M = \{x | x^T A x + 2b^T x + \alpha \leq 0\}$, for any point $z \in M$, we consider a ball $B_M(z)$ which can be viewed as a local approximation to the ellipsoid M at z . Some properties are needed for this set-mapping $B_M(z)$. For simplicity, we study those properties in terms of its center $c(z)$ and radius $\gamma(z)$, and thus the ball $B_M(z)$ is given by $B_M(z) := \{x | \|x - c(z)\| \leq \gamma(z)\}$. The center mapping, $c : M \rightarrow R^n$, and the radius mapping, $\gamma : M \rightarrow R_{++}$, are required to satisfy the following conditions:

1. Both c and γ are continuous mappings defined on M .
2. If $z \in \text{int}(M)$, then $z \in \text{int}(B_M(z))$.
3. If $z \in \partial M$, then $z \in \partial B_M(z)$, and $c(z) = z - \lambda(Az + b)$ for some fixed $\lambda > 0$.

It is not particularly difficult to construct mappings that can satisfy the above requirements and the class of methods differ in the choices of them. We give two examples below.

Scheme 1. For any two positive numbers λ and β ,

$$\begin{aligned} c(z) &= z - \lambda(Az + b), \\ \gamma(z) &= \lambda\|Az + b\| + \beta\sqrt{-(z^T Az + 2b^T z + \alpha)}. \end{aligned}$$

The advantage of this scheme is that it involves only matrix-vector products and does not need any matrix factorization. Hence, it is very suitable for large sparse A .

Scheme 2. For any two positive numbers λ and β ,

$$\begin{aligned} c(z) &= z - \lambda(Az + b), \\ \gamma(z) &= \lambda\|Az + b\| + \sqrt{\beta} \left(\sqrt{b^T A^{-1} b - \alpha} - \sqrt{z^T Az + 2b^T z + b^T A^{-1} b} \right). \end{aligned}$$

In this case, we may need to solve a linear system to get $A^{-1}b$, but when an ellipsoid is given in the form $M = \{x | (x - d)^T A (x - d) \leq \rho\}$, as in many applications, no linear system needs to be solved. Another advantage of this scheme is that the ball $B_M(z)$ will be contained in M when the parameters λ and β are suitably chosen. Therefore, the solution of our subproblems will be feasible to (1) and some feasibility-restoration procedure can be avoided. Without loss of generality, we assume that the center of M is the origin, i.e., $M = \{x | x^T A x + \alpha \leq 0\}$. We have the following lemma.

LEMMA 2.1. *Given $z \in M$, if we use Scheme 2 with $0 < \lambda, \beta \leq \frac{1}{\rho(A)}$, then $B_M(z) \subset M$, where $\rho(A)$ is the spectral radius of A .*

Proof. Let $c = c(z)$ and $\gamma = \gamma(z)$. To prove $B_M(z) \subset M$, we just need to show that $\forall u \in \partial M$, $\|u - c\| \geq \gamma$.

Let $M_z = \{u \in R^n | u^T A u \leq z^T A z\}$ and let N denote the ball centered at c with radius $\lambda \|Az\|$. From Lemma 2.1 in [7], we know $N \subset M_z$, and thus $c \in M_z \subset M$. Let u be any point on ∂M . Then the closed line segment $[c, u]$ has a unique intersection point with ∂M_z which we use w to denote. Obviously, $\|u - c\| = \|u - w\| + \|w - c\|$. Since $N \subset M_z$, $\|w - c\| \geq \lambda \|Az\|$. Notice $u^T A u = -\alpha$ and $w^T A w = z^T A z$, and so we have

$$\begin{aligned} \rho(A) \|u - w\|^2 &\geq (u - w)^T A (u - w) \\ &= u^T A u - 2u^T A w + w^T A w \\ &= -\alpha - 2(A^{\frac{1}{2}} u)^T (A^{\frac{1}{2}} w) + w^T A w \\ &\geq -\alpha - 2\sqrt{\|A^{\frac{1}{2}} u\|^2 \|A^{\frac{1}{2}} w\|^2} + w^T A w \\ &= -\alpha - 2\sqrt{(-\alpha)w^T A w + w^T A w} \\ &= \left(\sqrt{-\alpha} - \sqrt{w^T A w}\right)^2 \\ &= \left(\sqrt{-\alpha} - \sqrt{z^T A z}\right)^2, \end{aligned}$$

which leads to

$$\|u - w\| \geq \frac{\sqrt{-\alpha} - \sqrt{z^T A z}}{\sqrt{\rho(A)}} \geq \sqrt{\beta} \left(\sqrt{-\alpha} - \sqrt{z^T A z}\right),$$

and hence

$$\begin{aligned} \|u - z\| &= \|u - w\| + \|w - z\| \\ &\geq \sqrt{\beta} \left(\sqrt{-\alpha} - \sqrt{z^T A z}\right) + \lambda \|Az\| \\ &= \gamma. \quad \square \end{aligned}$$

2.2. Feasibility-restoration procedure. For any closed convex set S and two points u and v with $u \in S$, we define a step $\tau(u, v, S)$ as

$$\tau(u, v, S) = \max\{\tau | u + \tau(v - u) \in S \text{ and } \tau \in [0, 1]\}.$$

Therefore, the point $u + \tau(u, v, S)(v - u)$ is the closest point of S to v on the line segment $[u, v]$. For the set E in our problem,

$$\tau(u, v, E) = \min\{\tau(u, v, E_i) | i = 1, 2, \dots, m\},$$

where $\tau(u, v, E_i) = 1$ when $v \in E_i$ and, otherwise, $\tau(u, v, E_i)$ can be computed by solving a one-dimensional quadratic equation as

$$\frac{1}{d^T A_i d} \left\{ d^T (A_i u + b_i) + \sqrt{[d^T (A_i u + b_i)]^2 - (d^T A_i d)(u^T A_i u + 2b_i^T u + \alpha_i)} \right\},$$

where $d = u - v$. From the above formula, it is easy to see that $\tau(\cdot, \cdot, E_i)$ is continuous at any (u, v) where $u \in E_i$, $v \in R^n$, and $u \neq v$. Therefore, the mapping $\tau(\cdot, \cdot, E)$ is continuous on the set $\{(x, y) | x \in E, y \in R^n \text{ and } x \neq y\}$. In the class of algorithms,

having z^k , y^k , and a , we use the mapping $\tau(\cdot, \cdot, E)$ to generate the next feasible estimate z^{k+1} . Furthermore, if $x \neq y$ and $y - x$ is a feasible direction of E at x , then $\tau(x, y, E) > 0$. This follows directly from the definition of a feasible direction: a vector d is a feasible direction of a set S at x if $x + \lambda d \in S \forall \lambda \in [0, \epsilon]$ with some $\epsilon > 0$. The above continuity and positivity properties of the mapping τ will be needed in our convergence analysis.

2.3. Description of the algorithms. We now describe the class of algorithms. At an iteration of the algorithm, having a feasible point $z \in E$ as an estimate of the solution, we compute a better estimate $\bar{z}$ by first generating a ball $B_i(z) := \{x \mid \|x - c_i(z)\| \leq \gamma_i(z)\}$ for each ellipsoid E_i as described in section 2.1. Then we solve the following projection problem:

$$(3) \quad \begin{array}{ll} \text{Minimize} & \|x - a\|^2 \\ \text{subject to} & x \in B_i(z), \quad i = 1, 2, \dots, m. \end{array}$$

When the solution y of the above subproblem is infeasible for the original problem, we compute a feasible point v by

$$v = z + \tau(z, y, E)(y - z).$$

When y is feasible, we just simply set $v = y$. Since we are interested in points closer to a , we can easily improve v further by setting our next estimate $\bar{z}$ as

$$\bar{z} = v + \tau(v, a, E)(a - v).$$

The algorithm can now be summarized as follows.

1. Start from a feasible point z^0 .
2. At the k th iteration, having a feasible point z^k , we do the following.
 - (i) For each ellipsoid E_i , generate a ball $B_i(z^k)$ which satisfies the three conditions in section 2.1 or, in special cases, using one of the two schemes there.
 - (ii) Find the projection y^k of the given point a onto $\cap_{i=1}^m B_i(z^k)$.
 - (iii) If $y^k \in E$, then set $v^k = y^k$; otherwise, compute $v^k = z^k + \tau(z^k, y^k, E)(y^k - z^k)$.
 - (iv) Compute a new estimate z^{k+1} by setting $z^{k+1} = v^k + \tau(v^k, a, E)(a - v^k)$.

3. Convergence analysis. We first establish that our subproblems are well defined and satisfy the Slater condition whenever the original problem does. For simplicity, let $B(z)$ denote the intersection $\cap_{i=1}^m B_i(z)$. We note here an interesting fact that will be used very often in the following discussion: Both E and $B(z)$ have the same feasible direction cone at the point z . This is because, for $i \in \mathcal{A}(z)$, both E_i and $B_i(z)$ have the same normal vector $A_i z + b_i$ and, for $i \notin \mathcal{A}(z)$, z is an interior point for both E_i and $B_i(z)$.

LEMMA 3.1. *If $\text{int}(E) \neq \emptyset$, then, for any $z \in E$, $\text{int}(B(z)) \neq \emptyset$.*

Proof. Notice that it follows from our assumptions that all ellipsoids and balls are nondegenerate. Therefore, it suffices to show that the set $B(z)$ contains more than one point. Since $\text{int}(E) \neq \emptyset$, there exists $y \in E$ and $y \neq z$. By the convexity of E , the vector $y - z$ is a feasible direction of E at z . Therefore, by the remark preceding this lemma, it is also a feasible direction of $B(z)$ at z and hence $B(z)$ has more than one point, and therefore $\text{int}(B(z)) \neq \emptyset$. $\square$

We now establish a key lemma that shows the point y obtained from (3) is, indeed, a substantial improvement over the current point z .

LEMMA 3.2. *For any $z \in E$ and the solution y of (3),*

$$\|z - a\|^2 \geq \|y - a\|^2 + \|y - z\|^2.$$

Proof. If any two of z , y , and a are the same, the conclusion is obviously true.

If they are three distinct points, then they form a triangle. If $a - y$ and $z - y$ make an acute angle, then y cannot be the optimal solution of (3), since $y + t(z - y)$ is better for small $t > 0$. Therefore, the angle between $a - y$ and $z - y$ is greater than or equal to $\frac{\pi}{2}$, and hence the conclusion holds. $\square$

Using the same notation as in section 2.3, we know that the feasible point v lies on the line segment $[z, y]$. Therefore, by the convexity of the objective function, we have

$$\|z - a\| \geq \|v - a\| \geq \|y - a\|.$$

Moreover, the new point $\bar{z}$ is a further improvement over v , and hence we also have

$$\|z - a\| \geq \|v - a\| \geq \|\bar{z} - a\|.$$

Consequently, both $\{\|z^k - a\|\}$ and $\{\|v^k - a\|\}$ are nonincreasing and have the same limit. We summarize the above results in the following lemma.

LEMMA 3.3. *The following statements are true for the three sequences $\{z^k\}$, $\{v^k\}$, and $\{y^k\}$.*

- (i) $\|z^k - a\| \geq \|v^k - a\| \geq \|y^k - a\|$.
- (ii) $\|z^k - a\| \geq \|v^k - a\| \geq \|z^{k+1} - a\|$.
- (iii) $\lim \|z^k - a\| = \lim \|v^k - a\|$.

We also need the following lemma about the relation between the sequences $\{y^k\}$ and $\{z^k\}$.

LEMMA 3.4. $\lim(z^k - y^k) = 0$.

Proof. By Lemma 3.2, we only need to show $\lim(\|z^k - a\|^2 - \|y^k - a\|^2) = 0$. We prove it by contradiction. Assume, however, that this is not true. Then there exists a subsequence $\{n_k\}$ such that $\lim \|z^{n_k} - a\| > \lim \|y^{n_k} - a\|$. Because both $\{z^k\}$ and $\{y^k\}$ are bounded, we may assume, by extracting a further subsequence if necessary, that $\lim z^{n_k} = z^*$ and $\lim y^{n_k} = y^*$ for some points $z^* \in E$ and y^* . Of course, $z^* \neq y^*$. We first show that the step $\tau(z^*, y^*, E)$ is positive. It follows from the continuity of the ball mapping $B_i(z)$ and $y^{n_k} \in B(z^{n_k})$ that $y^* \in B(z^*)$. Therefore, the nonzero vector $y^* - z^*$ is a feasible direction of $B(z^*)$ at z^* , and hence it is also a feasible direction of E at z^* . Thus, the step $\tau(z^*, y^*, E) > 0$. It follows from $\|z^* - a\| > \|y^* - a\|$, $\tau(z^*, y^*, E) > 0$, and the strict convexity of the objective function that $\|v^* - a\| < \|z^* - a\|$ where $v^* = z^* + \tau(z^*, y^*, E)(y^* - z^*)$. By the continuity of the mapping $\tau(\cdot, \cdot, E)$ at (z^*, y^*) , we have $v^{n_k} \rightarrow v^*$ and therefore, for sufficiently large n_k , $\|v^{n_k} - a\| < \|z^* - a\|$. This is a contradiction, because from Lemma 3.3 we have $\|z^{n_k+1} - a\| \leq \|v^{n_k} - a\|$ and $\|z^k - a\| \geq \|z^* - a\|$ for all k . $\square$

The following lemma shows the close connection between (1) and the subproblem.

LEMMA 3.5. *A vector u^* is the unique optimal solution of (1) if and only if $u^* \in E$ and u^* is the optimal solution of the problem $\min\{\|x - a\| \mid x \in B(u^*)\}$.*

Proof. This follows directly from comparing the Karush–Kuhn–Tucker conditions of the two problems. $\square$

The following lemma is a special case of the standard parametric optimization. The results for much more general cases are contained in Proposition 23 of [2, p. 120] and the maximum theorem of [3, p. 116].

LEMMA 3.6. *Let $\{u^k\}$ be a sequence in E convergent to a point u^* . If, for each k , w^k solves $\min\{\|x - a\| \mid x \in B(u^k)\}$ and $\lim w^k = w^*$, then the point w^* solves the problem $\min\{\|x - a\| \mid x \in B(u^*)\}$.*

We now give our main convergence theorem below.

THEOREM 3.7. *The three sequences $\{z^k\}$, $\{v^k\}$, and $\{y^k\}$ all converge to the solution of (1).*

Proof. It follows from Lemma 3.4 that all three sequences have the same set of accumulation points because $\lim(z^k - y^k) = 0$ and v^k lies on the line segment $[z^k, y^k]$. Let u^* be an accumulation point; we want to show that u^* solves the problem. Let $\{n_k\}$ be a subsequence such that $\lim z^{n_k} = \lim y^{n_k} = u^*$. It follows from Lemma 3.6 that u^* solves $\min\{\|x - a\| \mid x \in B(u^*)\}$, and by Lemma 3.5, that the vector u^* is optimal to the original problem. By the strict convexity of the objective function, u^* is the unique optimal solution. It then follows from the boundedness of the three sequences that they all converge to u^* . $\square$

4. Projection on the intersection of several balls. In this section we briefly discuss how to project a point onto the intersection of several balls, an operation needed for Step 2(ii) of our projection algorithm. The problem can be expressed as

$$(4) \quad \begin{array}{ll} \text{Minimize} & \|x - a\|^2 \\ \text{subject to} & x \in B_i := \{x \mid \|x - a_i\|^2 \leq r_i^2\}, \quad i = 1, 2, \dots, m. \end{array}$$

Letting $B = \cap_{i=1}^m B_i$, we assume $\text{int}(B) \neq \emptyset$ and $a \notin B$.

As usual, the Lagrangian function is defined as

$$\begin{aligned} L(x, \lambda) &= \|x - a\|^2 + \sum_{i=1}^m \lambda_i (\|x - a_i\|^2 - r_i^2) \\ &= \left(1 + \sum_{i=1}^m \lambda_i\right) \left\|x - \frac{a + \sum_{i=1}^m \lambda_i a_i}{1 + \sum_{i=1}^m \lambda_i}\right\|^2 - \frac{\|a + \sum_{i=1}^m \lambda_i a_i\|^2}{1 + \sum_{i=1}^m \lambda_i} \\ &\quad + \sum_{i=1}^m \lambda_i (\|a_i\|^2 - r_i^2) + \|a\|^2, \end{aligned}$$

where $\lambda \in R_+^m$. The dual function is

$$\begin{aligned} g(\lambda) &= \inf_{x \in R^n} L(x, \lambda) \\ &= L\left(\frac{a + \sum_{i=1}^m \lambda_i a_i}{1 + \sum_{i=1}^m \lambda_i}, \lambda\right) \\ &= -\frac{\|a + \sum_{i=1}^m \lambda_i a_i\|^2}{1 + \sum_{i=1}^m \lambda_i} + \sum_{i=1}^m \lambda_i (\|a_i\|^2 - r_i^2) + \|a\|^2, \end{aligned}$$

and thus, the dual problem is defined as

$$(5) \quad \begin{array}{ll} \text{Maximize} & g(\lambda) \\ \text{subject to} & \lambda \in R_+^m. \end{array}$$

The primal problem is a convex programming problem satisfying the Slater condition, so strong duality holds [9]. Let x^* denote the primal optimal solution which is

unique due to the strict convexity of the primal objective function, and let λ^* denote a dual optimal solution. Since strong duality holds, we have

$$\begin{aligned}
& \|x^* - a\|^2 \\
&= g(\lambda^*) \\
&= \inf_{x \in R^n} L(x, \lambda^*) \\
&\leq L(x^*, \lambda^*) \\
&= \|x^* - a\|^2 + \sum_{i=1}^m \lambda_i^* (\|x^* - a_i\|^2 - r_i^2) \\
&\leq \|x^* - a\|^2,
\end{aligned}$$

where the last inequality is due to the fact that x^* is a feasible point of the primal problem and $\lambda^* \geq 0$. Thus, we get

$$L(x^*, \lambda^*) = \inf_{x \in R^n} L(x, \lambda^*) = L\left(\frac{a + \sum_{i=1}^m \lambda_i^* a_i}{1 + \sum_{i=1}^m \lambda_i^*}, \lambda^*\right).$$

Note that $L(x, \lambda^*)$ is a strictly convex function of x , so we must have

$$x^* = \frac{a + \sum_{i=1}^m \lambda_i^* a_i}{1 + \sum_{i=1}^m \lambda_i^*}.$$

From the above discussion, we see that to find x^* , we only need to solve the dual problem. The dual problem is a simple convex programming problem with only non-negative constraints and readily computable gradients and Hessians. Many standard algorithms can solve this problem very efficiently. For example, the nonmonotone spectral projected gradient method proposed in [4, 5] is very fast and involves only simple arithmetic computation, i.e., no matrix factorization is needed. In our current implementation, we use this method to solve the ball constrained subproblem.

Due to its special structure, the primal problem can also be attacked directly, for example, by interior point methods. There is a huge difference between applying interior point methods to (1) directly and applying them to (3) in terms of memory usage. For instance, on our computer, when $n = 500$ and $m = 50$, SeDuMi [10] (an excellent interior point method solver) cannot be applied to the general ellipsoid constrained projection problem since it needs too much memory, but it can solve the ball constrained projection problem in a few seconds and our algorithm can still proceed.

Both the nonmonotone spectral projected gradient method and the interior point method are excellent algorithms, but we do not think they fully exploit the special structure of this ball constrained subproblem. It will not be surprising that a special method dedicated to this problem can outperform these general methods. And this will be one of our future research topics.

5. Computational experiments. In this section, we present results from some preliminary numerical experiments. The testing environment is MATLAB 6 on a PC with 256 MB memory and dual Pentium III 800 MHz processors running the Red Hat Linux 7.2 operating system. We compared our algorithms with SeDuMi of version 1.05, which can be obtained from Dr. Jos F. Sturm's homepage <http://fewcal.uvt.nl/sturm/software/sedumi.html>.

In our implementation, we use the nonmonotone spectral projected gradient method to solve (3) and Scheme 1 with $\lambda = 1/\|A\|_\infty$ and $\beta = \sqrt{10}/\|A\|_\infty$ to generate the balls. The advantages and disadvantages of Schemes 1 and 2 are discussed in section 2.1. More analysis and comparison about the two schemes and different values of λ and β will be done later.

There are two possible ways to use SeDuMi. It can either be used directly to solve (1) [8] or as an intermediate step to solve (3). In this comparison, we use SeDuMi in the first way, i.e., to solve (1) directly.

To simplify our description, we use $a \in \text{Rand}(n, l, u)$ to indicate that a is a vector in R^n and that each entry of a is independently uniformly randomly chosen from (l, u) .

To create instances of our problem, we need a way to generate symmetric positive definite matrices given the dimension parameter n . We use the following method to do so.

Let I be the n -by- n identity matrix. For $i = 1, 2, 3$, let $w_i \in \text{Rand}(n, -1, 1)$, and $Q_i = I - \frac{2}{w_i^T w_i} (w_i w_i^T)$. Then we choose D to be an n -by- n diagonal matrix with each diagonal entry $\in \text{Rand}(1, 0, 100)$. Finally we set $U = Q_1 Q_2 Q_3$ and $A = U D U^T$ and take A as the output. (This is a modification of a method used in [1], which also employs a product of three Q_i 's.)

Let m denote the number of ellipsoids; we generate m symmetric positive definite matrices $\{A_1, \dots, A_m\}$ using the method we just described. Since we want to guarantee that the feasible region has a nonempty interior, we first get a point $p \in \text{Rand}(n, -500, 500)$ and proceed to make p a common interior point for all ellipsoids. For each $i = 1, \dots, m$, we let $b_i \in \text{Rand}(n, -100, 100)$, $s_i \in \text{Rand}(1, 0, 10)$, and $c_i = -(0.5 * p^T A_i p + b_i^T p + s_i)$, since $s_i > 0$, $p \in \text{int}(E_i)$. Finally we choose $a \in \text{Rand}(n, -500, 500)$.

Tables 1 to 10 display the numerical results. Each table contains the results of 5 testing problems. In those tables, we use ReToBa (reduce to balls) and SeDuMi to denote the two methods, respectively. The numbers in the corresponding rows represent the amount of time (seconds) the specific method spent on those problems. Let x_1 and x_2 denote the approximated solutions found by each method; we use Δ to denote the relative difference, i.e., $\Delta = \frac{\|x_1 - a\| - \|x_2 - a\|}{\|x_2 - a\|}$.

First we tried experiments on a small (n, m) pair to make sure both methods work. Tables 1 to 4 show the results of $n = m = 10$ and $n = m = 50$. It is easy to see that both methods perform similarly while SeDuMi usually produces better solutions. However, what we are really interested in is their performance in large cases. Tables 5 and 6 show the results of $n = 1000$ and $m = 5$. Tables 7 and 8 show the results of $n = 5$ and $m = 1000$. We can draw the conclusion that for big n and small m , our method works significantly better than SeDuMi. But for small n and big m , the opposite occurs. This is because we solve (3) (an n -dimensional problem) by using the nonmonotone spectral projected gradient method to solve the dual problem (an m -dimensional problem). It does not seem to be a perfect idea. A dedicated algorithm for (3) is indeed necessary.

Since our method uses only matrix-vector multiplication, its memory usage is very low. We tried one set of problems with parameters $n = 500$ and $m = 50$; SeDuMi always reports "Out of Memory" before the first iteration starts, while our method is still able to proceed and solve the problems.

These preliminary numerical tests by no means prove that our method is superior to the interior point methods. That was not our intention. Our research goal was to

TABLE 1
 $n = 10, m = 10.$

Method	Problem 1	Problem 2	Problem 3	Problem 4	Problem 5
ReToBa	0.12	0.13	0.12	2.81	0.01
SeDuMi	0.41	0.40	0.43	0.39	0.27
Δ	0.000068	0.000093	0.000042	0.000098	-1.00

TABLE 2
 $n = 10, m = 10.$

Method	Problem 6	Problem 7	Problem 8	Problem 9	Problem 10
ReToBa	0.26	0.15	2.17	0.18	0.11
SeDuMi	0.41	0.44	0.40	0.35	0.35
Δ	0.000097	0.000098	0.000052	0.000081	0.000071

TABLE 3
 $n = 50, m = 50.$

Method	Problem 1	Problem 2	Problem 3	Problem 4	Problem 5
ReToBa	12.33	0.77	16.76	0.04	0.68
SeDuMi	12.52	9.84	11.92	5.15	10.13
Δ	0.000098	0.000093	0.000094	-1	0.000032

TABLE 4
 $n = 50, m = 50.$

Method	Problem 6	Problem 7	Problem 8	Problem 9	Problem 10
ReToBa	45.47	13.29	39.93	0.04	24.08
SeDuMi	11.76	10.80	11.00	5.65	12.12
Δ	0.000094	0.000071	0.000073	-1	0.000061

TABLE 5
 $n = 1000, m = 5.$

Method	Problem 1	Problem 2	Problem 3	Problem 4	Problem 5
ReToBa	15.93	357.53	16.08	18.00	209.73
SeDuMi	2448.0	2024.2	2168.6	2171.45	2309.29
Δ	0.000044	0.000046	0.000032	0.000044	0.000067

TABLE 6
 $n = 1000, m = 5.$

Method	Problem 6	Problem 7	Problem 8	Problem 9	Problem 10
ReToBa	11.39	15.97	161.71	16.01	14.03
SeDuMi	1428.28	2181.75	2548.68	2166.71	2172.40
Δ	-1.00	0.000028	0.000065	0.000067	0.000072

TABLE 7
 $n = 5, m = 1000.$

Method	Problem 1	Problem 2	Problem 3	Problem 4	Problem 5
ReToBa	679.09	1001.40	88.72	229.05	64.76
SeDuMi	4.90	4.91	5.62	5.04	5.46
Δ	0.000089	0.000052	0.000089	0.000093	0.000095

TABLE 8
 $n = 5$, $m = 1000$.

Method	Problem 6	Problem 7	Problem 8	Problem 9	Problem 10
ReToBa	49.65	557.96	1003.82	1000.15	138.82
SeDuMi	4.99	5.40	5.16	4.74	5.57
Δ	0.000083	0.000094	0.00025	0.000098	0.000097

TABLE 9
 $n = 500$, $m = 50$.

Method	Problem 1	Problem 2	Problem 3	Problem 4	Problem 5
ReToBa	773.24	218.34	1113.88	1179.42	687.92
SeDuMi	failed	out	of	memory	!

TABLE 10
 $n = 500$, $m = 50$.

Method	Problem 6	Problem 7	Problem 8	Problem 9	Problem 10
ReToBa	741.47	210.58	51.03	245.87	1457.49
SeDuMi	failed	out	of	memory	!

find novel and promising ideas where the major focus is on the novelty. However, it is evident from the testing results that this new idea does have some potential and an advantage over the well-developed interior point methods, and is worth more research.

Acknowledgments. We thank the referees and the associate editor for the thoughtful comments and constructive suggestions that put this paper into its present form. The authors are particularly grateful to one referee for giving a better proof (the current one) of Lemma 3.2.

REFERENCES

- [1] L. T. H. AN, *An efficient algorithm for globally minimizing a quadratic function under convex quadratic constraints*, Math. Program., 87 (2000), pp. 401–426.
- [2] J.-P. AUBIN AND I. EKELAND, *Applied Nonlinear Analysis*, John Wiley, New York, 1984.
- [3] C. BERGE, *Topological Spaces*, Dover, Mineola, NY, 1997.
- [4] E. G. BIRGIN AND J. M. MARTÍNEZ, *Large-scale active-set box-constrained optimization method with spectral projected gradients*, Comput. Optim. Appl., 23 (2002), pp. 101–125.
- [5] E. G. BIRGIN, J. M. MARTÍNEZ, AND M. RAYDAN, *Nonmonotone spectral projected gradient methods on convex sets*, SIAM J. Optim., 10 (2000), pp. 1196–1211.
- [6] A. LIN, *Projection Algorithms in Nonlinear Programming*, Ph.D. thesis, Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, 2003.
- [7] A. LIN AND S.-P. HAN, *On the distance between two ellipsoids*, SIAM J. Optim., 13 (2002), pp. 298–308.
- [8] M. S. LOBO, L. VANDENBERGHE, S. BOYD, AND H. LEBRET, *Applications of second-order cone programming*, Linear Algebra Appl., 284 (1998), pp. 193–228 (special issue on linear algebra in control, signals, and image processing).
- [9] O. L. MANGASARIAN, *Nonlinear Programming*, SIAM, Philadelphia, 1994.
- [10] J. F. STURM, *Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones*, Optim. Methods Softw., 11-12 (1999), pp. 625–653 (version 1.05 available from <http://fewcal.kub.nl/sturm>).

ON SIZING AND SHIFTING THE BFGS UPDATE WITHIN THE SIZED-BROYDEN FAMILY OF SECANT UPDATES*

H. YABE[†], H. J. MARTÍNEZ[‡], AND R. A. TAPIA[§]

Abstract. In a numerical study, Contreras and Tapia demonstrated the value of selectively sizing the BFGS secant update. A conclusion of the current study is that there is value in following a sizing of the BFGS update with a shifting of the BFGS update. Our motivation for such a strategy comes from the known fact that the BFGS update is prone to producing updates with large eigenvalues. Hence, sizing may contribute to near singularity, and shifting can help compensate for this deficiency. We follow Contreras and Tapia in our form of selective sizing. Our form of shifting can be viewed as switching from the BFGS update to an alternative member of the Broyden family of updates. Our shift is obtained by minimizing a weighted form of the Byrd–Nocedal measure over a sized Broyden family of secant updates. Numerically, our most effective shift is the one that gives the member of the class which is closest to steepest descent in the Byrd–Nocedal measure.

Key words. sizing, BFGS update, secant update, quasi-Newton update, Newton’s method

AMS subject classifications. 49M15, 65K05, 65K10

DOI. 10.1137/S1052623496308801

1. Introduction. We consider quasi-Newton methods for the unconstrained optimization problem

$$\min_{x \in \mathcal{R}^n} f(x),$$

where f is twice continuously differentiable. Given x , an approximation to a minimizer x_* , to compute x_+ , a new approximation to x_* , these methods use a local quadratic model of the form

$$f(x+s) \approx f(x) + \nabla f(x)^T s + \frac{1}{2} s^T B s,$$

where B is the current approximation to $\nabla^2 f(x)$, the Hessian of f at x , and $\nabla f(x)$ is the gradient of f at x . Our particular interest is in *secant methods*, i.e., methods that have the property that the subsequent Hessian approximation B_+ satisfies the secant equation

$$(1.1) \quad B_+ s = y \equiv \nabla f(x_+) - \nabla f(x).$$

A well-known class of such approximations is the *Broyden family*, and it is generally accepted that the BFGS method is computationally the best method among

*Received by the editors September 3, 1996; accepted for publication (in revised form) March 5, 2004; published electronically October 14, 2004. This research was sponsored by NSF cooperative agreement CCR-88-09615, AFOSR grant 89-0363, DOE grant DEFG05-86-ER25017, and ARO grant 9DAAL03-90-G-0093.

<http://www.siam.org/journals/siopt/15-1/30880.html>

[†]Department of Mathematical Information Science, Tokyo University of Science, Tokyo, 1-3, Kagurazaka, Shinjuku-ku, Tokyo, Japan (yabe@rs.kagu.tus.ac.jp).

[‡]Departamento de Matemáticas, Universidad del Valle, A.A. 25360, Cali, Colombia (hector@mafalda.univalle.edu.co). The research of this author was partially supported by COLCIENCIAS CO: 1106-05-307-93.

[§]Department of Computational & Applied Mathematics, Rice University, Houston, TX 77251-1892 (rat@rice.edu).

these secant methods. However, the BFGS update may be unsatisfactory when B has very large eigenvalues. In order to overcome this difficulty, two remedies have been considered. The first is *sizing* (multiplying the approximate Hessian by an appropriate scalar before it is updated), and the second is *switching* (updating the approximate Hessian by a different member of the Broyden class).

In this paper, we combine both remedies and consider a new type of *sized Broyden family update*. We propose several sizing strategies as a safeguard for the BFGS method and present numerical experimentation designed to identify the most effective approach. We are interested in determining whether any of our approaches compare favorably with the numerically successful selective sizing suggested by Contreras and Tapia [7]. Preliminary numerical experiments suggest that one of our approaches is superior to the others, and this approach is also superior to the Contreras and Tapia BFGS centered Oren–Luenberger approach.

In the remainder of this section, we introduce our notation and present some history concerning sizing and switching. In section 2, we collect our main tools and discuss some of their properties. In section 3, we derive our sizing strategy. The formulas used in the numerical testing are summarized in section 4, and our numerical results are presented in section 5. Finally, in section 6 we give some concluding remarks.

Following Al-Baali [1], we consider the map

$$(1.2) \quad \text{BFGS}(X, s, y) = X - \frac{Xss^T X}{s^T X s} + \frac{yy^T}{y^T s}.$$

In terms of this map, the BFGS update of B can be written

$$B_+ = \text{BFGS}(B, s, y) = B - \frac{Bss^T B}{s^T B s} + \frac{yy^T}{y^T s},$$

and the Broyden family of updates of B can be written

$$(1.3) \quad B_+ = \text{BFGS}(B + \phi vv^T, s, y) = B - \frac{Bss^T B}{s^T B s} + \frac{yy^T}{y^T s} + \phi vv^T,$$

where ϕ is a scalar parameter and

$$(1.4) \quad v = \sqrt{s^T B s} \left(\frac{y}{y^T s} - \frac{Bs}{s^T B s} \right).$$

The equality in (1.3) is immediately transparent if one observes that $v^T s = 0$. Two important subclasses of this family are the *convex class*, $\phi \in [0, 1]$, and the *preconvex class*, $\phi < 0$. The BFGS update corresponds to the choice $\phi = 0$, and the DFP update corresponds to the choice $\phi = 1$.

Sufficient conditions for B_+ , obtained from the Broyden family given by (1.3), to be symmetric and positive definite are that B is symmetric and positive definite, $y^T s > 0$, and $\phi > \phi^\#$, where

$$(1.5) \quad \phi^\# = -\frac{1}{\tau - 1}, \quad \text{with } \tau = \frac{(s^T B s)(y^T B^{-1} y)}{(y^T s)^2}.$$

Note that $\tau \geq 1$, and if s and y are linearly independent, then $\tau > 1$.

In 1973/1974, Oren [15] and Luenberger [16] proposed a class of secant methods that they referred to as *self-scaling variable metric methods* (SSVMs). They modified

the convex class and obtained, for $\phi \in [0, 1]$,

$$\begin{aligned}
 B_+ &= \gamma \left(B - \frac{Bss^TB}{s^TBs} + \phi vv^T \right) + \frac{yy^T}{y^Ts} \\
 (1.6) \quad &= \text{BFGS}(\gamma(B + \phi vv^T), s, y) \\
 &= \text{BFGS}(\gamma B, s, y) + \gamma \phi vv^T.
 \end{aligned}$$

For a fixed value of ϕ , the parameter γ was chosen so that $\kappa_2(Q^{-1/2}B_+Q^{-1/2}) \leq \kappa_2(Q^{-1/2}BQ^{-1/2})$, where $\kappa_2(A)$ is the l_2 -norm condition number of A , and the matrix Q is the symmetric positive definite Hessian of f (which is constant because they assumed f to be quadratic).¹ They suggested two choices for the parameter γ :

$$(1.7) \quad \gamma_{OL} = \frac{y^Ts}{s^TBs} \quad \text{and} \quad \gamma_{IOL} = \frac{y^TB^{-1}y}{y^Ts}.$$

Soon after, in 1976, Oren and Spedicato [17] used the l_2 -norm condition number, proposed earlier by Davidon [8], to derive optimal parameters ϕ^* and γ^* for the sized-Broyden family (1.6). Since these choices were, in some sense, optimal, they intended them to be used at every iteration. However, in 1978, Shanno and Phua [18] claimed that choosing ϕ^* and γ^* fresh at each iteration did not perform well and recommended always using the BFGS update ($\phi = 0$) and sizing only at the first iteration. They suggested sizing with γ_{OL} , given in (1.7).

In 1981, Dennis, Gay, and Welsch [9] proposed their NL2SOL algorithm for the nonlinear least-squares problem. Their basic algorithm was a Gauss–Newton, trust-region method. However, in order to handle large residual problems, they maintained a structured DFP secant approximation to the second-order part of the least-squares Hessian and adaptively decided when to use this approximation. A major concern was that, since secant methods do not generate Hessian approximations that become arbitrarily accurate as the iteration proceeds, the approximation to the second-order information may impede the fast convergence for zero-residual problems. They observed that γ_{OL} in (1.7) had the property that the interval spectrum of $\gamma_{OL}B$ overlaps the interval spectrum of $\nabla^2 f(x + \theta s)$ for some $\theta \in (0, 1)$. For this reason, they decided to call multiplication by γ_{OL} “sizing” instead of “scaling.” More importantly, they realized that in their application sizing with γ_{OL} would guarantee that their approximation to the second-order information would converge to zero when the second-order information was zero at the solution.

Contreras and Tapia [7] proposed a selective sizing technique in the framework of the BFGS trust-region method using the centered Oren–Luenberger sizing factor, defined as

$$(1.8) \quad \gamma_{COL} = \frac{\frac{y_-^Ts_-}{s_-^Ts_-} + \frac{y^Ts}{s^Ts}}{\frac{s_-^TBs_-}{s_-^Ts_-} + \frac{s^TBs}{s^Ts}},$$

where s_- and y_- are calculated at the previous iteration. Essentially, Contreras and Tapia suggested sizing only when $\gamma_{OL} < 1$. They felt that γ_{OL} sized the BFGS update too much and suggested γ_{COL} because, for $\gamma_{OL} < 1$, it follows that $\gamma_{OL} < \gamma_{COL} < 1$. Hence, they viewed γ_{COL} as a damping or softening of γ_{OL} .

¹For a positive definite matrix A , the notation $A^{1/2}$ will be used to denote the symmetric positive definite matrix such that $(A^{1/2})^2 = A$.

Their numerical experiments showed that, without sizing, DFP is vastly inferior to BFGS; however, when selectively sized, DFP is competitive with BFGS. In their experiments, DFP performed best with the Oren–Luenberger factor γ_{OL} , and BFGS fit well with the centered Oren–Luenberger factor γ_{COL} for medium dimensional problems. They concluded that their selective sizing with factor γ_{COL} was superior to the alternatives of never sizing, sizing only at the first iteration, or sizing at every iteration.

Several other measures have been used to derive similar updates. These scalings are optimal with respect to the measures used to derive them.

Byrd and Nocedal [4] introduced the measure $\psi(A) = \text{Tr}(A) - \ln[\det(A)]$ as a tool for analyzing the global behavior of quasi-Newton methods. The ψ -measure was a critical ingredient in the elegant global convergence theory established for methods from the convex Broyden class (except DFP) by Byrd, Nocedal, and Yuan [5]. Fletcher [12] studied the ψ -measure and used it to establish yet another optimality characterization of the BFGS and DFP secant updates.

Dennis and Wolkowicz [11] used the measure $\omega(A) = \frac{\text{Tr}(A)}{n[\det(A)]^{1/n}}$ to give a specific sense in which γ_{OL} (see (1.7)) is the optimal sizing factor for the DFP update, and γ_{IOL} (see (1.7)) is the optimal sizing factor for the BFGS update.

Todd [20], following Fletcher’s approach, used the measure $\varphi(A) = n \ln[n\omega(A)] = n \ln \text{Tr}(A) - \ln \det(A)$ to provide an alternative proof of the result of Dennis and Wolkowicz [11], mentioned above.

Wolkowicz [21] discussed the measure $\sigma(A) = \frac{\lambda_1(A)}{\det(A)^{1/n}}$, which yields optimally conditioned, sized, symmetric rank-one updates. In this measure, $\lambda_1(A)$ denotes the largest eigenvalue of A .

Byrd, Liu, and Nocedal [3] used the ψ -function to study the preconvex class, which was first analyzed by Zhang and Tewarson [23]. They claimed that negative values of ϕ in (1.3) are desirable. They also gave conditions on ϕ that guaranteed superlinear convergence. Moreover, they minimized $\psi(D^{-1/2}B_+D^{-1/2})$ with respect to ϕ (with $\gamma = 1$) to obtain the optimal parameter ϕ^* . Byrd, Liu, and Nocedal used $D = \nabla^2 f(x_c)$ in their numerical experiments. Clearly, this D is not available in a secant environment. However, since its use led to good numerical results, a certain validity of the theory has been demonstrated.

Wolkowicz and Zhao [22] presented another measure and derived least change secant updates by using the framework of multicriteria optimization. They also presented the Broyden efficient class and the self-scaling efficient region.

Other studies of scaling may be found in Siegel [19] and Hu and Storey [13].

2. Preliminaries. The previous comments lead us to consider obtaining secant updates by embedding the Broyden family of updates in a larger parametric class of secant updates and then selecting the parameters that optimize a “goodness” or “conditioning” measure. Toward this end, we now motivate our choice for the measure function and our choice for the parametric class.

Our choice for the measure is the Byrd–Nocedal ψ -measure

$$(2.1) \quad \psi(A) = \text{Tr}(A) - \ln[\det(A)].$$

This measure involves all the eigenvalues of A ; more traditional condition measures involve only the largest and smallest eigenvalue. The numerical results contained in section 5 lead us to believe that our idea has considerable merit and deserves further study.

The following known properties of ψ will be important in our theory. A proof can be found in Fletcher [12].

THEOREM 2.1. *The function ψ defined by (2.1) is strictly convex on the set*

$$\Gamma = \{M \in \mathcal{R}^{n \times n} \mid M \text{ is symmetric and positive definite}\}.$$

Furthermore, $\psi(A)$ is globally and uniquely minimized by $A = I$ over the set Γ .

We now motivate our parametric family of secant updates. Historically, it has been observed that the DFP update tends to produce updates with small eigenvalues, while the BFGS update tends to produce updates with large eigenvalues. Hence, in working with matrices that have been generated by the BFGS update, there is a danger that a particular sizing factor may be excessively small, and the sized matrix γB may be nearly singular. In order to overcome this difficulty, in addition to sizing, we consider shifting the sized matrix γB by adding a matrix M . We begin with an update of the form

$$(2.2) \quad B_+ = \text{BFGS}(\gamma B + M, s, y).$$

In (2.2), γ is a sizing factor that makes the large eigenvalues small. Note that, for any γ and M , the update $\text{BFGS}(\gamma B + M, s, y)$ satisfies the secant condition (1.1). If we make an appropriate choice of M , then the shift can be performed *after* the sized matrix is updated. We will use $M = \phi vv^T$, where v is given by (1.4). Then

$$(2.3) \quad B_+ = \text{BFGS}(\gamma B + \phi vv^T, s, y) = \text{BFGS}(\gamma B, s, y) + \phi vv^T.$$

Note that the new sized-Broyden family (2.3) is somewhat different from the Oren–Luenberger family (1.6). The essential difference between them is that in (2.3) the sizing and shifting are independent of each other. Shifting the spectrum by a rank-one update was discussed by Dennis and Wolkowicz [11]. While they tried to avoid a drastic change in the whole spectrum by shifting after sizing within the framework of the ω -measure, our approach uses the Byrd–Nocedal measure to deal with the parameters ϕ and γ independently.

For the sized-Broyden family (2.3), we have the following lemma. Its proof follows from the sufficient conditions stated above (see (1.5)) for B_+ in the Broyden family (1.3) to be symmetric and positive definite.

LEMMA 2.2. *Assume that B is symmetric and positive definite and that $y^T s > 0$. If $\gamma > 0$ and $\phi > \gamma \phi^\#$, then B_+ given by (2.3) is symmetric and positive definite.*

Our objective is to make the best choice of parameters γ and ϕ in (2.3). Oren and Spedicato [17] chose parameters γ and ϕ for the Oren–Luenberger sized-Broyden class (1.6) that minimized an upper bound on $\kappa_2(D^{-1/2}B_+D^{-1/2})$. Here, $D = B$ was used.

We will use the ψ -measure (2.1) instead of κ_2 to compute optimal values of γ and ϕ in (2.3). We consider minimizing $\psi(D^{-1/2}B_+D^{-1/2})$ for some symmetric positive definite matrix D and investigate the performance for $D = B$, $D = I$, and $D = \nabla^2 f(x)$.

We will consider three types of minimization problems applied to our sized-Broyden class (2.3). First, for a given sizing factor γ , we find the optimal secant update in the sized classes by minimizing ψ over the appropriate class to determine the optimal value of ϕ . Second, we do the above with the roles of γ and ϕ interchanged to find an optimal sizing factor γ for a fixed member of the Broyden class (i.e., fixed ϕ). Finally, we determine the optimal sized member of each sized-Broyden class by minimizing ψ over (2.3) in both γ and ϕ .

The following algebraic properties of the map (1.2) will be very useful in the next two sections. Their proofs can be obtained from straightforward linear algebra manipulations.

PROPOSITION 2.3. *Assume that $s^T X s \neq 0$ and $y^T s \neq 0$. For any $n \times n$ matrix X ,*

1. *the secant condition $[\text{BFGS}(X, s, y)]s = y$ is satisfied.*
2. *if X is symmetric, then $\text{BFGS}(X, s, y)$ is also symmetric.*
3. *if X is symmetric and positive definite and $y^T s > 0$, then $\text{BFGS}(X, s, y)$ is also symmetric and positive definite.*
4. *if X is symmetric, then*

$$\text{Tr}[\text{BFGS}(X, s, y)] = \text{Tr}(X) - \frac{\|Xs\|^2}{s^T X s} + \frac{\|y\|^2}{y^T s}.$$

5. *we have*

$$\det[\text{BFGS}(X, s, y)] = \det(X) \frac{y^T x}{s^T X s}.$$

3. ψ -Optimal values for the new sized-Broyden family. In this section, we derive weighted ψ -optimal values for γ and ϕ in our new sized-Broyden family (2.3). Recall that this new family is different from the Oren–Luenberger family (1.6). The important difference between (1.6) and (2.3) is that in (2.3), the matrix $\phi v v^T$ is used as a shifting operator, and only the matrix B is sized. By doing so, we can separately control the sizing parameter and the shifting term.

Let D be a given symmetric positive definite matrix. We introduce D into our optimization problem by working with the quantities

$$(3.1) \quad \begin{aligned} \hat{B} &= D^{-1/2} B D^{-1/2}, & \hat{y} &= D^{-1/2} y, \\ \hat{B}_+ &= D^{-1/2} B_+ D^{-1/2}, & \hat{s} &= D^{1/2} s \end{aligned}$$

and then considering the D -weighted Byrd–Nocedal measure $\psi(\hat{B}_+)$. When $n = 1$, the secant condition uniquely determines $B_+ = y/s$, so we consider the case $n \geq 2$ in the following discussion.

Observe that, since the Byrd–Nocedal measure ψ is uniquely and globally minimized by the identity matrix, there will be a tendency to choose parameters that, in some sense, force $\hat{B}_+$ to approximate the identity. Now, if $\hat{B}_+ \approx I$, then $B_+ \approx D$. Hence the choice $D = B$ gives a least change or secant flavor, the choice $D = \nabla^2 f(x)$ gives a Newton’s method flavor, and the choice $D = I$ gives a steepest descent method flavor. We ask the reader to keep these comments in mind throughout the remainder of the paper and, in particular, in the numerical experiments section. At this point in the journey, it is not at all clear which choice of D will win numerically. Indeed, this was a significant part of our motivation for including all three choices in the present study.

We have

$$(3.2) \quad \begin{aligned} \hat{B}_+ &= \text{BFGS}(\gamma \hat{B} + \phi \hat{v} \hat{v}^T, \hat{s}, \hat{y}) \\ &= \gamma \left(\hat{B} - \frac{\hat{B} \hat{s} \hat{s}^T \hat{B}}{\hat{s}^T \hat{B} \hat{s}} \right) + \phi \hat{v} \hat{v}^T + \frac{\hat{y} \hat{y}^T}{\hat{y}^T \hat{s}}, \end{aligned}$$

where

$$\hat{v} = \sqrt{\hat{s}^T \hat{B} \hat{s}} \left(\frac{\hat{y}}{\hat{y}^T \hat{s}} - \frac{\hat{B} \hat{s}}{\hat{s}^T \hat{B} \hat{s}} \right) = D^{-1/2} v,$$

and v is given by (1.4). Note that some quantities are invariant under the scaling (3.1); i.e.,

$$\hat{y}^T \hat{s} = y^T s, \quad \hat{s}^T \hat{B} \hat{s} = s^T B s, \quad \hat{y}^T \hat{B}^{-1} \hat{y} = y^T B^{-1} y, \quad \hat{v}^T \hat{B}^{-1} \hat{v} = v^T B^{-1} v = \tau - 1.$$

However, others are not invariant; i.e.,

$$\begin{aligned} \|\hat{y}\|^2 &= y^T D^{-1} y, & \|\hat{B} \hat{s}\|^2 &= \|D^{-1/2} B s\|^2 = s^T B D^{-1} B s, \\ \|\hat{v}\|^2 &= v^T D^{-1} v = (s^T B s) \left[\frac{y^T D^{-1} y}{(y^T s)^2} - 2 \frac{y^T D^{-1} B s}{(y^T s)(s^T B s)} + \frac{s^T B D^{-1} B s}{(s^T B s)^2} \right]. \end{aligned}$$

Thus an application of Proposition 2.3 to expression (3.2) yields

$$\begin{aligned} \text{Tr}(\hat{B}_+) &= \gamma \left(\text{Tr}(\hat{B}) - \frac{\|\hat{B} \hat{s}\|^2}{\hat{s}^T \hat{B} \hat{s}} \right) + \phi \|\hat{v}\|^2 + \frac{\|\hat{y}\|^2}{\hat{y}^T \hat{s}} \\ &= \gamma \left(\text{Tr}(\hat{B}) - \frac{s^T B D^{-1} B s}{s^T B s} \right) + \phi v^T D^{-1} v + \frac{y^T D^{-1} y}{y^T s}, \end{aligned}$$

and

$$\begin{aligned} \det(\hat{B}_+) &= \det[\text{BFGS}(\gamma \hat{B} + \phi \hat{v} \hat{v}^T, \hat{s}, \hat{y})] \\ &= \gamma^{n-1} \det(\hat{B}) \left(1 + \frac{\phi}{\gamma} (\tau - 1) \right) \frac{y^T s}{s^T B s}, \end{aligned}$$

where τ is defined in (1.5).

We will be interested in minimizing $\psi(\hat{B}_+)$ in γ or ϕ or both for a fixed matrix D . Hence, we must study $\psi(\hat{B}_+)$. Using the preceding results, we have

$$\begin{aligned} \psi(\hat{B}_+) &= \text{Tr}(\hat{B}_+) - \ln[\det(\hat{B}_+)] \\ (3.3) \quad &= \gamma \xi + \phi v^T D^{-1} v + \frac{y^T D^{-1} y}{y^T s} \\ &\quad - (n-2) \ln \gamma - \ln[\det(\hat{B})] - \ln[\gamma + \phi(\tau - 1)] - \ln \left[\frac{y^T s}{s^T B s} \right], \end{aligned}$$

where

$$(3.4) \quad \xi = \text{Tr}(\hat{B}) - \frac{s^T B D^{-1} B s}{s^T B s}.$$

Since $\text{Tr}(\hat{B})$ is the sum of the eigenvalues of $D^{-1/2} B D^{-1/2}$ and the Rayleigh quotient

$$\frac{s^T B D^{-1} B s}{s^T B s} = \frac{(B^{1/2} s)^T B^{1/2} D^{-1} B^{1/2} (B^{1/2} s)}{(B^{1/2} s)^T (B^{1/2} s)}$$

lies between the smallest and the largest eigenvalues of $B^{1/2} D^{-1} B^{1/2}$, or equivalently² of $D^{-1/2} B D^{-1/2}$, we have

$$\xi = \text{Tr}(\hat{B}) - \frac{s^T B D^{-1} B s}{s^T B s} > 0.$$

²If (λ, x) is an eigenvalue-eigenvector pair of $B^{1/2} D^{-1} B^{1/2}$, then $(\lambda, D^{1/2} B^{-1/2} x)$ is an eigenvalue-eigenvector pair of $D^{-1/2} B D^{-1/2}$.

Then we have

$$\begin{aligned}\frac{\partial\psi}{\partial\phi} &= v^T D^{-1}v - \frac{\tau-1}{\gamma + \phi(\tau-1)}, \\ \frac{\partial\psi}{\partial\gamma} &= \xi - \frac{n-2}{\gamma} - \frac{1}{\gamma + \phi(\tau-1)}, \\ \frac{\partial^2\psi}{\partial\phi^2} &= \frac{(\tau-1)^2}{[\gamma + \phi(\tau-1)]^2} > 0, \\ \frac{\partial^2\psi}{\partial\gamma^2} &= \frac{n-2}{\gamma^2} + \frac{1}{[\gamma + \phi(\tau-1)]^2} > 0,\end{aligned}$$

and

$$\frac{\partial^2\psi}{\partial\phi\partial\gamma} = \frac{\partial^2\psi}{\partial\gamma\partial\phi} = \frac{\tau-1}{[\gamma + \phi(\tau-1)]^2} > 0.$$

We will consider the following minimization problems.

MINIMIZATION PROBLEM 1.

$$(3.5) \quad \min_{\phi > \gamma\phi^\#} \psi(\hat{B}_+) \quad \text{for a given } \gamma(>0).$$

Of course, we need an optimal value ϕ^* such that $\phi^* > \gamma\phi^\#$ (see Lemma 2.2), which will be the case. Since $d^2\psi/d\phi^2 > 0$, ψ is a strictly convex function in ϕ . Hence $d\psi/d\phi = 0$ yields the unique global solution of Minimization Problem 1

$$\phi^* = \frac{1}{v^T D^{-1}v} - \frac{\gamma}{\tau-1} = \frac{1}{v^T D^{-1}v} + \gamma\phi^\# > \gamma\phi^\#.$$

The choice $D = B$ gives

$$\phi^* = \frac{1-\gamma}{\tau-1}.$$

Note that $\gamma = 1$ (no sizing) implies $\phi^* = 0$ (BFGS). For $\gamma = \gamma_{OL}$ and $\gamma = \gamma_{IOL}$, we have

$$\phi^* = \frac{(y^T s)^2 s^T (Bs - y)}{(s^T Bs)[(s^T Bs)(y^T B^{-1}y) - (y^T s)^2]}$$

and

$$\phi^* = \frac{(y^T s)y^T (s - B^{-1}y)}{(s^T Bs)(y^T B^{-1}y) - (y^T s)^2},$$

respectively.

MINIMIZATION PROBLEM 2.

$$(3.6) \quad \min_{\gamma > 0} \psi(\hat{B}_+) \quad \text{for a given } \phi(>\gamma\phi^\#).$$

As before, since $d^2\psi/d\gamma^2 > 0$, ψ is a strictly convex function of γ ; hence the equation $d\psi/d\gamma = 0$ implies

$$(3.7) \quad \xi - \frac{n-2}{\gamma} - \frac{1}{\gamma + \phi(\tau-1)} = 0.$$

For the case $n = 2$, we obtain

$$\gamma^* = \frac{1}{\xi} - \phi(\tau - 1)$$

for a given constant ϕ satisfying $1/\xi(\tau - 1) > \phi \geq 0$. In this case, $\gamma^* > 0$ and $\phi > \gamma^*\phi^\#$.

When $n > 2$, (3.7) yields a quadratic equation in γ :

$$\xi\gamma^2 + [\xi\phi(\tau - 1) - (n - 1)]\gamma - \phi(\tau - 1)(n - 2) = 0.$$

Thus we have

$$\gamma = \frac{1}{2\xi}[(n - 1) - \xi\phi(\tau - 1) \pm \sqrt{\delta}],$$

where

$$\delta = [\xi\phi(\tau - 1) - (n - 1)]^2 + 4\xi\phi(\tau - 1)(n - 2).$$

Considering the positivity of γ , we choose

$$\gamma^* = \frac{1}{2\xi}[(n - 1) - \xi\phi(\tau - 1) + \sqrt{\delta}].$$

Now we investigate the positivity of γ . For the choice $\phi > 0$, we have

$$\delta = [\xi\phi(\tau - 1) - (n - 1)]^2 + 4\xi\phi(\tau - 1)(n - 2) > 0.$$

For the choice $\phi \leq 0$, we have

$$\delta = [\xi\phi(\tau - 1) + (n - 1)]^2 - 4\xi\phi(\tau - 1) > 0.$$

Then, for any ϕ ,

$$\gamma^* = \gamma^*(\phi) = \frac{1}{2\xi}[(n - 1) - \xi\phi(\tau - 1) + \sqrt{\delta}] > 0,$$

so we must choose a parameter ϕ such that $\phi > \gamma^*(\phi)\phi^\#$.

The choice $D = B$ gives $\xi = n - 1$; therefore

$$\gamma^* = \frac{1}{2} \left(1 - \phi(\tau - 1) + \sqrt{[\phi(\tau - 1) - 1]^2 + 4\phi(\tau - 1)\frac{n-2}{n-1}} \right).$$

Note that $\phi = 0$ implies that $\gamma = 1$.

MINIMIZATION PROBLEM 3.

$$(3.8) \quad \min_{\gamma > 0, \phi > \gamma\phi^\#} \psi(\hat{B}_+).$$

For convenience, let

$$\alpha = \frac{1}{\gamma + \phi(\tau - 1)}.$$

Then we can write

$$\nabla^2 \psi(\phi, \gamma) = \begin{bmatrix} \alpha^2(\tau - 1)^2 & \alpha^2(\tau - 1) \\ \alpha^2(\tau - 1) & \alpha^2 + \frac{(n-2)}{\gamma^2} \end{bmatrix}.$$

Consider first the case $n = 2$. We note that $\nabla^2 \psi(\phi, \gamma)$ is a positive semidefinite matrix. Setting $\nabla \psi = 0$, we obtain

$$(3.9) \quad \begin{aligned} v^T D^{-1} v - \alpha(\tau - 1) &= 0, \\ \xi - \alpha &= 0. \end{aligned}$$

This is a degenerate system, and the equations are consistent if $\xi(\tau - 1) = v^T D^{-1} v$. Any stationary point (ϕ^*, γ^*) satisfying $\xi = \alpha$ is a global minimizer. In the case $D = B$, we have

$$\xi = 1 = \frac{v^T D^{-1} v}{\tau - 1},$$

and (3.9) is a consistent system.

Now assume that $n > 2$. Since $\partial^2 \psi / \partial \phi^2 = \alpha^2(\tau - 1)^2 > 0$, and

$$\det[\nabla^2 \psi(\phi, \gamma)] = \alpha^2(\tau - 1)^2 \left(\frac{n-2}{\gamma^2} \right) > 0,$$

the function $\psi(\hat{B}_+)$ given by (3.3) is strictly convex with respect to (ϕ, γ) . Thus, for Minimization Problem 3, a stationary point is the unique optimal point. The equations

$$\frac{\partial \psi}{\partial \phi} = 0 \quad \text{and} \quad \frac{\partial \psi}{\partial \gamma} = 0$$

yield

$$\frac{1}{\gamma + \phi(\tau - 1)} = \frac{v^T D^{-1} v}{\tau - 1}$$

and

$$\xi - \frac{n-2}{\gamma} - \frac{1}{\gamma + \phi(\tau - 1)} = 0.$$

Therefore, we obtain the unique global solution for Minimization Problem 3:

$$\begin{aligned} \gamma^* &= \frac{n-2}{\left[\xi - \frac{v^T D^{-1} v}{\tau-1} \right]} = \left(\frac{1}{\phi^\# - \phi^{\#\#}} \right) \left(\frac{n-2}{v^T D^{-1} v} \right), \\ \phi^* &= \frac{1}{v^T D^{-1} v} - \frac{\gamma^*}{\tau-1} = \frac{1}{v^T D^{-1} v} + \gamma^* \phi^\# > \gamma^* \phi^\#. \end{aligned}$$

In the expression for γ^* , we have used the new quantity $\phi^{\#\#}$, defined by

$$\phi^{\#\#} = -\frac{\xi}{v^T D^{-1} v}.$$

The choice $D = B$ gives $\gamma^* = 1$ and $\phi^* = 0$. This implies that the BFGS update without sizing is optimal for this choice of weighting matrix.

Summarizing the preceding results, we obtain the following proposition.

PROPOSITION 3.1.

1. The parameter ϕ^* obtained as the solution of Minimization Problem 1 is greater than $\gamma\phi^\#$.
2. The parameter γ^* obtained as the solution of Minimization Problem 2 is positive for all ϕ . (Recall that for B_+ to be positive definite, the parameter ϕ must be chosen so that $\phi > \gamma^*(\phi)\phi^\#$.)
3. The parameter ϕ^* obtained as the solution of Minimization Problem 3 is greater than $\gamma^*\phi^\#$. If $\phi^\# > \phi^{\#\#}$, then the parameter γ^* obtained as the solution of Minimization Problem 3 is positive.

4. Summary of results. Assume that B is symmetric and positive definite, y and s are linearly independent, and $y^T s > 0$. For Minimization Problem 1, we obtained, for $\gamma > 0$,

$$\phi^* = \frac{1}{v^T D^{-1} v} + \gamma^* \phi^\#.$$

This optimal value always produces symmetric positive definite updates.

For Minimization Problem 2, using a given ϕ , we obtained

$$\gamma^* = \frac{1}{\xi} - \phi(\tau - 1),$$

when $n = 2$, and

$$\gamma^* = \frac{1}{2\xi} \left[(n-1) - \xi\phi(\tau-1) + \sqrt{[\xi\phi(\tau-1) - (n-1)]^2 + 4\xi\phi(\tau-1)(n-2)} \right],$$

when $n > 2$. In this case, the optimal value does not always produce a positive definite update. However, $\phi > 0$ is sufficient, but not necessary, for the optimal value to produce symmetric positive definite updates.

For Minimization Problem 3 in the case $n > 2$, we obtained

$$\phi^* = \frac{1}{v^T D^{-1} v} + \gamma^* \phi^\#$$

and

$$\gamma^* = \frac{n-2}{\xi + \phi^\# (v^T D^{-1} v)}.$$

When $n = 2$, we require that ϕ^* and γ^* satisfy

$$\gamma^* + \phi^*(\tau - 1) = \frac{1}{\xi}.$$

Once again, even though these optimal values do not always produce positive definite updates, $\phi \geq 0$ is sufficient, but not necessary, for these optimal values to produce symmetric positive definite updates.

Finally, observe that, for the choice $D = B$, the standard BFGS update ($\phi = 0, \gamma = 1$) is optimal for Minimization Problem 1 with $\gamma = 1$, for Minimization Problem 2 with $\phi = 0$, and for Minimization Problem 3. This is certainly an interesting finding and calls for more understanding.

5. Numerical experiments. In this section, we present numerical experimentation using various forms of the sizing strategies described in section 3. We compare the basic BFGS method with approaches using numerous combinations of sizing, shifting, and weighting factors. Our objective in these preliminary tests is to identify the most effective of these combinations. In addition to the approaches based on the minimization problems in section 3, we include in our numerical studies the Contreras–Tapia approach and approaches derived by applying the ideas of section 3 to the Oren–Luenberger family (1.6). Contreras and Tapia [7] obtained impressive numerical results for the BFGS trust-region method using selective sizing and a sizing factor γ_{COL} that they called *centered Oren–Luenberger*. In our testing, we compared γ_{COL} with γ_{OL} and with optimal values of γ derived from Minimization Problem 2 and Minimization Problem 3. Our sizing strategies were adapted to both the trust-region and line-search algorithms in an enhanced MATLAB version of the well-known Dennis–Schnabel code [10]. For the sake of convenience, we chose to carry over to our code from the Dennis–Schnabel code the double-dogleg implementation of trust-region globalization and the backtracking implementation of line-search globalization.

Our preliminary numerical experiments indicate that one approach is superior to the others, and this approach is also superior to the Contreras–Tapia BFGS approach. The best sizing procedure was derived from our new sized-Broyden family (2.3) as the solution of Minimization Problem 1 with $D = I$ and $\gamma = \gamma_{COL}$, the centered Oren–Luenberger sizing introduced by Contreras and Tapia. Later on, we call this approach I-COL-P1-new. In fact, method I-COL-P1-new was efficient (see Table 5.3), robust to parameter changes (see Tables 5.5 and 5.6), and reliable as the problem size increased (see Table 5.7).³ See section 6 for a plausible explanation as to why I-COL-P1-new is our most effective sizing strategy.

5.1. Optimal parameters for the Oren–Luenberger SSVM method. In section 3, we discussed a new sizing strategy using optimal parameters obtained by solving three kinds of minimization problems. We can derive optimal parameters for the Oren–Luenberger SSVM method in the same way. In order to distinguish these parameters from those computed as part of our “new” sizing strategy, we will use the subscript “old” to denote the optimal parameters computed using the Oren–Luenberger family. For Minimization Problem 1, assuming $\gamma > 0$,

$$\phi_{old}^* = \frac{1}{\gamma(v^T D^{-1}v)} + \phi^\#.$$

For Minimization Problem 2 we have, for given ϕ ,

$$\gamma_{old}^* = \frac{n-1}{\xi + \phi(v^T D^{-1}v)},$$

where ξ is given by (3.4). For Minimization Problem 3, assuming that $n > 2$ and $\xi + \phi^\#(v^T D^{-1}v) > 0$, we find that

$$\begin{aligned} \gamma_{old}^* &= \frac{n-2}{\xi + \phi^\#(v^T D^{-1}v)}, \\ \phi_{old}^* &= \frac{1}{\gamma_{old}^*(v^T D^{-1}v)} + \phi^\#. \end{aligned}$$

³Other methods were “equally efficient” but not as robust and reliable. See methods BFGS-COL, I-OL-P1-old, and I-OL-P1-new, for example.

When $n = 2$, the system obtained by setting the first partial derivatives of ψ to zero is degenerate. The equations are consistent if $\xi(\tau - 1) = v^T D^{-1}v$, and the equations are solved by any (ϕ^*, γ^*) that satisfies

$$(v^T D^{-1}v)\phi - \frac{1}{\gamma} + \xi = 0, \quad \phi > \phi^\#, \quad \gamma > 0.$$

The optimal value ϕ_{old}^* obtained from Minimization Problem 1 always produces symmetric positive definite updates. The updates obtained from the solutions of Minimization Problem 2 and Minimization Problem 3 may not be positive definite. However, the condition $\phi \geq 0$ is sufficient, although not necessary, for the optimal values in these two cases to produce symmetric positive definite updates.

5.2. Implementation. We incorporated into our code the options of *shifting* and *sizing* the approximation to the Hessian before it is updated. Thus, before the BFGS updating formula is applied to a matrix B , we have the option of modifying B by either

$$(5.1) \quad B \leftarrow \gamma(B + \phi vv^T) \quad (\text{Oren-Luenberger (old) sizing (1.6)})$$

or

$$(5.2) \quad B \leftarrow \gamma B + \phi vv^T \quad (\text{new sizing (2.3)}).$$

Here, γ and ϕ are real scalars, and $v \in \mathcal{R}^n$ is given by (1.4).

We also included in the code the ability to accept an arbitrary initial matrix B_0 . Our choices for the initial matrix B_0 were

$$(5.3) \quad \begin{aligned} B_0^0 &= |f(x_0)|I, \quad (x_0 \text{ is the starting point}) \\ B_0^1 &= \text{diag}(10, 10^4, 1, \dots, 1), \\ B_0^2 &= \text{diag}(10^{-1}, 10^{-4}, 1, \dots, 1), \\ B_0^3 &= \text{diag}(0.1, 2, 0.5, 10, 1, \dots, 1). \end{aligned}$$

Observe that B_0^0 has representative eigenvalues, B_0^1 has large eigenvalues, B_0^2 has small eigenvalues, and B_0^3 has moderate-sized eigenvalues. In all of our experiments, we allowed a maximum of 500 iterations. The stopping criterion employed is the same as that used in the Dennis-Schnabel optimization code. In particular, convergence is reached when the relative gradient is less than 10^{-7} . The machine used to obtain these numerical results was a SUN SPARC-10/Model 41.

Our test set consists of five problems. Four of them are problems 1, 2, 5, and 21 (with $n = 6$ and $n = 20$) from the standard set of test functions in Moré, Garbow, and Hillstom [14]. The remaining problem is the strictly convex quartic function of Byrd, Liu, and Nocedal [3]. This latter function was also used by Al-Baali [2] in similar numerical testing. For each problem, we chose the standard starting point included in the relevant reference.

In (5.1) and (5.2), two values were considered for ϕ : 0 (BFGS) and ϕ^* (the minimizer; see section 3).⁴ If the minimizer was computed using the new strategy, we denote it ϕ_{new}^* ; if it was computed using the Oren-Luenberger strategy, we denote it ϕ_{old}^* .

⁴Since we do not always have $\phi^* > \phi^\#$ (or $\phi^* > \gamma\phi^\#$), ϕ^* was used only if $\phi^* > 0.95\phi^\#$ (or $\phi^* > 0.95\gamma\phi^\#$).

Three values were used for γ : the Oren–Luenberger sizing factor γ_{OL} (1.7), the centered Oren–Luenberger sizing factor γ_{COL} (1.8), and γ^* (the minimizer; see section 3). Again, the minimizers will be denoted γ_{new}^* and γ_{old}^* , depending on the sizing strategy used. As suggested by Contreras and Tapia [7], to prevent us from inadvertently sizing with an excessively small constant, the actual sizing factor used was $\max\{\gamma_{\min}, \gamma\}$, where γ is the computed sizing factor, and $\gamma_{\min}$ is a small positive constant.

Our numerical experimentation concerned the following 14 sizing procedures and the standard BFGS method for comparison purposes.

METHOD	D	γ	ϕ	DESCRIPTION
0. BFGS	–	1	0	Standard BFGS
1. BFGS-OL	–	γ_{OL}	0	Oren–Luenberger sized-BFGS
2. BFGS-COL	–	γ_{COL}	0	Centered Oren–Luenberger sized-BFGS (Contreras and Tapia [7])
3. B-OL-P1-old	B	γ_{OL}	ϕ_{old}^*	Old strategy; Min. Prob. 1 using γ_{OL}
4. B-COL-P1-old	B	γ_{COL}	ϕ_{old}^*	Old strategy; Min. Prob. 1 using γ_{COL}
5. B-OL-P1-new	B	γ_{OL}	ϕ_{new}^*	New strategy; Min. Prob. 1 using γ_{OL}
6. B-COL-P1-new	B	γ_{COL}	ϕ_{new}^*	New strategy; Min. Prob. 1 using γ_{COL}
7. I-OL-P1-old	I	γ_{OL}	ϕ_{old}^*	Old strategy; Min. Prob. 1 using γ_{OL}
8. I-COL-P1-old	I	γ_{COL}	ϕ_{old}^*	Old strategy; Min. Prob. 1 using γ_{COL}
9. I-OL-P1-new	I	γ_{OL}	ϕ_{new}^*	New strategy; Min. Prob. 1 using γ_{OL}
10. I-COL-P1-new	I	γ_{COL}	ϕ_{new}^*	New strategy; Min. Prob. 1 using γ_{COL}
11. I-opt-P2-old	I	γ_{old}^*	0	Old strategy; Min. Prob. 2 using $\phi = 0$
12. I-opt-P2-new	I	γ_{new}^*	0	New strategy; Min. Prob. 2 using $\phi = 0$
13. I-opt-P3-old	I	γ_{old}^*	ϕ_{old}^*	Old strategy; Min. Prob. 3
14. I-opt-P3-new	I	γ_{new}^*	ϕ_{new}^*	New strategy; Min. Prob. 3

5.3. Preliminary results. An important observation from our initial experimentation concerns how often sizing should be performed. If used at every iteration, the sized secant strategies suggested in section 3 (whether applied using the Oren–Luenberger family (1.6) or the new family (2.3)) were inferior to the standard BFGS in all problems except the last one.⁵ Therefore, sizing should be used as a *safeguarding* of the BFGS secant method. In other words, we size or shift B before it is updated only when a need is detected. We will refer to this as Selective Sizing/Shifting. It is interesting that this is what has been concluded by numerous authors, including Carter [6], Contreras and Tapia [7], Al-Baali [1, 2], and Wolkowicz [21]. Following Contreras and Tapia [7], to decide when to size, we use the fact that both γ_{OL} and γ_{COL} should be “close” to 1 whenever B has “good” information relative to $\nabla^2 f(x)$ along the direction s . We consider the following two Selective Sizing/Shifting criteria:

1. Selective Sizing/Shifting Criterion 1

For $k = 1, 2, \dots$

If $1 - \bar{\gamma} > \epsilon$, then size or shift B .

⁵For this particular problem, suggested by Byrd, Liu, and Nocedal [3], sizing or shifting the matrix in the ways proposed above did improve the behavior of BFGS. This agrees with the results reported by Byrd, Liu, and Nocedal [3] and by Al-Baali [2], using a special member of the Broyden class different from BFGS (shifting).

2. Selective Sizing/Shifting Criterion 2

For $k = 1, 2, \dots$

If $|1 - \bar{\gamma}| > \epsilon$, then size or shift B .

We should mention that our preliminary experimentation included other choices of ϕ , such as $\phi = 1$ (DFP), and other choices of γ , such as $\gamma = \gamma_{IOL} = y^T B^{-1} y / y^T s$ (inverse Oren–Luenberger [16]). The results of these experiments showed that selectively sizing the approximation matrix B using γ^* as the sizing factor (see Minimization Problem 2) improved the behavior of DFP much more than it improved the behavior of BFGS. However, these sizing strategies never made DFP competitive with the standard BFGS. This concurs with the common belief that BFGS is a more robust update than DFP and, therefore, that DFP is more likely to be improved by a safeguarding technique, such as sizing. We believe that further tuning of these sizing procedures is needed in order to obtain the competitive behavior of DFP reported by Contreras and Tapia [7]. Since no better results than those of the standard BFGS method are expected for the sized-DFP method, we excluded it from further numerical experiments.

These results also showed that the choice of γ_{IOL} gave similar results to the choice γ_{OL} . This is not surprising since both quantities contain essentially the same information concerning B relative to $\nabla^2 f(x)$. For this reason, and because it takes $\mathcal{O}(n^2)$ more operations to compute, we excluded γ_{IOL} from further numerical experiments.

5.4. Results. Each of the 14 procedures described in section 5.2 was used to solve the five problems mentioned above. Each problem was solved with the four initial matrix approximations listed in (5.3), using both line-search and trust-region globalization techniques. We use three measures for comparisons: number of iterations, number of function evaluations, and number of function-gradient evaluations (function evaluations plus n times gradient evaluations, where n is the dimension of the problem). The results are presented as the average of each measure over the 20 outputs for each procedure. Our comparisons will be given relative to the control (standard BFGS). We accomplish this by listing all quantities normalized (divided) by the corresponding quantity for the BFGS control.

For each method and each globalization technique, the tables in this section present the average of the normalized data for each of our three measures. Tables 5.1 and 5.2 use $\bar{\gamma} = \gamma_{OL}$ in Selective Sizing/Shifting Criteria 1 and 2, respectively. Similarly, Tables 5.3 and 5.4 use $\bar{\gamma} = \gamma_{COL}$. All four tables use $\epsilon = 0.05$ and $\gamma_{\min} = 0.1$ in the Selective Sizing/Shifting Criteria, as suggested by Contreras and Tapia [7].

From these results, we can see that, in general, the following are true:

1. The improvement for the line-search globalization technique is better than it is for the trust-region globalization technique. This indicates that trust region is a more robust technique and is therefore less sensitive to changes such as selective sizing or shifting.

2. Selective Sizing/Shifting Criterion 1 (Tables 5.1 and 5.3) is much better than Criterion 2 (Tables 5.2 and 5.4). This supports Contreras and Tapia’s reason for sizing only when the sizing factor is less than one [7].

3. The number of iterations and function-gradient evaluations (which are similar) improved much more than the number of function evaluations. This provides further support for the explanation above. Both globalization techniques are implemented to take care of “large” steps (“small” B ’s or $\bar{\gamma} > 1$) at the expense of more than one function evaluation per iteration. Sizing only when $\bar{\gamma} < 1$ increases the number of “large” steps. This decreases the number of iterations, but it may increase the total number of function iterations.

TABLE 5.1
Normalized averages, ($1 - \gamma_{OL} > 0.05$, $\gamma_{\min} = 0.1$).

METHOD	LINE SEARCH			TRUST REGION		
	Iterations	Functions	Fun.-Grad.	Iterations	Functions	Fun.-Grad.
0. BFGS	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1. BFGS-OL	0.8659	0.9677	0.8662	0.9533	1.0392	0.9538
2. BFGS-COL	0.9340	0.9848	0.9337	1.0041	1.1300	1.0038
3. B-OL-P1-old	0.9139	0.9777	0.9141	0.9551	0.9971	0.9555
4. B-COL-P1-old	0.9248	0.9663	0.9250	0.9681	1.0115	0.9684
5. B-OL-P1-new	0.9138	0.9747	0.9141	0.9592	1.0048	0.9596
6. B-COL-P1-new	0.9285	0.9709	0.9287	0.9589	1.0056	0.9593
7. I-OL-P1-old	0.8497	1.0944	0.8500	0.9577	1.1834	0.9575
8. I-COL-P1-old	0.8919	1.0690	0.8922	1.0063	1.2197	1.0058
9. I-OL-P1-new	0.8601	0.9917	0.8607	0.9375	1.0820	0.9378
10. I-COL-P1-new	0.8944	0.9832	0.8949	0.9460	1.0194	0.9463
11. I-opt-P2-old	1.4030	1.6330	1.3930	1.0336	1.0635	1.0316
12. I-opt-P2-new	1.4030	1.6330	1.3930	1.0336	1.0635	1.0316
13. I-opt-P3-old	1.0270	1.1068	1.0269	0.9908	1.1426	0.9904
14. I-opt-P3-new	1.0036	1.0308	1.0038	0.9463	1.0074	0.9466

TABLE 5.2
Normalized averages, ($|1 - \gamma_{OL}| > 0.05$, $\gamma_{\min} = 0.1$).

METHOD	LINE SEARCH			TRUST REGION		
	Iterations	Functions	Fun.-Grad.	Iterations	Functions	Fun.-Grad.
0. BFGS	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1. BFGS-OL	1.2112	1.2071	1.2088	1.0961	1.1650	1.0959
2. BFGS-COL	1.1274	1.1746	1.1259	1.1440	1.2255	1.1423
3. B-OL-P1-old	1.2679	1.2184	1.2671	1.1662	1.1671	1.1653
4. B-COL-P1-old	1.1689	1.1364	1.1682	1.0831	1.0975	1.0825
5. B-OL-P1-new	1.2833	1.2407	1.2823	1.1897	1.1878	1.1886
6. B-COL-P1-new	1.1884	1.1543	1.1875	1.0821	1.0837	1.0815
7. I-OL-P1-old	1.6885	2.0295	1.6853	1.4871	1.9763	1.4831
8. I-COL-P1-old	1.1712	1.5095	1.1695	1.4075	1.8285	1.4048
9. I-OL-P1-new	1.3490	1.4927	1.3462	1.2107	1.4682	1.2073
10. I-COL-P1-new	1.1823	1.3410	1.1800	1.1997	1.4089	1.1967
11. I-opt-P2-old	1.2986	3.4839	1.2877	1.5460	2.1660	1.5425
12. I-opt-P2-new	1.2986	3.4839	1.2877	1.5460	2.1660	1.5425
13. I-opt-P3-old	0.9455	2.2305	0.9447	1.3927	2.0251	1.3907
14. I-opt-P3-new	0.9299	1.9855	0.9295	1.2172	1.4421	1.2177

4. Using the identity matrix as a weighting matrix in the computation of the optimal values γ^* or ϕ^* gives better results than using the matrix B . A plausible reason for this is that when $\bar{\gamma} \neq 1$, the matrix B does not have “good” information relative to $\nabla^2 f(x)$. Therefore, it is desirable to de-emphasize the role of B and use a step that has some “gradient” flavor (weighted matrix equal to the identity) to safeguard the BFGS method.

5. The optimal sizing factor given by Minimization Problem 2 in section 3 always failed as a safeguard for the BFGS method. However, Minimization Problem 1 and Minimization Problem 3 sometimes improved the behavior of the standard BFGS method, and this improvement was slightly more than that obtained from standard

TABLE 5.3
Normalized averages, ($1 - \gamma_{COL} > 0.05$, $\gamma_{\min} = 0.1$).

METHOD	LINE SEARCH			TRUST REGION		
	Iterations	Functions	Fun.-Grad.	Iterations	Functions	Fun.-Grad.
0. BFGS	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1. BFGS-OL	0.9336	1.0110	0.9337	0.9706	1.0501	0.9710
2. BFGS-COL	0.8686	0.9613	0.8692	0.9430	1.0633	0.9429
3. B-OL-P1-old	0.9851	1.0050	0.9853	0.9828	1.0087	0.9832
4. B-COL-P1-old	0.9029	0.9651	0.9032	0.9651	1.0101	0.9654
5. B-OL-P1-new	0.9326	0.9869	0.9328	0.9729	1.0083	0.9732
6. B-COL-P1-new	0.8989	0.9619	0.8992	0.9713	1.0138	0.9717
7. I-OL-P1-old	1.0802	1.3142	1.0796	1.0255	1.1725	1.0251
8. I-COL-P1-old	0.8791	1.1248	0.8791	0.9837	1.1691	0.9836
9. I-OL-P1-new	0.9895	1.1661	0.9887	1.0047	1.1365	1.0044
10. I-COL-P1-new	0.8524	1.0177	0.8556	0.9442	1.0900	0.9463
11. I-opt-P2-old	1.4086	1.6368	1.3984	1.0454	1.0798	1.0431
12. I-opt-P2-new	1.4086	1.6368	1.3984	1.0454	1.0798	1.0431
13. I-opt-P3-old	1.0227	1.1233	1.0222	0.9864	1.1181	0.9861
14. I-opt-P3-new	1.0044	1.0307	1.0043	0.9854	1.0342	0.9854

TABLE 5.4
Normalized averages, ($|1 - \gamma_{COL}| > 0.05$, $\gamma_{\min} = 0.1$).

METHOD	LINE SEARCH			TRUST REGION		
	Iterations	Functions	Fun.-Grad.	Iterations	Functions	Fun.-Grad.
0. BFGS	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1. BFGS-OL	1.2112	1.2071	1.2088	1.0961	1.1650	1.0959
2. BFGS-COL	1.1274	1.1746	1.1259	1.1440	1.2255	1.1423
3. B-OL-P1-old	1.2747	1.2165	1.2740	1.1430	1.1212	1.1420
4. B-COL-P1-old	1.1329	1.1056	1.1322	1.0978	1.0956	1.0970
5. B-OL-P1-new	1.1704	1.1270	1.1696	1.1465	1.1388	1.1454
6. B-COL-P1-new	1.1606	1.1315	1.1596	1.1228	1.1151	1.1220
7. I-OL-P1-old	1.9117	2.3055	1.9075	1.5109	2.0212	1.5054
8. I-COL-P1-old	1.2292	1.6407	1.2266	1.5080	1.9200	1.5027
9. I-OL-P1-new	1.2907	1.4182	1.2876	1.2032	1.3705	1.1999
10. I-COL-P1-new	1.1703	1.3073	1.1679	1.2347	1.4600	1.2307
11. I-opt-P2-old	1.3083	3.5063	1.2973	1.6134	2.3020	1.6086
12. I-opt-P2-new	1.3083	3.5063	1.2973	1.6134	2.3020	1.6086
13. I-opt-P3-old	0.9656	2.2585	0.9642	1.4929	2.2196	1.4886
14. I-opt-P3-new	0.9481	2.0051	0.9471	1.2705	1.5207	1.2690

sizing techniques. This says that to safeguard the BFGS method, it is better to size another member of the Broyden class of secant updates than to size the BFGS itself. That is, when the BFGS is not performing well, “move on” to another update instead of trying to patch up the BFGS. A similar conclusion was reached by Al-Baali [2] in numerical experimentation using his *switching methods*.

6. The best results are obtained when there is a direct correlation between $\bar{\gamma}$, the quantity used to decide when to size or shift, and γ , the sizing factor actually used. When $\bar{\gamma} = \gamma_{OL}$, Methods 7 and 9, which used $\gamma = \gamma_{OL}$ as the sizing factor, saved 15% and 6.3% over BFGS with line search and trust region, respectively (Table 5.1). Similarly, when $\bar{\gamma} = \gamma_{COL}$, the best method, I-COL-P1-new, used γ_{COL} as the sizing

TABLE 5.5
Normalized averages, $(1 - \gamma_{COL} > 0.1, \gamma_{\min} = 0.1)$.

METHOD	LINE SEARCH			TRUST REGION		
	Iterations	Functions	Fun.-Grad.	Iterations	Functions	Fun.-Grad.
0. BFGS	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1. BFGS-OL	0.9303	1.0077	0.9305	0.9834	1.0682	0.9837
2. BFGS-COL	0.8745	0.9678	0.8747	0.9776	1.0473	0.9775
3. B-OL-P1-old	0.9369	0.9917	0.9371	0.9700	1.0190	0.9703
4. B-COL-P1-old	0.9059	0.9686	0.9062	0.9577	1.0108	0.9580
5. B-OL-P1-new	0.9279	0.9891	0.9281	0.9666	0.9971	0.9669
6. B-COL-P1-new	0.9027	0.9672	0.9030	0.9646	1.0203	0.9649
7. I-OL-P1-old	0.9920	1.1871	0.9915	1.0113	1.1796	1.0109
8. I-COL-P1-old	0.8732	1.1066	0.8733	0.9670	1.1475	0.9669
9. I-OL-P1-new	0.9746	1.1410	0.9738	1.0090	1.1335	1.0087
10. I-COL-P1-new	0.8540	1.0057	0.8543	0.9456	1.0897	0.9457
11. I-opt-P2-old	1.2608	1.6543	1.2508	1.0602	1.2437	1.0577
12. I-opt-P2-new	1.2608	1.6543	1.2508	1.0602	1.2437	1.0577
13. I-opt-P3-old	0.8634	1.0813	0.8634	0.9347	1.1377	0.9349
14. I-opt-P3-new	0.8549	1.0034	0.8552	0.9605	1.0601	0.9608

factor and saved 14.5% and 5.4%, respectively (Table 5.3). Notice that the same is true not only for the choices that gave the best results.

7. Notice that the results given above are slightly better than those for Method 2, which was the choice considered by Contreras and Tapia [7] (see Tables 5.1 and 5.3).

5.5. More results. Further numerical experimentation was done to determine the sensitivity of our results to both the dimension of the problem and choices of ϵ and $\gamma_{\min}$. For the second analysis, we computed the same type of information as in the tables above, but using different values of ϵ and $\gamma_{\min}$ with Selective Sizing/Shifting Criterion 1. Tables 5.5 and 5.6 show this information for $\gamma = \gamma_{COL}$ and $(\epsilon, \gamma_{\min})$ equal to $(0.1, 0.1)$ and $(0.05, 0.0001)$, respectively. These are typical results for this sensitivity analysis.

From these data, we can see that the results obtained from the standard sizing procedures (Methods 1 and 2) became worse as ϵ increased or $\gamma_{\min}$ decreased. The results obtained from the other procedures remained mostly unchanged (with the exception of Methods 13 and 14, which improved with these changes). Of course, this holds within some ranges of ϵ and $\gamma_{\min}$. In fact, with $\epsilon > .2$, the normalized data is close to one,⁶ and it is greater than one when $\gamma_{\min}$ goes to zero.⁷

To check the sensitivity of our results to the dimension of the problems, we ran our code for problem number 21 with several values of n between 10 and 30, using the same starting points and initial matrices as Contreras and Tapia [7]. Again, we used Selective Sizing/Shifting Criterion 1 with $\bar{\gamma} = \gamma_{COL}$ and $\gamma_{\min} = 0.1$.

The issue now is reliability rather than efficiency. So Table 5.7 presents, for $n = 20$, the number of iterations required by each procedure for convergence from the different starting points and initial matrices used in trust-region sized-Broyden

⁶This means that, if ϵ is not small enough, there is not enough sizing to really safeguard (improve) the behavior of the standard BFGS method.

⁷This shows the importance of preventing the sizing strategies from using sizing factors that are too small.

TABLE 5.6
Normalized averages, $(1 - \gamma_{COL} > 0.05, \gamma_{\min} = 0.001)$.

METHOD	LINE SEARCH			TRUST REGION		
	Iterations	Functions	Fun.-Grad.	Iterations	Functions	Fun.-Grad.
0. BFGS	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1. BFGS-OL	0.9856	1.0245	0.9857	1.0100	1.0670	1.0103
2. BFGS-COL	0.8714	0.9584	0.8716	0.9728	1.0542	0.9727
3. B-OL-P1-old	0.9387	0.9859	0.9388	0.9731	1.0167	0.9734
4. B-COL-P1-old	0.9067	0.9671	0.9070	0.9651	1.0101	0.9654
5. B-OL-P1-new	0.9380	0.9920	0.9382	0.9729	1.0083	0.9732
6. B-COL-P1-new	0.9075	0.9659	0.9078	0.9713	1.0138	0.9717
7. I-OL-P1-old	1.0337	1.2743	1.0331	1.0255	1.1725	1.0251
8. I-COL-P1-old	0.8849	1.1249	0.8848	0.9837	1.1691	0.9836
9. I-OL-P1-new	1.0352	1.2263	1.0340	1.0047	1.1365	1.0044
10. I-COL-P1-new	0.8721	1.0278	0.8722	0.9462	1.0900	0.9463
11. I-opt-P2-old	1.2284	1.9570	1.2184	1.1360	1.3849	1.1332
12. I-opt-P2-new	1.2284	1.9570	1.2184	1.1360	1.3849	1.1332
13. I-opt-P3-old	0.8648	1.3652	0.8647	0.9857	1.1960	0.9857
14. I-opt-P3-new	0.8523	1.2269	0.8525	0.9960	1.1475	0.9961

methods. The starting points are

$$\begin{aligned}
x_0^1 &= (-1.2, 1, \dots, -1.2, 1), \\
x_0^2 &= (-4000, 1, -1.2, 1, \dots, -4000, 1, -1.2, 1), \\
x_0^3 &= (-100.5, 40, \dots, -100.5, 40), \\
x_0^4 &= (-40, 20, 20, 20, \dots, -40, 20, 20, 20), \\
x_0^5 &= (-300, -100, \dots, -300, -100),
\end{aligned}$$

and the initial matrices are

$$\begin{aligned}
B_0^0 &= I, \\
B_0^1 &= I + (10^{12} - 1) * \text{diag} \left(1 - \frac{i-1}{n-1} \right), \\
B_0^2 &= \text{diag}(10^7, 10^{-7}, \dots, 10^7, 10^{-7}), \\
B_0^3 &= \text{diag}(10^5, 10^{-5}, \dots, 10^5, 10^{-5}).
\end{aligned}$$

From these results, we can see that an appropriate sizing procedure is needed to improve the reliability of the standard BFGS method, as was also concluded by Contreras and Tapia [7]. Even though none of the 14 procedures tested here were 100% reliable, the most reliable procedures were those derived from Minimization Problem 1. This fact further supports observation 5 above, which says that when BFGS is “bad” it is better to size a different update instead of attempting to patch up the BFGS update.

Observe also that, ignoring the cases for which all procedures failed, the most reliable procedures were “close” or “very close” to the solution in the cases where they failed, and the standard BFGS failed completely for these cases. It should be noted that this was not true for Method 2, the Contreras–Tapia choice (see case x_0^4 with B_0^3).

TABLE 5.7
Number of iterations (problem 21 with $n = 20$, $1 - \gamma_{COL} > 0.05$, $\gamma_{\min} = 0.1$).

INITIAL VALUES		M E T H O D S															
		BFGS				Old Strategy						New Strategy					
						Min. Prob. 1				Min P2	Min P3	Min. Prob. 1				Min P2	Min P3
						using D=B	using D=I	using D=B	using D=I	using D=B	using D=I						
x_0	B_0	Stand	OL	COL	OL	COL	OL	COL	D=I	D=I	OL	COL	OL	COL	D=I	D=I	
x_0^1	B_0^1	138	147	140	149	142	144	139	142	138	152	147	146	137	142	169	
	B_0^2	F	197	149	167	146	G**	152	G**	143	157	143	197	147	G**	G**	
	B_0^3	183	375	360	286	384	206	292	F*	356	219	394	196	313	F*	348	
	B_0^4	206	279	382	294	345	219	293	407	350	301	334	212	297	407	346	
x_0^2	B_0^1	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	
	B_0^2	F	284	253	204	221	425	298	F	F	213	218	284	294	F	F	
	B_0^3	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	
	B_0^4	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	
x_0^3	B_0^1	F	472	487	F**	464	498	413	F	F*	F**	486	472	487	F	F**	
	B_0^2	F	372	393	417	344	446	373	280	G**	392	342	442	364	280	295	
	B_0^3	F	F	F	F*	F*	F	F**	F	F	F	F*	F	F*	F	F	
	B_0^4	F	F	F*	463	447	F*	465	489	F	472	423	498	484	489	F	
x_0^4	B_0^1	334	258	263	267	259	264	263	247	238	251	255	257	261	247	250	
	B_0^2	F	367	291	319	268	367	310	221	216	319	255	367	291	221	221	
	B_0^3	416	F	F	359	426	335	446	F*	F*	391	443	389	430	F*	F	
	B_0^4	372	453	440	314	317	332	328	339	338	299	339	325	317	339	333	
x_0^5	B_0^1	F*	430	390	397	365	428	390	330	348	382	363	430	390	330	369	
	B_0^2	F	414	274	303	245	231	343	F	F	331	245	237	346	F	F	
	B_0^3	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	
	B_0^4	F*	F	F	413	386	381	377	375	385	407	390	393	372	375	384	

F: Maximum number of iterations (500) reached.

G: Global strategy failed to compute a “good” step.

*: Last iterate was “close” to the solution ($error < 10^{-2}$).

** : Last iterate was “very close” to the solution ($error \leq 10^{-5}$).

6. Summary and concluding remarks. In this study, we have presented a considerable amount of theoretical discussion concerning sizing members of the Broyden family of secant updates, followed by results of preliminary testing on sample problems. The study has led us to the following conclusions. The BFGS secant update should continue to be the secant method of choice for unconstrained optimization. However, the BFGS update is occasionally faulty. This occurrence can be detected using either the Oren–Luenberger sizing factor (1.7) or the centered Oren–Luenberger sizing factor (1.8). When such a situation is detected, the approximate Hessian should be sized with the sizing factor used to detect the anomaly. These findings mirror those of Contreras and Tapia [7]. However, we have discovered more.

The BFGS update is prone to producing updates with large eigenvalues. Indeed, in some ways it was designed to do this, since it was presented as a fix for the DFP method, which produced excessively small eigenvalues. Hence, sizing the BFGS update may lead to an excessively small sizing factor and near singularity of the sized matrix. Contreras and Tapia were aware of this deficiency and suggested a safeguard for their sizing strategy. Another way to compensate for this negative feature is to shift, i.e., add a matrix to the sized matrix. The effective implementation of this shift has been the primary subject of the present study. We first considered a two-parameter family consisting of sized-Broyden family secant updates. It was instructive for us to view this two-parameter family as a BFGS update of a matrix $\gamma B + \phi vv^T$

that had been sized and then shifted. The shift is of the form ϕvv^T and corresponds to switching from the BFGS update ($\phi = 0$) to an alternative member of the Broyden family of updates ($\phi \neq 0$). Considerable theory and validating numerical experimentation was performed in an effort to find good choices for the sizing parameter γ and the shifting parameter ϕ . After much study, we concluded (as mentioned above) that the sizing factor should be either the Oren–Luenberger or the centered Oren–Luenberger sizing factor and that the sizing should be done selectively, as suggested by Contreras and Tapia [7]. Our numerical experiments indicated that an effective way of determining the shift factor ϕ was to minimize the Byrd–Nocedal ψ -measure over our sized-Broyden family with the sizing factor fixed. We considered weighting the Byrd–Nocedal ψ -measure by a matrix D and studied the following three choices for D : $D = I$, the identity; $D = B$, the approximate Hessian; and $D = \nabla^2 f(x)$, the exact Hessian. The choice $D = I$ gives the member of the sized-Broyden family that is nearest the method of steepest descent (in the Byrd–Nocedal measure), the choice $D = B$ gives the least change member, and the choice $D = \nabla^2 f(x)$ gives the member that is closest to Newton’s method. Therefore, we were rather surprised our numerically most effective choice was $D = I$. In retrospect, we see that there is considerable consistency in this choice. Our sizing indicator has told us that we should size; hence there is a bad match of information in the pair B and $\nabla^2 f(x)$. The choice $D = I$ prevents this faulty information from further contaminating our update.

Acknowledgments. The authors would like to thank the referees for their constructive comments and Teresa Parks for her contributions to the revision of the manuscript.

REFERENCES

- [1] M. AL-BAALI, *Partial Self Scaling Variable Metric Algorithms*, Technical report N80, Dipartimento di Sistemi, Università della Calabria, Calabria, Italy, 1988.
- [2] M. AL-BAALI, *An Efficient Class of Quasi-Newton Algorithms in the Broyden Family*, Technical report N9, Dipartimento di Elettronica Informatica e Sistemistica, Università della Calabria, Calabria, Italy, 1992.
- [3] R. H. BYRD, D. C. LIU, AND J. NOCEDAL, *On the behavior of Broyden’s class of quasi-Newton methods*, SIAM J. Optim., 2 (1992), pp. 533–557.
- [4] R. H. BYRD AND J. NOCEDAL, *A tool for the analysis of quasi-Newton methods with application to unconstrained minimization*, SIAM J. Numer. Anal., 26 (1989), pp. 727–739.
- [5] R. H. BYRD, J. NOCEDAL, AND Y. X. YUAN, *Global convergence of a class of quasi-Newton methods on convex problems*, SIAM J. Numer. Anal., 24 (1987), pp. 1171–1190.
- [6] R. G. CARTER, *Safeguarding Hessian Approximations in Trust Region Algorithms*, Technical report TR87-12, Department of Mathematical Sciences, Rice University, Houston, TX, 1987.
- [7] M. CONTRERAS AND R. A. TAPIA, *Sizing the BFGS and DFP updates: Numerical study*, J. Optim. Theory Appl., 78 (1993), pp. 93–108.
- [8] W. C. DAVIDON, *Optimally conditioned optimization algorithms without line searches*, Math. Programming, 9 (1975), pp. 1–30.
- [9] J. E. DENNIS, JR., D. M. GAY, AND R. E. WELSCH, *An adaptive nonlinear least-squares algorithm*, ACM Trans. Math. Software, 7 (1981), pp. 348–368.
- [10] J. E. DENNIS, JR., AND R. B. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Classics Appl. Math. 16, SIAM, Philadelphia, 1996.
- [11] J. E. DENNIS, JR., AND H. WOLKOWICZ, *Sizing and least-change secant methods*, SIAM J. Numer. Anal., 30 (1993), pp. 1291–1314.
- [12] R. FLETCHER, *A new variational result for quasi-Newton formulae*, SIAM J. Optim., 1 (1991), pp. 18–21.
- [13] Y. F. HU AND C. STOREY, *Family of optimally conditioned quasi-Newton updates for unconstrained optimization*, J. Optim. Theory Appl., 83 (1994), pp. 421–431.

- [14] J. J. MORÉ, B. S. GARBOW, AND K. E. HILLSTROM, *Testing unconstrained optimization software*, ACM Trans. Math. Software, 7 (1981), pp. 17–41.
- [15] S. S. OREN, *Self-scaling variable metric (SSVM) algorithms*, II. *Implementation and experiments*, Management Sci., 20 (1973/74), pp. 863–874.
- [16] S. S. OREN AND D. G. LUENBERGER, *Self-scaling variable metric (SSVM) algorithms*, I. *Criteria and sufficient conditions for scaling a class of algorithms*, Management Sci., 20 (1973/74), pp. 845–862.
- [17] S. S. OREN AND E. SPEDICATO, *Optimal conditioning of self-scaling variable metric algorithms*, Math. Programming, 10 (1976), pp. 70–90.
- [18] D. F. SHANNO AND K. H. PHUA, *Matrix conditioning and nonlinear optimization*, Math. Programming, 14 (1978), pp. 149–160.
- [19] D. SIEGEL, *Modifying the BFGS update by a new column scaling technique*, Math. Programming, 66 (1994), pp. 45–78.
- [20] M. J. TODD, *Another Variational Derivation of a Self-scaling Quasi-Newton Update Formula*, Technical report 964, School of Operations Research and Industrial Engineering, College of Engineering, Cornell University, Ithaca, NY, 1991.
- [21] H. WOLKOWICZ, *Measures for symmetric rank-one updates*, Math. Oper. Res., 19 (1994), pp. 815–830.
- [22] H. WOLKOWICZ AND Q. ZHAO, *An all-inclusive efficient region of updates for least change secant methods*, SIAM J. Optim., 5 (1995), pp. 172–191.
- [23] Y. ZHANG AND R. P. TEWARSON, *Quasi-Newton algorithms with updates from the preconvex part of Broyden's family*, IMA J. Numer. Anal., 8 (1988), pp. 487–509.

PARTIALLY AUGMENTED LAGRANGIAN METHOD FOR MATRIX INEQUALITY CONSTRAINTS*

DOMINIKUS NOLL[†], MOUNIR TORKI[‡], AND PIERRE APKARIAN[§]

Abstract. We discuss a partially augmented Lagrangian method for optimization programs with matrix inequality constraints. A global convergence result is obtained. Applications to hard problems in feedback control are presented to validate the method numerically.

Key words. augmented Lagrangian, linear matrix inequalities, bilinear matrix inequalities, semidefinite programming

AMS subject classifications. 49N35, 90C22, 93B51

DOI. 10.1137/S1052623402413963

1. Introduction. The augmented Lagrangian method was proposed independently by Hestenes [34] and Powell [47] in 1969 and, since its inauguration, continues to be an important option in numerical optimization. With the introduction of successive quadratic programming (SQP) in the 1970s and the rise of interior point methods in the 1980s, the interest in the augmented Lagrangian somewhat declined but never completely ceased. For instance, in the 1980s, some authors proposed to combine SQP with augmented Lagrangian merit functions, and today the idea of the augmented Lagrangian is revived in the context of interior point methods, where it is one possible way to deal with nonlinear equality constraints. A history of the augmented Lagrangian from its beginning to the early 1990s is presented in [20].

Here we are concerned with a *partially augmented Lagrangian* method, a natural variation of the original theme. *Partial* refers to when some of the constraints are not included in the augmentation process but kept explicitly in order to exploit their structure. Surprisingly enough, this natural idea appears to have been overlooked before 1990. In a series of papers [20, 21, 22, 23] starting in the early 1990s, Conn et al. finally examined this approach, and a rather comprehensive convergence analysis for traditional nonlinear programming problems has been obtained in [23, 49].

In the present work we discuss optimization programs featuring matrix inequality constraints in addition to the traditional equality and inequality constraints. Such programs arise quite naturally in feedback control and have a large number of interesting applications. We propose a partially augmented Lagrangian strategy as one possible way to deal with these programs.

Semidefinite programming (SDP) is the most prominent example of a matrix inequality constrained program. With its link to integer programming [32] and because of a large number of applications in control [12], SDP has become one of the most

*Received by the editors September 3, 2002; accepted for publication (in revised form) March 8, 2004; published electronically October 14, 2004.

<http://www.siam.org/journals/siopt/15-1/41396.html>

[†]Université Paul Sabatier, Mathématiques pour l'Industrie et la Physique, CNRS UMR 5640, 118, route de Narbonne, 31062 Toulouse, France (noll@mip.ups-tlse.fr).

[‡]Université d'Avignon, Laboratoire d'Analyse non linéaire et Géométrie, Institut Universitaire Professionnalisé, 339 chm. des Meinajariés, Agroparc BP 1228, 84911 Avignon, France (mounir.torki@iup.univ-avignon.fr).

[§]ONERA-CERT, Centre d'études et de recherche de Toulouse, Control System Department, 2 av. Edouard Belin, 31055 Toulouse, France, and Université Paul Sabatier, Mathématiques pour l'Industrie et la Physique, CNRS UMR 5640, 118, route de Narbonne, 31062 Toulouse, France (apkarian@cert.fr).

active research topics in nonlinear optimization. During the 1990s, problems like H_2 - or H_∞ -synthesis, linear parameter varying (LPV) synthesis, robustness analysis, and analysis under integral quadratic constraints (IQCs), among others, have been identified as linear matrix inequality (LMI) feasibility or optimization problems, solvable therefore by SDP [2, 35, 50, 40, 30, 12].

It needs to be stressed, however, that the most important problems in feedback control *cannot* be solved by SDP. Challenging problems like parametric robust H_2 - or H_∞ -output feedback synthesis, reduced or fixed-order output feedback design, static output feedback control, multimodel design or synthesis under IQC-constraints, synthesis with parameter-dependent Lyapunov functions, robust controller design with generalized Popov multipliers, and stabilization of delayed systems are all known to be NP-hard problems, which are beyond convexity methods, and the list could be extended.

Most of these hard problems in control have been deemed largely inaccessible only a couple of years ago [6, 18, 48, 42]. In response to this challenge, we have proposed three different strategies beyond SDP which address these hard problems [28, 29, 4, 5], and one of the most promising approaches is the partially augmented Lagrangian discussed here. In this work we will mainly consider convergence issues, but several numerical test examples in reduced order H_∞ -synthesis and in robust H_∞ -control synthesis are included in order to validate the approach numerically. We mention related work on reduced order synthesis by Leibfritz and Mostafa [37, 38], and a very different algorithmic approach by Burke, Lewis, and Overton [15, 16] based on nonsmooth analysis techniques. The appealing aspect of their strategy is that it seems better adapted to large-size problems.

A general feature of the mentioned hard problems in feedback control is the fact that they may all be cast as minimizing a convex or even linear objective function subject to bilinear matrix inequality (BMI) constraints:

$$\begin{aligned} & \text{minimize} && c^T x, \quad x \in \mathbb{R}^n, \\ (B) \quad & \text{subject to} && A_0 + \sum_{i=1}^n x_i A_i + \sum_{1 \leq i < j \leq n} x_i x_j B_{ij} \preceq 0, \end{aligned}$$

where $\mathcal{B}(x) := A_0 + \sum_{i=1}^n x_i A_i + \sum_{1 \leq i < j \leq n} x_i x_j B_{ij}$ is a bilinear matrix function with values in a space $\mathbb{S}_p$ of symmetric $p \times p$ matrices, and where $\preceq 0$ means negative semidefinite. Such a program may be transformed to minimizing a linear objective subject to LMI-constraints in tandem with nonlinear equality constraints:

$$\begin{aligned} & \text{minimize} && c^T x, \quad x \in \mathbb{R}^n, \\ (S) \quad & \text{subject to} && g_j(x) = 0, \quad j = 1, \dots, m, \\ & && A_0 + \sum_{i=1}^n x_i A_i \preceq 0, \end{aligned}$$

where $\mathcal{A}(x) := A_0 + \sum_{i=1}^n x_i A_i$ is now an affine matrix function, and where the nonconvexity in (B) has been shifted to the equality constraints $g(x) = 0$. Notice, however, that the way in which the cast (S) is obtained from (B) is usually critical for a successful numerical solution.

Once a suitable form (S) has been found, the following partially augmented Lagrangian strategy seems near at hand. Augmenting the nonlinear equality constraints, $g(x) = 0$, and keeping the LMI-constraints, $\mathcal{A}(x) \preceq 0$, we expect to solve the difficult problem (S) through a succession of easier SDPs. This is convenient from a practical

point of view, as existing software for SDP may be exploited. We have successfully applied this strategy in [28, 5, 4, 29] to problems in robust control and static output feedback control. Here we shall corroborate our experience by presenting two applications in robust control classified as difficult.

The paper is structured as follows. Section 2 presents the partially augmented Lagrangian method, sections 3–5 provide convergence results, followed by a discussion in section 6. Applications are presented in section 7.

Our presentation is inspired by the work of Conn et al. [20, 21, 22, 23], even though their techniques strongly rely on the polyhedral nature of the constraint set C . It will become clear at which points new ideas are required to account for the more complicated boundary structure of LMI or BMI constrained sets.

2. Problem setting. Our convergence analysis applies to more general situations than program (S). We consider the program

$$(P) \quad \begin{array}{ll} \text{minimize} & f(x), x \in \mathbb{R}^n, \\ \text{subject to} & g_j(x) = 0, j = 1, \dots, m, \\ & x \in C, \end{array}$$

where C is a closed convex set, and the data f, g_j are of class $\mathcal{C}^2$. For later use, let $g(x) = [g_1(x), \dots, g_m(x)]^T$, and let $J(x)$ be the $m \times n$ Jacobian of $g(x)$, that is,

$$J(x)^T = [\nabla g_1(x), \dots, \nabla g_m(x)].$$

For a given penalty parameter $\mu > 0$ and a Lagrange multiplier estimate $\lambda \in \mathbb{R}^m$, we define the *partially augmented Lagrangian* of program (P) as

$$(1) \quad \Phi(x; \lambda, \mu) = f(x) + \sum_{j=1}^m \lambda_j g_j(x) + \frac{1}{2\mu} \sum_{j=1}^m g_j(x)^2.$$

Following the classical idea of the augmented Lagrangian method, we replace program (P) by the following approximation:

$$(P_{\lambda, \mu}) \quad \begin{array}{ll} \text{minimize} & \Phi(x; \lambda, \mu), x \in \mathbb{R}^n, \\ \text{subject to} & x \in C. \end{array}$$

The rationale of this choice is that $(P_{\lambda, \mu})$ is easier to solve than (P) and that, with appropriate λ and μ , a solution to $(P_{\lambda, \mu})$ will be close to a solution of the original program (P). The first condition is in particular met for polyhedral sets $C = \{x \in \mathbb{R}^n : A^{(1)}x \leq b^{(1)}, A^{(2)}x = b^{(2)}\}$. The application we have in mind is when C is an LMI-constrained set,

$$(2) \quad C = \{x \in \mathbb{R}^n : \mathcal{A}(x) \preceq 0\}, \quad \text{where } \mathcal{A}(x) = A_0 + \sum_{i=1}^n x_i A_i$$

with $A_i \in \mathbb{S}_p$, the space of $p \times p$ symmetric matrices, and where $\preceq 0$ means negative semidefinite.

For later use, let us fix some notation. Along with the augmented Lagrangian (1), we also consider the traditional Lagrangian of program (P), defined as

$$(3) \quad L(x; \lambda) = f(x) + \lambda^T g(x).$$

The first order Lagrange multiplier update rule is now defined as

$$(4) \quad \lambda^\sharp(x; \lambda, \mu) := \lambda + \frac{g(x)}{\mu}$$

and the useful relation

$$(5) \quad \nabla \Phi(x; \lambda, \mu) = \nabla L(x; \lambda^\sharp(x; \lambda, \mu))$$

is satisfied. Here and later on, whenever a gradient symbol as in $\nabla \Phi(x; \lambda, \mu)$ or $\nabla L(x; \lambda)$ occurs, it is applied to the variable x . For simplicity we will exclusively use the Euclidean norms on the spaces $\mathbb{R}^n$ and $\mathbb{S}_p$. The scalar product on $\mathbb{S}_p$ is $X \bullet Y = \text{tr}(XY)$.

We need some more notation from nonsmooth analysis. For $x \in C$ let $N(C, x)$ be the normal cone, and let $T(C, x)$ be the tangent cone of C at x . When there is no ambiguity as to the meaning of the set C , the orthogonal projection $P_{T(C, x)}$ onto the tangent cone $T(C, x)$ will be written as $P(x)$. Let $V(C, x)$ denote the largest linear subspace or *lineality space* of $T(C, x)$, that is, $V(C, x) = T(C, x) \cap -T(C, x)$ (see [19]). Again, if the meaning of the set C is clear, the orthogonal projection onto $V(C, x)$ will be denoted by $\Pi(x)$.

During our analysis we assume that the nonsmoothness in program (P) has been shifted to the set C , whereas the other program data are smooth. Throughout we make the following hypothesis.

($\mathcal{H}_1$) The functions f, g_j are of class $\mathcal{C}^1$.

The method we investigate is now the following, built along the models by Conn et al. [20, 21, 22, 23].

PARTIALLY AUGMENTED LAGRANGIAN

1. (Preparation) Fix an initial Lagrange multiplier estimate λ and an initial penalty $\mu > 0$, ($\mu < 1$), and let x be an initial guess of the solution. Fix the tolerance parameters $\omega, \eta > 0$ at the initial values $\omega_0, \eta_0 > 0$, and choose final tolerance values $\omega_* \ll \omega_0$, $\eta_* \ll \eta_0$. Let $0 < \tau < 1$ and $\alpha > 0$, $\beta > 0$. Let **success** be a boolean variable with the values **yes** and **no**, and initialize **success** = **no**.
2. (Stopping test) Stop the algorithm if **success** == **yes** and if ω and η are sufficiently small, that is, $\omega \leq \omega_*$ and $\eta \leq \eta_*$.
3. (Optimization step) Given the current λ and $\mu > 0$, $\omega > 0$, approximately solve program $(P_{\lambda, \mu})$, possibly using x as a starting value. Stop the optimization $(P_{\lambda, \mu})$ as soon as an iterate x^+ close to the true solution of $(P_{\lambda, \mu})$ has been found: Stop if the solution d^+ of

$$(6) \quad \inf \{ \| -\nabla \Phi(x^+; \lambda, \mu) - d \| : d \in T(C, x^+) \}$$

satisfies $\|d^+\| \leq \omega$.

4. (Decision step) If $\|g(x^+)\| \leq \eta$, put **success** = **yes** and do a multiplier update step:

$$\begin{aligned} \mu^+ &= \mu, \\ \lambda^+ &= \lambda + g(x^+)/\mu^+, \\ \omega^+ &= \omega\mu^\beta, \\ \eta^+ &= \eta\mu^\beta, \end{aligned}$$

else put **success** = **no** and do a constraint reduction step:

$$\begin{aligned}\lambda^+ &= \lambda, \\ \mu^+ &= \tau\mu, \\ \omega^+ &= \omega_0(\mu^+)^\alpha, \\ \eta^+ &= \eta_0(\mu^+)^\alpha.\end{aligned}$$

5. Go back to step 2.

The mechanism is the following. Having solved the approximate program $(P_{\lambda,\mu})$ within the allowed tolerance ω , we check whether the approximate solution x^+ in step 3 satisfies the constraints $g(x^+) = 0$ within the currently acceptable tolerance level η . If this is the case, we consider this step as successful and proceed to a new instance of $(P_{\lambda,\mu})$ with λ updated according to the first order multiplier update rule (4). On the other hand, if $g(x) = 0$ is significantly violated, the solution of $(P_{\lambda,\mu})$ is considered unsuccessful. Here we reduce μ and perform $(P_{\lambda,\mu})$ again with λ unchanged. The choice of the term *successful* versus *unsuccessful* is understood from the perspective that we want to update λ according to the first order rule in order to drive it toward an optimal Lagrange multiplier λ_* . Our convergence theorems, Theorems 4.4 and 5.1, will clarify in which sense the first order updates $\lambda^\#$ may be expected to converge.

3. Multiplier estimates. Let us suppose that $x_* \in C$ is a Karush–Kuhn–Tucker (KKT) point of program (P) in the sense that $g(x_*) = 0$ and there exist a Lagrange multiplier $\lambda_* \in \mathbb{R}^m$ and an exterior normal vector $y_* \in N(C, x_*)$ such that

$$(7) \quad \nabla f(x_*) + J(x_*)^T \lambda_* + y_* = 0.$$

Let us further assume that the linear subspace $V(C, x_*)$ has dimension $r \geq 1$, and let $\Pi_* : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be the orthogonal projection onto that subspace. Then Π_* may be decomposed as $\Pi_* = Z_* Z_*^T$, where the columns of the $n \times r$ matrix Z_* form an orthonormal basis of $V(C, x_*)$. Notice that $Z_*^T Z_* = I_r$. Since $y_* \in N(C, x_*)$, we have $\Pi_* y_* = 0$, and by the orthogonality of Z_* this gives $Z_*^T y_* = 0$. Hence from (7) we derive

$$Z_*^T \nabla f(x_*) + Z_*^T J(x_*)^T \lambda_* = 0,$$

which gives rise to the relation

$$(8) \quad \lambda_* = -(J(x_*) Z_* Z_*^T J(x_*)^T)^{-1} J(x_*) Z_* Z_*^T \nabla f(x_*),$$

valid as soon as $J(x_*) Z_*$ has column rank $\geq m$. This suggests that for vectors x in a neighborhood of x_* , where $J(x) Z_*$ also has column rank $\geq m$, a natural Lagrange multiplier estimate would be

$$(9) \quad \lambda_*(x) := -(J(x) Z_* Z_*^T J(x)^T)^{-1} J(x) Z_* Z_*^T \nabla f(x).$$

This estimate is indeed used by Conn et al. as the main analytical tool to analyze convergence of the partially augmented Lagrangian method for polyhedral sets C . In the case of LMI-constrained sets C , it encounters problems related to the more complicated boundary structure, and a better suited construction will be elaborated on below. First, let us observe that the following hypothesis was needed to introduce $\lambda_*(x)$.

$(\mathcal{H}_2)$ $J(x_*) Z_*$ has column rank $\geq m$.

During the following, assume that $(\mathcal{H}_1)$, $(\mathcal{H}_2)$ are satisfied.

Remark. Notice that $(\mathcal{H}_2)$ is a constraint qualification hypothesis. To see this consider the case where C is described by a finite set of inequality constraints, $h_1(x) \leq 0, \dots, h_s(x) \leq 0$, each of which is active at x . Moreover, assume that the active gradients at x , $\nabla h_1(x), \dots, \nabla h_s(x)$, are linearly independent.

LEMMA 3.1. *Under these circumstances, the validity of $(\mathcal{H}_2)$ at x is equivalent to linear independence of the set $\nabla h_1(x), \dots, \nabla h_s(x), \nabla g_1(x), \dots, \nabla g_m(x)$ of equality and active inequality constraint gradients at x .*

Proof. (1) To see that $(\mathcal{H}_2)$ implies linear independence of the active gradients, suppose these vectors were linearly dependent. Then $p := \sum_{i=1}^m \mu_i \nabla g_i(x) = \sum_{j=1}^s \nu_j \nabla h_j(x)$, where $\mu \neq 0$. Hence $v^T p = 0$ for every $v \in V(C, x)$, these v being orthogonal to all $\nabla h_j(x)$. However, then the image of $V(C, x)$ under the operator $J(x)$ could no longer have dimension m , as required by $(\mathcal{H}_2)$, because $J(x)p = 0$ with $p \neq 0$.

(2) Conversely, suppose $\nabla h_1(x), \dots, \nabla h_s(x), \nabla g_1(x), \dots, \nabla g_m(x)$ are linearly independent. Suppose $J(x)Z(x)$ is not of rank m ; then it is not surjective, so $Z(x)^T J(x)^T$ is not injective. Therefore there exists $\lambda \neq 0$ such that $Z(x)^T J(x)^T \lambda = 0$, but linear independence of $\nabla g_1(x), \dots, \nabla g_m(x)$ means $J(x)^T$ is injective, so $\mu = J(x)^T \lambda \neq 0$ with $Z(x)^T \mu = 0$. This means $\mu \perp V(C, x)$, so μ is a linear combination of the $\nabla h_i(x)$, a consequence of the special boundary structure of C at x . However, $\mu - J(x)^T \lambda = 0$, so $\nabla h_i(x), \nabla g_j(x)$ are linearly dependent, which is a contradiction. $\square$

Let us now resume our line of investigation and see in which way trouble with (9) could be avoided for a reasonable rich class of sets C . Suppose that for every x in a neighborhood $U(x_*)$ of x_* there exists a linear subspace $L(C, x)$ of $V(C, x)$ which depends smoothly on x and coincides with $V(C, x_*)$ at x_* . This means that $\dim L(C, x) = r$, and that the orthogonal projector $\tilde{\Pi}(x)$ onto $L(C, x)$ varies smoothly with x . We may represent $\tilde{\Pi}(x) = \tilde{Z}(x)\tilde{Z}(x)^T$, with an orthonormal $n \times r$ matrix $\tilde{Z}(x)$ varying also smoothly with x . Then we define

$$(10) \quad \tilde{\lambda}(x) := -(J(x)\tilde{Z}(x)\tilde{Z}(x)^T J(x)^T)^{-1} J(x)\tilde{Z}(x)\tilde{Z}(x)^T \nabla f(x),$$

which is now Lipschitz in a neighborhood of x_* . Moreover, $\tilde{\lambda}(x_*) = \lambda_*$. We observe that as a consequence of $(\mathcal{H}_2)$, the matrix $J(x)\tilde{Z}(x)\tilde{Z}(x)^T J(x)^T$ is invertible in a neighborhood $U(x_*)$ of x_* .

DEFINITION 1. *A closed convex set C is said to admit a stratification into differentiable layers at $x \in \partial C$ if for $x' \in C$ in a neighborhood of x there exists a linear subspace $L(C, x')$ of the tangent cone $T(C, x')$ varying smoothly with x' such that at $x' = x$, $L(C, x)$ coincides with the lineality space $V(C, x)$ of the tangent cone at x .*

Example 1. Let $C = \mathbb{S}_p^-$, the negative semidefinite cone. Let A be in the boundary of $\mathbb{S}_p^-$; then $V(\mathbb{S}_p^-, A) = \{Z \in \mathbb{S}_p : Y_1^T Z Y_1 = 0\}$, where the columns of the $p \times r$ matrix Y_1 form an orthonormal basis of the eigenspace of the leading eigenvalue $\lambda_1(A) = 0$ of A , whose multiplicity is r .

For a perturbation E of A , there exists a matrix $Y_1(A + E)$ whose columns form an orthonormal basis of the invariant subspace associated with the first r eigenvalues of $A + E$. Then (cf. [53])

$$Y_1(A + E) = Y_1 + (\lambda_1(A)I_p - A)^\dagger E Y_1 + o(\|E\|),$$

where $M^\dagger$ denotes the pseudoinverse of M . Then we define the subspace $L(C, A + E)$ as $L(C, A + E) = \{Z \in \mathbb{S}_p : Y_1(A + E)^T Z Y_1(A + E) = 0\}$. This means that the

semidefinite order cone $\mathbb{S}_p^-$ has a differentiable stratification in the sense of Definition 1. In this example the layers or strata are the sets $\mathcal{S}_r = \{A \in \mathbb{S}_p^- : \lambda_1(A) \text{ has multiplicity } r\}$.

Example 2. Now let C be an LMI-constrained set given by (2). Since C is the preimage of $\mathbb{S}_p^-$ under an affine operator $\mathcal{A} : \mathbb{R}^n \rightarrow \mathbb{S}_p$, the elements $V(C, x)$ and $L(C, x)$ are just the preimages of $V(\mathbb{S}_p^-, \mathcal{A}(x))$ and $L(\mathbb{S}_p^-, \mathcal{A}(x))$ under the linear part $\mathcal{A}_*$ of $\mathcal{A}$. Therefore, LMI-sets satisfy the condition in Definition 1.

4. Convergence. Consider a sequence of iterates x_k generated by our algorithm. Let λ_k be the corresponding multiplier estimates, μ_k be the penalty parameters, and ω_k, η_k be the tolerance parameters. Suppose $\omega_k \rightarrow 0$. Suppose x_* is an accumulation point of the sequence x_k , and select a subsequence $\mathcal{K} \subset \mathbb{N}$ such that $x_k, k \in \mathcal{K}$, converges to x_* . Suppose hypotheses $(\mathcal{H}_1)$, $(\mathcal{H}_2)$ are met at x_* . Moreover, suppose $x_* \in C$ admits a stratification into differentiable layers as in Definition 1.

LEMMA 4.1. *Suppose the x_k satisfy the stopping test (6) in step 3 of the algorithm. Then*

$$(11) \quad \|\tilde{Z}(x_k)^T \nabla \Phi(x_k; \lambda_k, \mu_k)\| \leq \omega_k.$$

Proof. Since $\tilde{\Pi}(x_k) = \tilde{Z}(x_k)\tilde{Z}(x_k)^T$ is the projection onto $L(C, x_k)$, and since $L(C, x) \subset T(C, x)$, we have $\|\tilde{\Pi}(x_k)\nabla \Phi(x_k; \lambda_k, \mu_k)\| \leq \|P(x_k)(-\nabla \Phi(x_k; \lambda_k, \mu_k))\|$, where $P(x)$ is the orthogonal projector onto the tangent cone $T(C, x)$ at x . However, now the stopping test (6) gives $\|P(x_k)(-\nabla \Phi(x_k; \lambda_k, \mu_k))\| \leq \omega_k$.

To conclude, observe that $\|\tilde{Z}(x_k)^T \nabla \Phi(x_k; \lambda_k, \mu_k)\| = \|\tilde{\Pi}(x_k)\nabla \Phi(x_k; \lambda_k, \mu_k)\|$, since $\tilde{Z}(x)$ is orthogonal. $\square$

LEMMA 4.2. *Under the same assumptions,*

1. $\lambda_k^\sharp := \lambda^\sharp(x_k; \lambda_k, \mu_k)$, $k \in \mathcal{K}$, converges to $\lambda_* = \tilde{\lambda}(x_*)$.
2. There exists a constant $K > 0$ such that

$$\|\lambda_k^\sharp - \lambda_*\| \leq K(\omega_k + \|x_k - x_*\|)$$

for every $k \in \mathcal{K}$.

3. $\nabla \Phi(x_k; \lambda_k, \mu_k) \rightarrow \nabla L(x_*; \lambda_*)$, $k \in \mathcal{K}$.
4. There exists a constant $K' > 0$ such that for every $k \in \mathcal{K}$,

$$\|g(x_k)\| \leq K' \mu_k (\omega_k + \|\lambda_k - \lambda_*\| + \|x_k - x_*\|).$$

Proof. (1) Starting out with

$$\|\lambda_k^\sharp - \lambda_*\| \leq \|\lambda_k^\sharp - \tilde{\lambda}(x_k)\| + \|\tilde{\lambda}(x_k) - \lambda_*\|,$$

we observe that since $\lambda_* = \tilde{\lambda}(x_*)$, the second term on the right-hand side is of the order $\mathcal{O}(\|x_k - x_*\|)$, since $\tilde{\lambda}$ is Lipschitz on $U(x_*)$. Let us say $\|\tilde{\lambda}(x_k) - \lambda_*\| \leq K_0 \|x_k - x_*\|$ for some $K_0 > 0$. So in order to establish items 1 and 2, it remains to estimate the first term on the right-hand side. We have

$$\begin{aligned} \|\lambda_k^\sharp - \tilde{\lambda}(x_k)\| &= \|[J(x_k)\tilde{Z}(x_k)\tilde{Z}(x_k)^T J(x_k)^T]^{-1} J(x_k)\tilde{Z}(x_k)\tilde{Z}(x_k)^T \nabla f(x_k) + \lambda_k^\sharp\| \\ &\leq K_1 \|J(x_k)\tilde{Z}(x_k)\tilde{Z}(x_k)^T \nabla f(x_k) + J(x_k)\tilde{Z}(x_k)\tilde{Z}(x_k)^T J(x_k)^T \lambda_k^\sharp\| \\ &\leq K_1 K_2 \|\tilde{Z}(x_k)^T \nabla f(x_k) + \tilde{Z}(x_k)^T J(x_k)^T \lambda_k^\sharp\| \\ &= K_3 \|\tilde{Z}(x_k)^T \nabla \Phi(x_k; \lambda_k, \mu_k)\| \\ &\leq K_3 \omega_k. \end{aligned}$$

Here the second line comes from $\| [J(x_k) \tilde{Z}(x_k) \tilde{Z}(x_k)^T J(x_k)^T]^{-1} \| \leq K_1$ on a neighborhood $U(x_*)$, which is guaranteed by the rank hypothesis $(\mathcal{H}_2)$. From the same reason, in line 3, $\| J(x_k) \tilde{Z}(x_k) \| \leq K_2$ on $U(x_*)$ for some $K_2 > 0$. We let $K_3 = K_1 K_2$ and use the definition of $\lambda^\sharp$, which gives line 4. Finally, the last line follows from Lemma 4.1. Altogether, we obtain the estimate in item 2 with $K = \max\{K_0, K_3\}$.

(2) Now consider item 3. Observe that by our assumptions $x_k \rightarrow x_*$, $(k \in \mathcal{K})$, and $\omega_k \rightarrow 0$, so item 2 gives $\lambda_k^\sharp \rightarrow \lambda_*$. Therefore, $\nabla \Phi(x_k; \lambda_k, \mu_k) = \nabla L(x_k; \lambda_k^\sharp)$ converges to $\nabla L(x_*; \lambda_*)$, $k \in \mathcal{K}$.

(3) Finally, to see estimate 4 we multiply (4) by μ_k and take norms, which gives

$$\|g(x_k)\| = \mu_k \|\lambda_k^\sharp - \lambda_k\| \leq \mu_k (K(\omega_k + \|x_k - x_*\|) + \|\lambda_k - \lambda_*\|).$$

This is just the desired estimate in item 4 with $K' = \max\{K, 1\}$. $\square$

LEMMA 4.3. *With the same hypotheses, suppose $g(x_*) = 0$; then x_* is a KKT point, with corresponding Lagrange multiplier λ_* .*

Proof. To prove that x_* is a KKT point, we must show $P(x_*)(-\nabla L(x_*; \lambda_*)) = 0$, i.e., that $-\nabla L(x_*; \lambda_*)$ is in the normal cone to C at x_* . Since C is convex, this is equivalent to proving that for every test point $y \in C$, the angle between $-\nabla L(x_*; \lambda_*)$ and $y - x_*$ is at least 90° , i.e., that $-\nabla L(x_*; \lambda_*)^T (y - x_*) \leq 0$. Writing $\nabla \Phi_k = \nabla \Phi(x_k; \lambda_k, \mu_k)$, we first observe that by the stopping test (6),

$$\|P(x_k)(-\nabla \Phi_k)\| \leq \omega_k \rightarrow 0.$$

Let us now decompose the vector $-\nabla \Phi_k$ into its normal and tangential components at x_k , that is,

$$-\nabla \Phi_k = P(x_k)(-\nabla \Phi_k) + P^+(x_k)(-\nabla \Phi_k),$$

where $P^+(x_k)$ denotes the orthogonal projection onto $N(C, x_k)$, $P(x_k)$ as before the orthogonal projection onto $T(C, x_k)$. Such a decomposition is possible because the normal and tangent cones are polar cones of each other. Using this decomposition gives

$$\begin{aligned} -\nabla \Phi_k^T (y - x_k) &= P(x_k)(-\nabla \Phi_k)^T (y - x_k) + P^+(x_k)(-\nabla \Phi_k)^T (y - x_k) \\ &\leq P(x_k)(-\nabla \Phi_k)^T (y - x_k) \\ &\leq \omega_k \|y - x_k\| \rightarrow 0, \quad (k \in \mathcal{K}), \end{aligned}$$

where the last line uses the stopping test and Cauchy-Schwarz, while the second line comes from $P^+(x_k)(-\nabla \Phi_k)^T (y - x_k) \leq 0$, which is a consequence of the definition of $P^+(x_k)$ and the convexity of C . Altogether the term $-\nabla \Phi_k^T (y - x_k)$ converges to a quantity ≤ 0 , but by item 3 in Lemma 4.2, the same term also converges to $-\nabla L(x_*; \lambda_*)^T (y - x_*)$, $(k \in \mathcal{K})$. This proves $-\nabla L(x_*; \lambda_*)^T (y - x_*) \leq 0$. $\square$

THEOREM 4.4. *Let x_* be an accumulation point of a sequence x_k generated by the partially augmented Lagrangian algorithm such that hypotheses $(\mathcal{H}_1)$, $(\mathcal{H}_2)$ are satisfied at x_* . Suppose further that C admits a stratification into differentiable layers at x_* . Let $\mathcal{K} \subset \mathbb{N}$ be the index set of a subsequence converging to x_* . Let $\lambda_* := \tilde{\lambda}(x_*)$. Then we have the following:*

1. $\lambda_k^\sharp$, $k \in \mathcal{K}$, converges to λ_* . In particular, there exists a constant $K > 0$ such that $\|\lambda_k^\sharp - \lambda_*\| \leq K(\omega_k + \|x_k - x_*\|)$ for every $k \in \mathcal{K}$.
2. x_* is a KKT point, and λ_* is an associated Lagrange multiplier.
3. $\nabla \Phi(x_k; \lambda_k, \mu_k)$, $k \in \mathcal{K}$, converges to $\nabla L(x_*; \lambda_*)$.

Proof. Suppose first that μ_k is bounded away from 0. Then the algorithm eventually decides to do a first order update step at each iteration. Then $\|g(x_k)\| \leq \eta_k$, eventually, and $\eta_{k+1} = \mu^\beta \eta_k$ with $\mu^\beta < 1$ implies $\eta_k \rightarrow 0$. Therefore $g(x_*) = 0$. However, now the assumptions of Lemma 4.3 are all met, so we have the correct conclusions.

Now assume μ_k is not bounded away from 0. Assume $\mu_k \rightarrow 0$ for a subsequence. Then the construction of the parameters μ_k ensures that $\mu_k \|\lambda_k - \lambda_*\| \rightarrow 0$. This is exactly the argument from [20, Lemma 4.2], whose statement we reproduce below for the reader's convenience. So we arrive at the same conclusions, because now estimate 4 in Lemma 4.2 implies $g(x_*) = 0$. $\square$

LEMMA 4.5 (see [20, Lemma 4.2]). *Suppose μ_k , $k \in \mathcal{K}$, converges to 0. Then $\mu_k \|\lambda_k\|$, $k \in \mathcal{K}$, also converges to 0.*

The proof of Lemma 4.5 uses the specific form of the parameter updates in step 4 of the augmented Lagrangian algorithm. Any other update $\mu \rightarrow \mu^+$ for which the statement of Lemma 4.5 remains correct gives the same convergence result.

Remark. Notice that the weak convergence statement of Theorem 4.4 in terms of subsequences is the best we can hope to achieve in general. Reference [20] gives an example where the sequence x_k generated by the augmented Lagrangian algorithm has two accumulation points. A strict convergence result requires strong additional assumptions, like, for instance, convexity, which is not satisfied in cases we are interested in. On the other hand, in our experiments the method often converges nicely even without these hypotheses, so we consider Theorem 4.4 a satisfactory result.

5. SDP-representable sets. In this section we indicate in which way Theorem 4.4 may be extended to a larger class of convex constraint sets C . The motivating example are *SDP-representable sets*, a natural extension of LMI-sets as in (2). Recall that a closed convex set C is SDP-representable [10, 11] if it may be written in the form

$$C = \{x \in \mathbb{R}^n : \mathcal{A}(x, u) \preceq 0 \text{ for some } u \in \mathbb{R}^q\},$$

where $\mathcal{A} : \mathbb{R}^n \times \mathbb{R}^q \rightarrow \mathbb{S}_p$ is an affine operator. In other terms, SDP-representable sets are orthogonal projections of LMI-sets and may be considered the natural class of sets described by semidefinite programs. Notice that despite the similarity to LMI-sets, SDP-representable sets are a much larger class, including very interesting examples (see [10, 11]).

More generally, we may consider the class of closed convex sets C which are orthogonal projections of sets $\tilde{C}$ admitting a stratification into differentiable layers according to Definition 1. It is not clear whether Definition 1 is invariant under projections, which means that sets C of this type do not necessarily inherit this structure, and we cannot apply Theorem 4.4 directly to this class. Nonetheless, there is an easy way in which the partially augmented Lagrangian method can be extended to this larger class of sets C .

Consider program (P) with C the orthogonal projection of a set $\tilde{C}$, which admits a stratification into differentiable layers. Suppose without loss that C is the set of $x \in \mathbb{R}^n$ such that there exists $u \in \mathbb{R}^q$ with $(x, u) \in \tilde{C}$. It seems natural to consider the following program $(\tilde{P})$, which contains u as a slack variable and is equivalent to (P) :

$$\begin{aligned} & \text{minimize} && f(x), \quad x \in \mathbb{R}^n, \quad u \in \mathbb{R}^q \\ (\tilde{P}) \quad & \text{subject to} && g_j(x) = 0, \quad j = 1, \dots, m, \\ & && (x, u) \in \tilde{C}. \end{aligned}$$

This program is amenable to our convergence theorem as soon as the corresponding constraint qualification hypothesis is satisfied. At first sight, replacing (P) by $(\tilde{P})$ does not seem attractive, because we have introduced a slack variable. On second sight, however, we see that the impact of adding u is moderate.

Suppose we apply the partially augmented Lagrangian algorithm to program (P) , generating iterates $x_k \in C$, so that $(x_k, u_k) \in \tilde{C}$ for suitable $u_k \in \mathbb{R}^q$. Can we interpret (x_k, u_k) as a sequence of iterates generated by the same algorithm, but running for program $(\tilde{P})$ in (x, u) -space? If so, then convergence could be proved in (x, u) -space and would immediately imply convergence in x -space. This idea requires that we analyze the different steps of the algorithm in both settings.

Let us begin with the augmented version $(\tilde{P}_{\lambda, \mu})$ of program $(P_{\lambda, \mu})$. Since the partially augmented Lagrangian $\Phi(x, \lambda, \mu)$ does not depend on u , we realize that these two programs are exactly the same. This is good news, because on solving $(P_{\lambda, \mu})$ in x -space, as we naturally plan to do, we also implicitly solve $(\tilde{P}_{\lambda, \mu})$ in (x, u) -space.

What really needs to be done in (x, u) -space and not in x -space is the stopping test (6) in step 3 of our algorithm. What we propose to do is to modify the augmented Lagrangian scheme and accept $x^+ \in C$ as an approximate solution of $(P_{\lambda, \mu})$, and hence as the new iterate in x -space, if there exists u^+ such that $(x^+, u^+) \in \tilde{C}$ satisfies the stopping test (6) for the lifted program $(\tilde{P}_{\lambda, \mu})$. Explicitly this leads to the following test. Accept x^+ as soon as the solution (d_x, d_u) of

$$(12) \quad \inf \left\{ \| (-\nabla \Phi(x^+, \lambda, \mu), 0) - (d_x, d_u) \| : (d_x, d_u) \in T(\tilde{C}, (x^+, u^+)) \right\}$$

satisfies $\|(d_x, d_u)\| \leq \omega$. For definiteness, we may require here that u^+ be the smallest element in norm satisfying $(x^+, u^+) \in \tilde{C}$.

The last element of the algorithm to analyze concerns the parameter updates in step 4, and in particular the first order update rule. This is again identical in both settings, because the variable u does not intervene.

Altogether we have the following consequence of Theorem 4.4.

THEOREM 5.1. *Let C be a closed convex set which is the orthogonal projection of a closed convex set $\tilde{C}$ admitting a stratification into differentiable layers. Generate sequences $x_k \in C$, ω_k , η_k , λ_k , $\lambda_k^\sharp$, μ_k according to the partially augmented Lagrangian algorithm, with the difference that the stopping test (12) is applied at the point $(x_k, u_k) \in \tilde{C}$. Suppose (x_*, u_*) is an accumulation point of (x_k, u_k) such that hypotheses $(\mathcal{H}_1)$, $(\mathcal{H}_2)$ are satisfied at (x_*, u_*) . Let $\mathcal{K} \subset \mathbb{N}$ the index set of a convergent subsequence. Let $\lambda_* = \tilde{\lambda}(x_*)$; then*

1. $\lambda_k^\sharp$, $k \in \mathcal{K}$, converge to λ_* . In particular, there exists a constant $K > 0$ such that $\|\lambda_k^\sharp - \lambda_*\| \leq K(\omega_k + \|x_k - x_*\|)$ for every $k \in \mathcal{K}$.
2. x_* is a KKT point for (P) , and λ_* is an associated Lagrange multiplier.
3. $\nabla \Phi(x_k; \lambda_k, \mu_k)$, $k \in \mathcal{K}$, converges to $\nabla L(x_*; \lambda_*)$.

One may wonder whether it is really necessary to solve the stopping test in (x, u) -space all the time. Obviously, as soon as the orthogonal projection of $T(\tilde{C}, (x^+, u^+))$ is identical with $T(C, x^+)$, solving (6) and (12) is equivalent. In general, however, this is not the case. We have only the trivial inclusion $\pi(T(\tilde{C}, (x^+, u^+))) \subset T(C, x^+)$, where π denotes the projection $(x, u) \mapsto x$, which also shows that the stopping test (12) is stronger than (6).

A particular case where equality holds is when (x^+, u^+) is a smooth point of the boundary of $\tilde{C}$, because then x^+ is also smooth for C . Since almost all points in the boundary of a convex set are smooth points, this is quite satisfactory.

6. Discussion. In this section we briefly discuss the hypotheses in Theorems 4.4 and 5.1 and then pass to practical aspects of the algorithm. Both results use the constraint qualification hypothesis $(\mathcal{H}_2)$, which as we have seen reduces to a familiar condition in the case of classical programming. Notice that for $m \geq 1$, $(\mathcal{H}_2)$ excludes in particular corner points x of the constraint set C , which would have $V(C, x) = \{0\}$. An assumption like $(\mathcal{H}_2)$ is already required to obtain suitable KKT conditions.

The additional hypothesis of boundedness of the gradients $\nabla\Phi(x_k; \lambda_k, \mu_k)$ has been made in several approaches (see [20]). Our present approach shows that this hypothesis can be avoided.

We recall that the original idea of the augmented Lagrangian method [47] was to improve on pure penalty methods insofar as the penalty parameter μ_k no longer needed to be driven to 0 to yield convergence—a major advantage because ill-conditioning is avoided. For the partially augmented Lagrangian method with polyhedral sets, a similar result is proved in [23]. We can establish such a result for matrix inequality constraints if a second order sufficient optimality condition stronger than the no-gap condition in [13] is satisfied. Details will be presented elsewhere. The phenomenon is confirmed by experiments, where μ_k is very often frozen at a moderately small size.

Let us now consider some practical aspects of the partially augmented Lagrangian for LMI constrained sets $C = \{x \in \mathbb{R}^n : \mathcal{A}(x) \preceq 0\}$. Observe that the stopping test (6) may be computed by solving an SDP. According to [52], the tangent cone at $x_0 \in C$ is $T(C, x_0) = \{d \in \mathbb{R}^n : Y_1^T(\mathcal{A}_*d)Y_1 \preceq 0\}$, where the columns of Y_1 form an orthonormal basis of the eigenspace of $\lambda_1(\mathcal{A}(x_0))$, and where $\mathcal{A}_*$ is the linear part of $\mathcal{A}$, i.e., $\mathcal{A}_*d = \sum_{i=1}^n A_i d_i$. Letting $\nabla\Phi := \nabla\Phi(x^+; \lambda, \mu)$, the stopping test (6) leads to the LMI constrained least squares program

$$(13) \quad \min\{\| -\nabla\Phi - d \| : Y_1^T(\mathcal{A}_*d)Y_1 \preceq 0, d \in \mathbb{R}^n\}.$$

An equivalent cast as an SDP is

$$(14) \quad \begin{array}{ll} \text{minimize} & t \\ \text{subject to} & \begin{pmatrix} I_n & \nabla\Phi + d \\ * & t \end{pmatrix} \succeq 0, \quad Y_1^T(\mathcal{A}_*d)Y_1 \preceq 0, \end{array}$$

where the decision variable is now $(t, d) \in \mathbb{R} \times \mathbb{R}^n$. Notice that in general the column rank r of Y_1 is much smaller than the size of $\mathcal{A}$, so a full spectral decomposition of $\mathcal{A}(x_0)$ is not required and the program data of (13) or (14) are obtained efficiently. For large-dimension n , it may therefore be interesting to solve the dual of (13), which is readily obtained as

$$\min \left\{ \frac{1}{2} \|\mathcal{A}_*^T(Y_1 Z Y_1^T)\|^2 + \mathcal{A}_* \nabla\Phi \bullet (Y_1 Z Y_1^T) : Z \succeq 0, Z \in \mathbb{S}_r \right\},$$

with return formula $d = -\nabla\Phi - \mathcal{A}_*^T(Y_1 Z Y_1^T)$ relating dual and primal optimal solutions.

Most of the time the multiplicity r of $\lambda_1(\mathcal{A}(x_0))$ even equals 1. Then the LMI-constraint $Y_1^T(\mathcal{A}_*d)Y_1 \preceq 0$ in (13) and (14) becomes the scalar constraint $e_1^T(\mathcal{A}_*d)e_1 \leq 0$, where e_1 is the normalized eigenvector of $\lambda_1(\mathcal{A}(x_0))$. This may also be written as $[\mathcal{A}_*^T e_1 e_1^T]^T d \leq 0$, where the adjoint $\mathcal{A}_*^T$ of the linear part of $\mathcal{A}(x)$ is defined as $\mathcal{A}_*^T Z = (A_1 \bullet Z, \dots, A_n \bullet Z)$, and where $e_1 e_1^T$ is of rank 1. Then (13) is an inequality constrained least squares program, $\min\{\|g - d\| : d \in \mathbb{R}^n, h^T d \leq 0\}$, which has an

explicit solution:

$$d = \begin{cases} g - \frac{g^T h}{\|h\|^2} h & \text{if } g^T h \geq 0, \\ g & \text{if } g^T h \leq 0, \end{cases} \quad \text{where } g := -\nabla \Phi, \quad h := \mathcal{A}_*^T e_1 e_1^T.$$

In practice $g = -\nabla \Phi$ clearly points away from the half space $h^T d \leq 0$, so that the first case occurs, which we recognize as the projection of g onto the hyperplane $h^T d = 0$.

To conclude, recall that the partially augmented Lagrangian scheme clearly hinges on the possibility of solving the approximate programs $(P_{\lambda,\mu})$ much faster than the full program (P) . To this end, the structure of C should be sufficiently simple, since $(P_{\lambda,\mu})$ has to be solved many times.

7. Applications. In our experimental section we test the augmented Lagrangian method on two typical applications of program (S) in feedback control synthesis. We start with static output-feedback H_∞ -synthesis in section 7.1 and present numerical tests in sections 7.2 and 7.3. A second application is parametric robust control design, which is considered in section 7.4. A case study in section 7.5 concludes the experimental part.

7.1. Static H_∞ -synthesis. Static H_∞ -control design is an NP-hard problem. Due to its great practical importance many heuristic approaches have been proposed; see, e.g., [8, 24, 41, 26]. Solutions based on nonlinear optimization are, for instance, [37, 38] or [16]. We have proposed several optimization-based approaches in [5, 4, 28, 29]. Here we show how this problem may be solved with the help of our augmented Lagrangian algorithm.

A detailed description of the static H_∞ -problem and a comprehensive discussion are presented in [5, 6]. Here we only briefly recall the outset. Consider a linear time-invariant plant described in standard form by its state-space equations:

$$(15) \quad P(s) : \quad \begin{bmatrix} \dot{x} \\ z \\ y \end{bmatrix} = \begin{bmatrix} A & B_1 & B_2 \\ C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{22} \end{bmatrix} \begin{bmatrix} x \\ w \\ u \end{bmatrix},$$

where $x \in \mathbb{R}^n$ is the state vector, $u \in \mathbb{R}^{m_2}$ are the control inputs, $w \in \mathbb{R}^{m_1}$ is an exogenous signal, $y \in \mathbb{R}^{p_2}$ is the vector of measurements, and $z \in \mathbb{R}^{p_1}$ is the vector of controlled or performance variables. After substitution into (15), any static output feedback control law $u = Ky$ induces a closed-loop transfer function $T_{w,z}(s)$ from w to z , called the *performance channel*. Our aim is now to compute a static controller K which meets the following design requirements:

Stability. It stabilizes the plant.

Performance. Among all stabilizing controllers, K minimizes the H_∞ -norm $\|T_{w,z}(s)\|_\infty$.

The closed-loop system is first transformed into a matrix inequality using the Bounded Real Lemma [1]. Then the Projection Lemma [31] is used to eliminate the unknown controller data K from the cast. We obtain the following.

PROPOSITION 7.1. *A stabilizing static output feedback controller K with H_∞ -gain $\|T_{w,z}(s)\|_\infty \leq \gamma$ exists provided there exist $X, Y \in \mathbb{S}^n$ such that*

TABLE 1
Problem dimensions.

pb.	n	m_2	p_2	m_1	p_1	var	LMI	const
pb5	5	2	2	2	2	31	25	25
pb10	10	2	3	3	3	111	48	100
pb15	15	3	3	3	3	241	67	225
pb20	20	3	4	5	5	421	94	400
pb25	25	3	4	5	5	651	114	625
pb30	30	5	6	6	7	931	136	900
pb35	35	5	6	6	7	1261	156	1225

$$(16) \quad \mathcal{N}_Q^T \begin{bmatrix} A^T X + X A & X B_1 & C_1^T \\ B_1^T X & -\gamma I & D_{11}^T \\ C_1 & D_{11} & -\gamma I \end{bmatrix} \mathcal{N}_Q \prec 0,$$

$$(17) \quad \mathcal{N}_P^T \begin{bmatrix} Y A^T + A Y & B_1 & Y C_1^T \\ B_1^T & -\gamma I & D_{11}^T \\ C_1 Y & D_{11} & -\gamma I \end{bmatrix} \mathcal{N}_P \prec 0,$$

$$(18) \quad X \succ 0, \quad Y \succ 0, \quad XY - I = 0,$$

where $\mathcal{N}_Q$ and $\mathcal{N}_P$ denote bases of the null spaces of $Q := \begin{bmatrix} C_1 & D_{21} & 0 \end{bmatrix}$ and $P := \begin{bmatrix} B_1^T & D_{12}^T & 0 \end{bmatrix}$.

It is convenient [26] to replace positive definiteness in (18) by

$$(19) \quad \begin{bmatrix} X & I \\ I & Y \end{bmatrix} \succ 0.$$

On the other hand, the nonlinear equality $XY - I = 0$ cannot be removed and renders the problem difficult. The cast (16)–(18) and (19) is now of the form (S) if we replace strict inequalities $\prec 0$ by suitable $\preceq -\varepsilon$. The objective to be minimized is γ . The dimension of the decision variable $x = (X, Y, \gamma)$ is $1 + n(n+1)$, displayed as var in Table 1. The size of the LMIs is displayed in the column labeled LMI. It depends on the dimensions of $\mathcal{N}_P$ and $\mathcal{N}_Q$ and due to possible rank deficiency cannot be computed in advance. The nonlinear equality constraint in the terminology of (S) corresponds to a function $g : \mathbb{R}^{n(n+1)} \rightarrow \mathbb{R}^{n^2}$. The last column const in Table 1 therefore displays n^2 .

Once solved via the augmented Lagrangian method, this procedure requires an additional step, where the controller K , which has been eliminated from the cast, needs to be restored from the decision parameters of (S). This last step may be based on the method in [31] and, as a rule, does not present any numerical difficulties.

7.2. Numerical experiment I. In our first experiment we solve a series of static output-feedback H_∞ -synthesis problems randomly generated via the procedure in [43] at different sizes n ranging from 5 to 35. In each case it is known that a stabilizing static controller K exists, but the global optimal gain $\gamma = \|T_{w,z}(s)\|_\infty$ is not known. Dimensions of our test problems are described in Table 1.

While n, m_2, p_2, m_1, p_1 refer to the plant (15), columns var, LMI, and const display for each problem the number of decision variables, the LMI size, and the number of nonlinear equality constraints in $g(x) = 0$.

In Table 2, the column $P_{\lambda,\mu}$ gives the number of instances of the augmented Lagrangian subproblem. Each of these programs is solved by a succession of SDPs, and the column labeled SDP therefore gives the total number of SDPs needed to solve

TABLE 2
Results of static H_∞ -synthesis.

pb.	$P_{\lambda,\mu}$	SDP	μ	ω	$\ g\ _\infty$	Full/static
pb5	16	20	$1.42e-2$	$1.12e-2$	$5.71e-6$	$9.63e-5$; 3.49
pb10	20	21	$5.58e-5$	$0.63e-2$	$2.08e-7$	3.44; 3.48
pb15	21	27	$2.18e-4$	$3.28e-2$	$7.55e-6$	3.44; 3.48
pb20	16	21	$9.07e-4$	$8.05e-3$	$9.14e-6$	3.25; 3.74
pb25	21	22	$5.50e-5$	$5.35e-2$	$5.90e-6$	4.60; 4.61
pb30	26	32	$3.34e-6$	$4.23e-2$	$9.77e-6$	1.099; 1.317
pb35	28	35	$8.26e-6$	$2.07e-2$	$4.43e-6$	6.47; 8.46

(P). As a rule, only between one and two SDPs per subproblem ($P_{\lambda,\mu}$) are needed. The number of SDPs needed to solve the augmented Lagrangian problem (P) may be considered the crucial parameter to judge the speed of our approach.

In our tests, SDPs are solved with an alpha version of our own spectral SDP code, which minimizes convex quadratic objectives subject to LMI-constraints

$$(20) \quad \begin{aligned} & \text{minimize} && c^T x + \frac{1}{2} x^T Q x \\ & \text{subject to} && \mathcal{A}(x) \preceq 0. \end{aligned}$$

In contrast, currently available SDP solvers are often based on the cast

$$\min\{c^T x : \mathcal{A}(x) \preceq 0\}.$$

We have observed that those run into numerical problems very early, since the quadratic term $x^T Q x$ in the objective of (20) has to be converted into an LMI via Schur complement. This leads to large-size LMIs very quickly. For the problems in Table 1 the corresponding augmented LMIs are of size 57×57 in pb5, 160×160 in pb10, 309×309 in pb15, 516×516 in pb20, 766×766 in pb25, 1068×1068 in pb30, and 1418×1418 in pb35.

The remaining entries in Table 2 are as follows. Column μ gives the final value of the penalty parameter, while $\|g\|_\infty$ gives the final precision in the equality constraint. In each of our test cases this precision was small enough in order to enable the procedure in [31] to find a controller K meeting both design specifications, stability and H_∞ -performance. This may be regarded as the ultimate test of success of the method. The column ω gives the final value $\|P(-\nabla\Phi)\|$ used in the stopping test (6). We have observed that (6) should be employed rather tolerantly, which suggests using a comparatively large stopping tolerance ω_* in step 2 of the augmented Lagrangian algorithm. (This is also reflected by the fact that the covering sequence ω_k converges to 0 fairly slowly.)

The column full/static should be interpreted with care. It compares the performance $\gamma = \|T_{w,z}(s)\|_\infty$ achieved by the solution of (S) to the lower bound γ_∞ of the full H_∞ -controller, computed by the usual SDP or Riccati method. In general γ_∞ cannot be a tight lower bound for the best possible γ in (S), but in a considerable number of cases both gains are fairly close. This indicates that our method, as a rule, gets close to the global minimum of (S), even though theoretical evidence for this is lacking. Notice here that even cases with a large gap between γ and γ_∞ do not contradict this supposition. One may always artificially arrange a large gap by creating a poorly actuated system, that is, a system where the number of control inputs is much smaller than the state of the system, $m_2 \ll n$.

For the set of test examples (S) in automatic control we observed that the partially augmented Lagrangian method required a rather limited number of SDP subproblems.

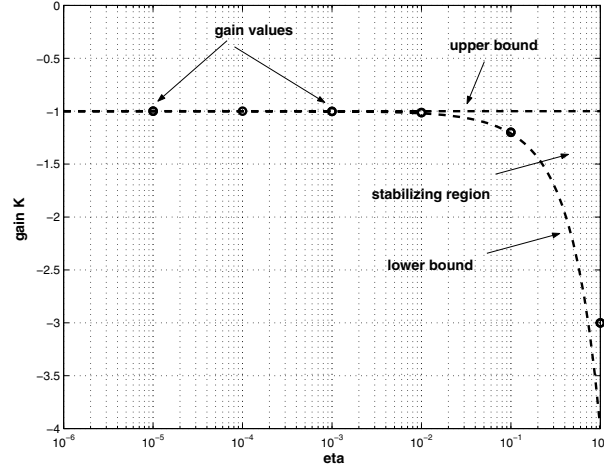


FIG. 1. Stabilizing static gains and stabilizing region.

This makes our approach based on a succession of SDPs attractive. Moreover, we observed that a solution to the control problem is practically always found provided the SDP subproblems can be solved in a reasonable time. The bottleneck of our approach is the SDP solver. The alpha version of our own solver performed well up to systems of size $n = 35$, which means 1261 decision variables and LMIs of size 156×156 (or of size 1418×1418 if the quadratic term is Schur complemented into an LMI).

7.3. Numerical experiment II. In this section we present an experiment imported from [41]. Numerical data are

$$A = \begin{bmatrix} 1 & 1 + \eta \\ -(1 + \eta) & 0 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad C_2 = [0 \quad 1],$$

where η is some small positive parameter. We consider the static stabilization problem for various values of η , so this example does not include performance, meaning $B_1 = 0$, $D_{11} = 0$, $C_1 = 0$, $D_{12} = 0$. The attractive feature of this example is that feasible values for the gain reduce to a small interval $K \in (-(1 + \eta)^2, -1)$ whose length decreases quickly as η tends to zero. As a consequence, most existing algorithms fail when the interval shrinks significantly. Also, this example provides an indication on the accuracy and reliability of the proposed augmented Lagrangian algorithm.

Using the proposed technique, we computed a stabilizing gain K for a set of values $\eta \in \{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$. The results are shown in Figure 1, where in logarithmic scale the tube of admissible values K appears as the zone between the horizontal line at $K = -1$ and the curved lower bound. The achieved values of K are indicated through small circles and, as expected, lie within the stabilizing region. Note that the stabilizing region shrinks dramatically as η gets smaller. For $\eta = 10^{-6}$ the algorithm fails since the system can be regarded as numerically unstabilizable. The admissible interval length has then been reduced to about $2.0 \cdot 10^{-6}$. The problem has 13 decision variables.

In parallel with the experiments presented in sections 7.2 and 7.3, we point the reader to the testing in [5], where reduced order synthesis (including the static case) is examined from various other points of view.

7.4. Robust synthesis with time-varying uncertainties. In this section we consider a second class of automatic control applications of program (S), the robust control problem for an uncertain plant subject to parametric uncertainties. These system uncertainties may be described by so-called linear fractional transforms (LFTs):

$$(21) \quad \begin{bmatrix} \dot{x} \\ z_\Delta \\ z \\ y \end{bmatrix} = \begin{bmatrix} A & B_\Delta & B_1 & B_2 \\ C_\Delta & D_{\Delta\Delta} & D_{\Delta 1} & D_{\Delta 2} \\ C_1 & D_{1\Delta} & D_{11} & D_{12} \\ C_2 & D_{2\Delta} & D_{21} & 0 \end{bmatrix} \begin{bmatrix} x \\ w_\Delta \\ w \\ u \end{bmatrix},$$

$$w_\Delta = \Delta(t) z_\Delta.$$

Here $\Delta(t)$ is a time-varying matrix-valued parameter, usually assumed to have a block-diagonal structure

$$(22) \quad \Delta(t) = \text{diag}(\dots, \delta_i(t)I, \dots, \Delta_j(t), \dots) \in \mathbf{R}^{N \times N}$$

normalized such that

$$(23) \quad \Delta(t)^T \Delta(t) \leq I \quad \forall t \geq 0.$$

According to the μ analysis and synthesis literature [27, 24], blocks $\delta_i I$ and Δ_j are referred to as repeated-scalar blocks and full blocks, respectively. Straightforward computations lead to the state-space representation

$$\begin{bmatrix} \dot{x} \\ z \\ y \end{bmatrix} = \left\{ \begin{bmatrix} A & B_1 & B_2 \\ C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & 0 \end{bmatrix} + \begin{bmatrix} B_\Delta \\ D_{1\Delta} \\ D_{2\Delta} \end{bmatrix} \Delta(t) (I - D_{\Delta\Delta} \Delta(t))^{-1} \begin{bmatrix} C_\Delta & D_{\Delta 1} & D_{\Delta 2} \end{bmatrix} \right\} \times \begin{bmatrix} x \\ w \\ u \end{bmatrix},$$

which means that the state-space data of the plant with inputs w and u and outputs z and y are fractional functions of the time-varying uncertain parameter $\Delta(t)$. Hence we have the name *LFT*. In (21), u is the control input, w denotes the exogenous input, z denotes the controlled performance variables, and y denotes the measurement signal or output. Given the uncertain plant (21)–(23), the robust control problem requires finding a linear time-invariant (LTI) controller

$$(24) \quad \begin{aligned} \dot{x}_K &= A_K x_K + B_K y, \\ u &= C_K x_K + D_K y, \end{aligned}$$

such that for *all* admissible parameter trajectories (22), (23),

- the closed-loop system, obtained by substituting (24) into (21)–(23), is internally stable.
- the L_2 -induced gain of the operator connecting w to z is bounded by γ .

Note that the performance specification says that

$$\int_0^\infty z(t)^T z(t) dt \leq \gamma^2 \int_0^\infty w(t)^T w(t) dt \quad \forall w \in L_2, \quad \forall \Delta(t) \text{ as in (22), (23)}.$$

It has been shown that such problems may be handled via a version of the Bounded Real Lemma [1], which translates these conditions into matrix inequalities. The resulting sufficient conditions are the following: Find a Lyapunov matrix X_{cl} , $X_{cl} \succ 0$ and appropriately structured scalings S and T such that

$$(25) \quad \begin{bmatrix} A_{cl}^T X_{cl} + X_{cl} A_{cl} & * & * \\ B_{cl}^T X_{cl} + \begin{bmatrix} T & 0 \\ 0 & 0 \end{bmatrix} C_{cl} & - \begin{bmatrix} S & 0 \\ 0 & \gamma I \end{bmatrix} & * \\ & + \begin{bmatrix} T & 0 \\ 0 & 0 \end{bmatrix} D_{cl} + D_{cl}^T \begin{bmatrix} T & 0 \\ 0 & 0 \end{bmatrix}^T & * \\ C_{cl} & D_{cl} & - \begin{bmatrix} S & 0 \\ 0 & \gamma I \end{bmatrix}^{-1} \end{bmatrix} \prec 0.$$

Here A_{cl} , B_{cl} , C_{cl} , and D_{cl} are the state-space data of the closed-loop system (21)–(24) with the loop $w_\Delta = \Delta(t) z_\Delta$ still open. *Appropriately structured* means that the scalings are compatible with the uncertainty structure (22), (23), for which we use the shorthand notation $\Delta \in \mathbf{\Delta}$. The set of symmetric scalings associated with the parameter structure $\mathbf{\Delta}$ is defined as

$$S_{\mathbf{\Delta}} := \{S : S^T = S, \quad S\Delta = \Delta S \quad \forall \Delta \in \mathbf{\Delta}\}.$$

Similarly, the set of skew-symmetric scalings associated with the parameter structure $\mathbf{\Delta}$ is defined as

$$T_{\mathbf{\Delta}} := \{T : T^T = -T, \quad T\Delta = \Delta^T T \quad \forall \Delta \in \mathbf{\Delta}\}.$$

The idea here is that as soon as $\Delta \in \mathbf{\Delta}$ and $S \succ 0$, $S \in S_{\mathbf{\Delta}}$, $T \in T_{\mathbf{\Delta}}$, the quadratic constraint

$$(26) \quad \begin{bmatrix} \Delta \\ I \end{bmatrix}^T \begin{bmatrix} -S & T^T \\ T & S \end{bmatrix} \begin{bmatrix} \Delta \\ I \end{bmatrix} \succeq 0$$

is automatically satisfied. With the above definitions and notation, the following algebraically constrained LMI characterization for the solvability of the problem can be established. The reader is referred to [45, 44, 2, 3, 33, 51, 25] for more details and additional results.

THEOREM 7.2. *Consider the plant (21) with uncertain trajectories $\Delta(t)$ as in (22), (23). The following conditions guarantee the existence of an LTI output-feedback controller (24), which stabilizes the closed-loop system (21)–(24) for all admissible trajectories $\Delta \in \mathbf{\Delta}$ such that, in addition, the L_2 gain performance γ is achieved: There exist symmetric matrices X, Y, S, Σ and skew-symmetric matrices T, Γ such that*

$$(27) \quad S, \Sigma \in S_{\mathbf{\Delta}} \quad \text{and} \quad T, \Gamma \in T_{\mathbf{\Delta}},$$

and such that the matrix inequalities

$$(28) \quad \mathcal{N}_1^T \begin{bmatrix} A^T X + X A & X B_\Delta + C_\Delta^T T^T & X B_1 & C_\Delta^T S & C_1^T \\ B_\Delta^T X + T C_\Delta & -S + T D_{\Delta\Delta} + D_{\Delta\Delta}^T T^T & T D_{\Delta 1} & D_{\Delta\Delta}^T S & D_{\Delta 1}^T \\ B_1^T X & D_{\Delta 1}^T T^T & -\gamma I & D_{\Delta 1}^T S & D_{11}^T \\ S C_\Delta & S D_{\Delta\Delta} & S D_{\Delta 1} & -S & 0 \\ C_1 & D_{1\Delta} & D_{11} & 0 & -\gamma I \end{bmatrix} \mathcal{N}_1 \prec 0,$$

$$(29) \quad \mathcal{N}_2^T \begin{bmatrix} A Y + Y A^T & Y C_\Delta^T + B_\Delta \Gamma^T & Y C_1^T & B_\Delta \Sigma & B_1 \\ C_\Delta Y + \Gamma B_\Delta^T & -\Sigma + \Gamma D_{\Delta\Delta}^T + D_{\Delta\Delta} \Gamma^T & \Gamma D_{\Delta 1}^T & D_{\Delta\Delta} \Sigma & D_{\Delta 1} \\ C_1 Y & D_{1\Delta} \Gamma^T & -\gamma I & D_{1\Delta} \Sigma & D_{11} \\ \Sigma B_\Delta^T & \Sigma D_{\Delta\Delta}^T & \Sigma D_{\Delta 1}^T & -\Sigma & 0 \\ B_1^T & D_{\Delta 1}^T & D_{11}^T & 0 & -\gamma I \end{bmatrix} \mathcal{N}_2 \prec 0,$$

$$(30) \quad \begin{bmatrix} X & I \\ I & Y \end{bmatrix} \succ 0,$$

$$(31) \quad \begin{bmatrix} S & 0 \\ 0 & \Sigma \end{bmatrix} \succ 0$$

in tandem with the algebraic constraints

$$(32) \quad (S + T)^{-1} = (\Sigma + \Gamma), \quad \text{or equivalently,} \quad \begin{bmatrix} S & T \\ T^T & -S \end{bmatrix}^{-1} = \begin{bmatrix} \Sigma & \Gamma^T \\ \Gamma & -\Sigma \end{bmatrix},$$

are satisfied. Here $\mathcal{N}_1$ is a basis of the null space of $[C_2, D_{2\Delta}, D_{21}, 0]$, and $\mathcal{N}_2$ is a basis of the null spaces of $[B_2^T, D_{\Delta 2}^T, D_{12}^T, 0]$. $\square$

Note that due to the algebraic constraints (32), the problem under consideration is NP-hard [9] and not solvable via SDP. Even simpler instances of this problem like those considered in [39] are already NP-hard. (This is in sharp contrast to the associated nominal H_∞ -synthesis problem, which reduces to a standard SDP since the nonlinear conditions (32) fully disappear.) This is our second example of how a program of type (S) may be obtained in lieu of a program of type (B), based directly on (25). Once again this rests on a diligent use of the Projection Lemma.

The explicit form (S) is obtained through the following steps. As previously done, replace strict inequalities $\prec 0$ by $\preceq -\varepsilon I$. For the structure (22), conditions (27) imply a typical block structure for the matrices S, Σ, T, Γ , so the conditions (31) reduce to blocks of LMIs, to which the nonlinear equality constraints (32) have to be added. The cost function to be minimized is γ , and the decision vector is $x = (X, Y, \gamma, S, T, \Sigma, \Gamma)$, which regroups the gain γ , the multiplier variables S, Σ, T, Γ , and the Lyapunov matrix variables X, Y . As before, due to the Projection Lemma, the controller data do not directly enter the decision vector x and have to be recovered from the optimal x through the procedure in [31]. With these elements, the problem is directly open to our partially augmented Lagrangian algorithm, and the numerical tests presented below have been obtained accordingly.

7.5. Numerical experiment III: Flexible satellite. We consider the design of a robust attitude control system for a flexible satellite, adopted from [14, 17]. Despite its seemingly moderate size, this problem has been identified as a difficult case,

where nominal H_∞ -synthesis fails and robust techniques are required. We confine our study to the yaw axis of the satellite, whose dynamics are of the form

$$M\ddot{\phi} + D\dot{\phi} + K\phi = Tu,$$

where $\phi = [\phi_1 \ \phi_2]^T$ and ϕ_1 is the yaw angle displacement in radians, ϕ_2 is the modal displacement used to represent the flexible dynamics, and u is the control torque (inch-pounds). Numerical data are given as

$$(33) \quad \begin{aligned} M &= \begin{bmatrix} 77511 & 248.1 \\ 248.1 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 0 & 0 \\ 0 & 0.002288 \end{bmatrix}, \\ K &= \begin{bmatrix} 0 & 0 \\ 0 & k_0 \end{bmatrix}, \quad T = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \end{aligned}$$

with $k_0 = 0.104124$. As is often convenient in applications, we augment the model by the (redundant) integral of ϕ_1 . This introduces a new variable $\phi_3 = \phi_1$, whose role will become clear when performance variables will be specified. The system can then be rewritten in first order form as

$$\dot{x} = Ax + Bu,$$

where $x = [\phi_1, \phi_2, \dot{\phi}_1, \dot{\phi}_2, \int \phi_1 dt] \in \mathbb{R}^5$ and

$$A = \begin{bmatrix} 0_{2 \times 2} & I_2 & 0_{2 \times 1} \\ -M^{-1}K & -M^{-1}D & 0_{2 \times 1} \\ [1 \ 0] & 0_{1 \times 2} & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0_{2 \times 1} \\ M^{-1}T \\ 0 \end{bmatrix}.$$

The measured variables are $y = [\phi_1, \dot{\phi}_1, \int \phi_1 dt] \in \mathbb{R}^3$. Specifications in this design problem are twofold:

- We wish to maintain a pointing accuracy of $\leq 4.0 \cdot 10^{-4}$ radian (0.023 degree) in the yaw angle.
- This pointing accuracy must be guaranteed in the presence of 25% variation in the structural frequency due to uncertainties in the stiffness matrix M .

It was shown in [17] that the requested 25% variation in the structural frequency corresponds to the uncertainty $0.053296 \leq k \leq 0.154953$ in the parameter k_0 . This in turn leads to the uncertain model $k = k_0 + W_\delta \delta$ with $k_0 = 0.104124$, the nominal value, $\delta \in [-1, 1]$, and $W_\delta = 0.0050829$. The satellite model incorporating uncertainty is now obtained as

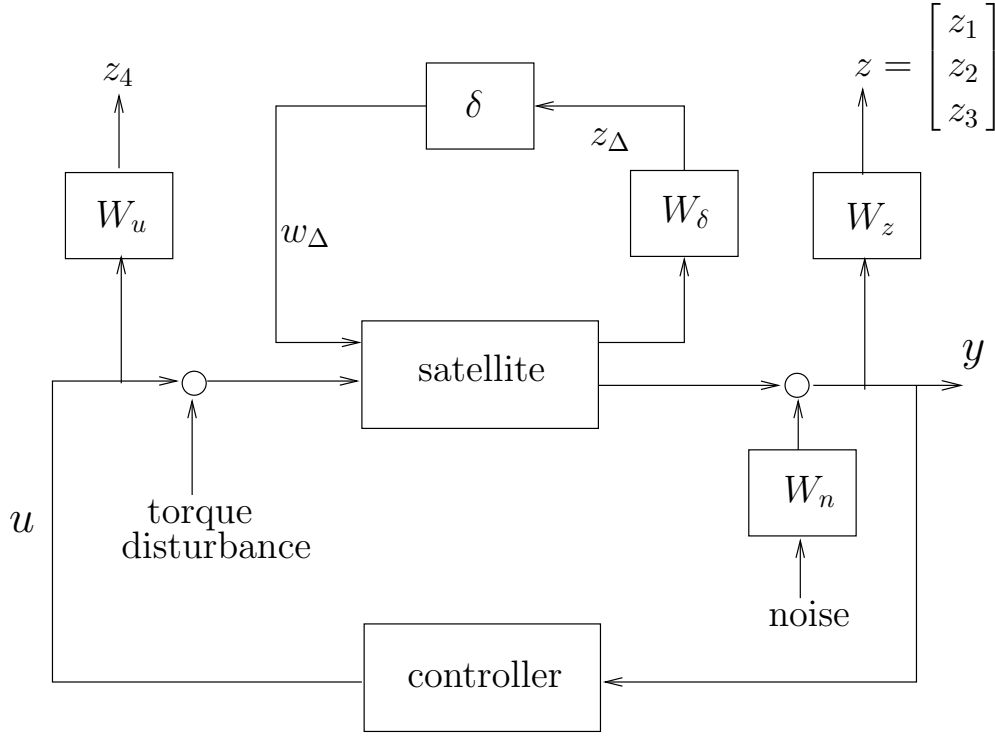
$$\begin{bmatrix} \dot{x} \\ z_\Delta \\ y \end{bmatrix} = \begin{bmatrix} A & B_\Delta & B \\ C_\Delta & D_{\Delta\Delta} & D_{\Delta 2} \\ C & D_{2\Delta} & D_{22} \end{bmatrix} \begin{bmatrix} x \\ w_\Delta \\ u \end{bmatrix},$$

$$w_\Delta = \delta z_\Delta, \quad \delta \in [-1, 1],$$

where A, B are as before and

$$\begin{aligned} B_\Delta &= [0 \ 0 \ m_{12} \ m_{22} \ 0], \\ C_\Delta &= W_\delta [0 \ 1 \ 0 \ 0 \ 0], \\ D_{\Delta\Delta} &= 0, \ D_{\Delta 2} = 0, \ D_{2\Delta} = 0, \ D_{22} = 0, \end{aligned}$$

and where m_{12} and m_{22} are the (1, 2) and (2, 2) elements of $-M^{-1}$, respectively. The matrix $C \in \mathbb{R}^{3 \times 5}$ is such that $y = Cx$.

FIG. 2. *Synthesis architecture.*

This uncertain model has now to be completed by specifying exogenous input w and controlled output signals z . The data of the synthesis structure were all taken from [14] except for the noise weighting, W_n , which we have increased from 10^{-6} to 10^{-4} in order to comply with the increased performance request. This modification will highlight the differences between nominal and robust syntheses.

Inspecting the overall synthesis architecture in Figure 2, we see in which way the controller interacts with the satellite, and also how exogenous signals and performance signals are specified. We note that as is common in control system design, the measurements y are corrupted by noise. The noise magnitude is determined through the so-called weighting filter $W_n = 10^{-4}$ as discussed above, and this represents our first exogenous signal $\text{noise} = (w_1, w_2, w_3) \in \mathbb{R}^3$. The satellite is also subject to torque disturbance, which tends to deviate the yaw angle offset from its nominal zero level. This disturbance represents the fourth exogenous input w_4 .

The loop featuring the δ block corresponds to the diagram representation of the LFT uncertainty in the structural frequency. The magnitude of the uncertainty is specified by the weighting $W_\delta = 0.0050829$ introduced to obtain the normalization $|\delta| \leq 1$.

A performance vector $(z_1, z_2, z_3) \in \mathbb{R}^3$ is introduced to reduce the impact of both noise and torque disturbance on the yaw angle offset. The corresponding channel is $(z_1, z_2, z_3) = W_z([\phi_1; \dot{\phi}_1; \int \phi_1 dt] + W_n \text{ noise})$, where the weighting is defined as $W_z = \text{diag}(400; 1; 20)$ and serves to specify the relative importance of the entries z_1, z_2, z_3 in the minimization. Indeed, z_1 specifies high-, z_2 specifies medium-, and z_3 specifies low-frequency parts of the yaw angle offset, and W_z allows us to address these components

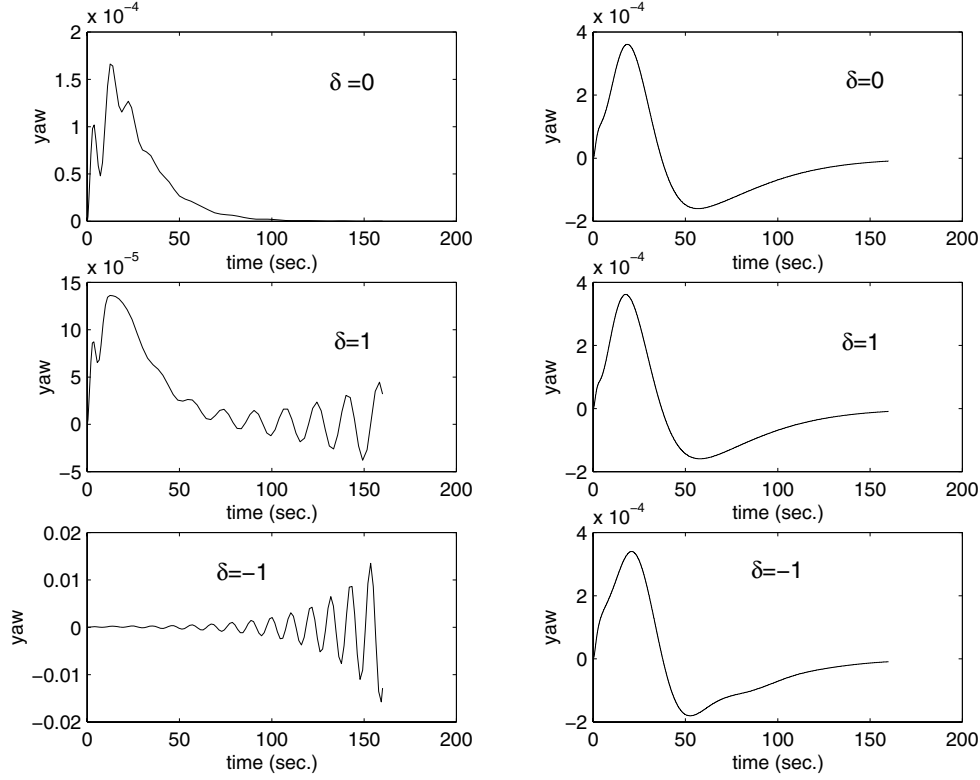


FIG. 3. Simulations of yaw angle for $\delta = 0; 1; -1$. Left column: nominal controller. Right column: robust controller.

individually. The effect of controlling (z_1, z_2, z_3) on the torque disturbance w_4 is indirect, as w_4 acts on the entry signal u to the satellite (see Figure 2).

Finally, in order to prevent unrealistic controller gains, the controller output u is given an additional cost by introducing a performance variable $z_4 = W_u u$. The weighting associated with this specification has been set to $W_u = 0.1$. We stress that choosing the weighting matrices reflects practical engineering specifications and is not a trivial task, involving both engineering insight and trial and error at this stage.

Note that two controllers were designed in this example. A nominal H_∞ -controller was synthesized (corresponding to $\delta = 0$ above). A second robust H_∞ -controller was synthesized using our augmented Lagrangian algorithm which explicitly accounts for the uncertainty in the flexible dynamics. Simulations were then performed with a torque step disturbance of 1 in-lbf at the plant input. The results are compared in Figure 3 for a set of values of δ . We observe that as expected the nominal controller (left-column simulations) is satisfactory in the nominal case ($\delta = 0$) but exhibits significant loss of performance and even of stability in the nonnominal situations $\delta = \pm 1$. This is in strong contrast with the robust controller obtained by our method, which meets the prescribed performance requirements despite the uncertainties in the flexible dynamics (right-column simulations).

REFERENCES

- [1] B. ANDERSON AND S. VONGPANITLERD, *Network Analysis and Synthesis: A Modern System Theory Approach*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
- [2] P. APKARIAN AND P. GAHINET, *A convex characterization of gain-scheduled $\mathcal{H}_\infty$ controllers*, IEEE Trans. Automat. Control, 40 (1995), pp. 853–864.
- [3] P. APKARIAN, P. GAHINET, AND G. BECKER, *Self-scheduled H_∞ control of linear parameter-varying systems: A design example*, Automatica J. IFAC, 31 (1995), pp. 1251–1261.
- [4] P. APKARIAN, D. NOLL, AND H. D. TUAN, *A Prototype Primal-Dual LMI-Interior Algorithm for Nonconvex Robust Control Problems*, Rapport Interne 01-08, MIP, UMR 5640, Toulouse, France, 2001. Available online at <http://mip.ups-tlse.fr/publi/publi.html>.
- [5] P. APKARIAN, D. NOLL, AND H. D. TUAN, *Fixed-order $\mathcal{H}_\infty$ control design via an augmented Lagrangian method*, Internat. J. Robust Nonlinear Control, 13 (2003), pp. 1137–1148.
- [6] P. APKARIAN AND H. D. TUAN, *Concave programming in control theory*, J. Global Optim., 15 (1999), pp. 343–370.
- [7] V. I. ARNOLD, *On matrices depending on parameters*, Russian Math. Surveys, 26 (1971), pp. 29–43.
- [8] G. J. BALAS, J. C. DOYLE, K. GLOVER, A. PACKARD, AND R. SMITH, *μ -Analysis and Synthesis Toolbox: User's Guide*, The MathWorks, Inc., Natick, MA, 1991.
- [9] K. P. BENNETT AND O. L. MANGASARIAN, *Bilinear separation of two sets in n -space*, Comput. Optim. Appl., 2 (1993), pp. 207–227.
- [10] A. BEN-TAL AND A. NEMIROVSKI, *Lectures on Modern Convex Optimization*, MPS/SIAM Ser. Optim., SIAM, Philadelphia, 2001.
- [11] A. BEN-TAL, *Conic and Robust Optimization*, Lecture Notes, Università di Roma La Sapienza, Rome, Italy, 2002.
- [12] S. BOYD, L. EL GHAOU, E. FERON, AND V. BALAKRISHNAN, *Linear Matrix Inequalities in System and Control Theory*, SIAM Stud. Appl. Math. 15, SIAM, Philadelphia, 1994.
- [13] J. F. BONNANS AND A. SHAPIRO, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, New York, 2000.
- [14] H. BUSCHEK AND H. A. J. CALISE, *Mu controllers: Mixed and fixed*, AIAA J. Guidance Control Dyn., 20 (1995), pp. 43–41.
- [15] J. BURKE, A. S. LEWIS, AND M. L. OVERTON, *Optimizing matrix stability*, Proc. Amer. Math. Soc., 129 (2000), pp. 1635–1642.
- [16] J. BURKE, A. S. LEWIS, AND M. L. OVERTON, *Two numerical methods for optimizing matrix stability*, Linear Algebra Appl., 351/352 (2002), pp. 117–145.
- [17] H. A. J. CALISE AND E. V. BYRN, *Parameter sensitivity reduction in fixed-order dynamic compensation*, AIAA J. Guidance Control Dyn., 15 (1992), pp. 440–447.
- [18] R. Y. CHIANG AND M. G. SAFONOV, *Real K_m -synthesis via generalized Popov multipliers*, in Proceedings of the American Control Conference, Chicago, IL, 1992, pp. 2417–2418.
- [19] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, SIAM Class. Appl. Math. 5, SIAM, Philadelphia, 1990.
- [20] A. R. CONN, N. I. M. GOULD, AND PH. L. TOINT, *A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds*, SIAM J. Numer. Anal., 28 (1991), pp. 545–572.
- [21] A. R. CONN, N. I. M. GOULD, A. SARTENAER, AND PH. L. TOINT, *Global Convergence of Two Augmented Lagrangian Algorithms for Optimization with a Combination of General Equality and Linear Constraints*, Technical report TR/PA/93/26, Toulouse, France, 1993.
- [22] A. R. CONN, N. I. M. GOULD, A. SARTENAER, AND PH. L. TOINT, *Local Convergence Properties of Two Augmented Lagrangian Algorithms for Optimization with a Combination of General Equality and Linear Constraints*, Technical report TR/PA/93/27, Toulouse, France, 1993.
- [23] A. R. CONN, N. I. M. GOULD, A. SARTENAER, AND PH. L. TOINT, *Convergence properties of an augmented Lagrangian algorithm for optimization with a combination of general equality and linear constraints*, SIAM J. Optim., 6 (1996), pp. 674–703.
- [24] J. C. DOYLE, A. PACKARD, AND K. ZHOU, *Review of LFT's, LMI's and μ* , in Proceedings of the IEEE Conference on Decision and Control, Vol. 2, Brighton, UK, 1991, pp. 1227–1232.
- [25] G. E. DULLERUD AND F. PAGANINI, *A Course in Robust Control Theory: A Convex Approach*, Springer Texts Appl. Math. 36, Springer-Verlag, New York, 2000.
- [26] L. EL GHAOU, F. OUSTRY, AND M. AIT RAMI, *An algorithm for static output-feedback and related problems*, IEEE Trans. Automat. Control, 42 (1997), pp. 1171–1176.

- [27] M. K. H. FAN, A. L. TITS, AND J. C. DOYLE, *Robustness in the presence of mixed parametric uncertainty and unmodeled dynamics*, IEEE Trans. Automat. Control, 36 (1991), pp. 25–38.
- [28] B. FARES, P. APKARIAN, AND D. NOLL, *An augmented Lagrangian method for a class of LMI constrained problems in robust control theory*, Internat. J. Control, 74 (2001), pp. 348–360.
- [29] B. FARES, D. NOLL, AND P. APKARIAN, *Robust control via sequential semidefinite programming*, SIAM J. Control Optim., 40 (2002), pp. 1791–1820.
- [30] E. FERON, P. APKARIAN, AND P. GAHINET, *Analysis and synthesis of robust control systems via parameter-dependent Lyapunov functions*, IEEE Trans. Automat. Control, 41 (1996), pp. 1041–1046.
- [31] P. GAHINET AND P. APKARIAN, *A linear matrix inequality approach to H_∞ control*, Internat. J. Nonlinear Control, 4 (1994), pp. 421–448.
- [32] M. X. GOEMANS AND D. P. WILLIAMSON, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, J. ACM, 42 (1995), pp. 1115–1145.
- [33] A. HELMERSSON, *Methods for Robust Gain-Scheduling*, Ph.D. thesis, Linköping University, Linköping, Sweden, 1995.
- [34] M. R. HESTENES, *Multiplier and gradient methods*, J. Optim. Theory Appl., 4 (1969), pp. 303–320.
- [35] T. IWASAKI AND R. E. SKELTON, *All controllers for the general $\mathcal{H}_\infty$ control problem: LMI existence conditions and state space formulas*, Automatica J. IFAC, 30 (1994), pp. 1307–1317.
- [36] J. LASALLE AND S. LEFSCHETZ, *Stability by Lyapunov's Direct Method*, Academic Press, New York, 1961.
- [37] F. LEIBFRTZ AND E. M. E. MOSTAFA, *Trust region methods for solving the optimal output feedback design problem*, Int. J. Control, 76 (2003), pp. 501–519.
- [38] F. LEIBFRTZ AND E. M. E. MOSTAFA, *An interior point constrained trust region method for a special class of nonlinear semidefinite programming problems*, SIAM J. Optim., 12 (2002), pp. 1048–1074.
- [39] O. L. MANGASARIAN AND J. S. PANG, *The extended linear complementarity problem*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 359–368.
- [40] A. MEGRETZKI AND A. RANTZER, *System analysis via integral quadratic constraints*, IEEE Trans. Automat. Control, 42 (1997), pp. 819–830.
- [41] M. MESBASHI, *A semi-definite programming solution of the least order dynamic output feedback synthesis problem*, in Proceedings of the IEEE Conference on Decision and Control, Phoenix, AZ, 1999.
- [42] A. NEMIROVSKII, *Several NP-hard problems arising in robust stability analysis*, Math. Control Signals Systems, 6 (1994), pp. 99–105.
- [43] M. C. DE OLIVEIRA AND J. C. GEROMEL, *Numerical comparison of output feedback design methods*, in Proceedings of the American Control Conference, Albuquerque, NM, 1997, pp. 72–76.
- [44] A. PACKARD, *Gain scheduling via linear fractional transformations*, Systems Control Lett., 22 (1994), pp. 79–92.
- [45] A. PACKARD AND G. BECKER, *Quadratic stabilization of parametrically-dependent linear systems using parametrically-dependent linear dynamic feedback*, Adv. Robust Nonlinear Control Syst., 43 (1992), pp. 29–36.
- [46] A. PACKARD, K. ZHOU, P. PANDEY, AND G. BECKER, *A collection of robust control problems leading to LMI's*, in Proceedings of the IEEE Conference on Decision and Control, Brighton, UK, 1991, pp. 1245–1250.
- [47] M. J. D. POWELL, *A method for nonlinear constraints in minimization problems*, in Optimization, R. Fletcher, ed., Academic Press, London, New York, 1969.
- [48] M. G. SAFONOV AND G. P. PAPAVASSILOPOULOS, *The diameter of an intersection of ellipsoids and BMI robust synthesis*, in Proceedings of the IFAC Symposium on Robust Control Design, Rio de Janeiro, Brazil, 1994.
- [49] A. SARTENAER, *Développement d'algorithmes en programmation non-linéaire: de l'étude théorique à l'expérimentation numérique*, Habilitation à diriger des recherches, Université Paul Sabatier, Toulouse, France, 1999.
- [50] C. W. SCHERER, *Robust mixed control and linear parameter-varying control with full block scaling*, in Advances in Linear Matrix Inequality Methods in Control, Adv. Des. Control 2, L. El Ghaoui and S.-I. Niculescu, eds., SIAM, Philadelphia, 1999, pp. 187–207.

- [51] G. SCORLETTI AND L. EL GHAOU, *Improved linear matrix inequality conditions for gain-scheduling*, in Proceedings of the IEEE Conference on Decision and Control, New Orleans, 1995, pp. 3626–3631.
- [52] A. SHAPIRO, *First and second order analysis of nonlinear semidefinite programs*, Math. Program., 77 (1997), pp. 301–320.
- [53] M. TORKI, *First- and second-order epi-differentiability in eigenvalue optimization*, J. Math. Anal. Appl., 234 (1999), pp. 391–416.
- [54] K. ZHOU, J. C. DOYLE, AND K. GLOVER, *Robust and Optimal Control*, Prentice–Hall, Upper Saddle River, NJ, 1996.

PROBABILISTIC COMBINATORIAL OPTIMIZATION: MOMENTS, SEMIDEFINITE PROGRAMMING, AND ASYMPTOTIC BOUNDS*

DIMITRIS BERTSIMAS[†], KARTHIK NATARAJAN[‡], AND CHUNG-PIAW TEO[§]

Abstract. We address the problem of evaluating the expected optimal objective value of a 0-1 optimization problem under uncertainty in the objective coefficients. The probabilistic model we consider prescribes limited marginal distribution information for the objective coefficients in the form of moments. We show that for a fairly general class of marginal information, a tight upper (lower) bound on the expected optimal objective value of a 0-1 maximization (minimization) problem can be computed in polynomial time if the corresponding deterministic problem is solvable in polynomial time. We provide an efficiently solvable semidefinite programming formulation to compute this tight bound. We also analyze the asymptotic behavior of a general class of combinatorial problems that includes the linear assignment, spanning tree, and traveling salesman problems, under knowledge of complete marginal distributions, with and without independence. We calculate the limiting constants exactly.

Key words. combinatorial optimization, probabilistic analysis, convex optimization, moments problem

AMS subject classifications. 90C22, 90C27

DOI. 10.1137/S1052623403430610

1. Introduction. We analyze the optimal objective value of the generic 0-1 optimization problem with random objective coefficients. Let $\mathbf{X} = \{1, \dots, n\}$ and P be a nonnegative integer that denotes the number of feasible solutions to the combinatorial optimization problem. Let $\{B_p, p = 1, \dots, P\}$ be nonempty subsets of $\mathbf{X}$ that correspond to the set of feasible solutions. Without loss of generality, we assume that the problem does not contain redundant variables, namely, for each $i \in \mathbf{X}$, there exists at least one feasible solution that contains i and at least one feasible solution that does not contain i . The nominal 0-1 optimization problem in maximization form is

$$(1.1) \quad Z_{\max}^*(\mathbf{c}) = \max_{p=1, \dots, P} \sum_{i \in B_p} c_i.$$

Traditional models solve the nominal 0-1 optimization problem under the assumption that the objective vector $\mathbf{c}$ is completely deterministic. However, such models ignore the inherent data uncertainty that may affect the quality of the output solution. To overcome this shortcoming, an increasing research effort has focused on addressing uncertainty in the objective function. The standard approach to tackle such uncertainty is to assume an underlying distribution for the objective vector $\mathbf{c}$. Let θ represent this prescribed multivariate distribution. For example, when θ is specified to be uniform on $[0, 1]$ and i.i.d. for each coefficient, it reduces to the classical

*Received by the editors June 24, 2003; accepted for publication (in revised form) March 25, 2004; published electronically November 17, 2004. This research was partially supported by the Singapore-MIT Alliance.

<http://www.siam.org/journals/siopt/15-1/43061.html>

[†]Boeing Professor of Operations Research, Sloan School of Management and Operations Research Center, Massachusetts Institute of Technology, E53-363, Cambridge, MA 02139 (dbertsim@mit.edu).

[‡]High Performance Computation for Engineered Systems, Singapore-MIT Alliance, Singapore 119260 (karthik.natarajan@yahoo.com).

[§]Department of Decision Sciences, NUS Business School, National University of Singapore, Singapore 117591 (bizteocp@nus.edu.sg).

model studied in probabilistic combinatorial optimization [28]. Examples of problems that have been studied under this model are the linear assignment [1], quadratic assignment [6], and spanning tree problems [10]. The primary focus of such techniques is to estimate the expected optimal objective value:

$$(1.2) \quad E_{\theta}[Z_{\max}^*(\mathbf{c})].$$

Evaluating this expectation exactly is nontrivial. Moreover, computing bounds on this expected optimal value is already quite challenging.

In this paper, we study this problem under a more general setting. Specifically, we consider a class of feasible distributions Θ , and our objective is to find a tight upper bound on the expected optimal objective value of the nominal 0-1 maximization problem:

$$(1.3) \quad \begin{aligned} Z_{\max}^* = \sup_{\theta} \quad & E_{\theta}[Z_{\max}^*(\mathbf{c})] \\ \text{s.t.} \quad & \theta \in \Theta. \end{aligned}$$

By a tight upper bound we refer to a valid upper bound on the expected objective value that is achieved either exactly or asymptotically by a set of feasible distributions. Problem (1.3) arises naturally in applications where the distribution for the objective coefficients is not completely known. As an example, consider the problem of estimating an upper bound on the expected time of completion of a project that consists of activities with precedence relationships for which we have information on the expected duration and variability of individual activities but no information on their correlation (see also section 4). Specifying the exact distribution in such problems is a potentially difficult task. Most traditional approaches make the additional assumption of independence and try to estimate the expected optimal value. We, however, drop this assumption of independence completely and provide insights into the performance of the problem under dependence. Obtaining good lower bounds on $E_{\theta}[Z_{\max}^*(\mathbf{c})]$ is relatively easier by using Jensen's inequality. Our results can be easily extended to the analogous problem of finding a tight lower bound $Z_{\min}^*$ on the expected optimal objective value of the 0-1 minimization problem by setting $\mathbf{c} := -\mathbf{c}$.

1.1. Structure of data uncertainty. The feasible set of multivariate distributions Θ we consider is characterized by known information on the marginal distributions of each objective coefficient. Let $F_i(c) = \mathcal{P}(c_i \leq c)$ denote the complete marginal distribution function of c_i . We let $\Theta(F_1, \dots, F_n)$ denote the set of multivariate distributions compatible with the marginal distributions F_i . Note that this model does not assume that objective coefficients are independently distributed, and hence Θ is not uniquely specified. Under complete knowledge of marginal distributions F_i , problem (1.3) is formulated as

$$(1.4) \quad \begin{aligned} Z_{\max}^* = \sup_{\theta} \quad & E_{\theta}[Z_{\max}^*(\mathbf{c})] \\ \text{s.t.} \quad & \theta \in \Theta(F_1, \dots, F_n). \end{aligned}$$

Motivated by an application in project management, problem (1.4) was studied by Meilijson and Nadas [21] in the context of computing the longest path in a directed acyclic graph. Their main result shows that $Z_{\max}^*$ can be computed by solving a convex minimization reformulation of this problem. Weiss [31] observed that this approach generalizes to the maximum flow and shortest route problem.

In this paper, we generalize this result to problems with limited information on the distribution F_i . We assume that the objective coefficient c_i takes value in Ω_i (i.e.,

the support of the random variable is possibly a subset of Ω_i). The distribution is also assumed to satisfy the moment equality constraints on real-valued functions of c_i in the form $E_{F_i}[f_{ik}(c_i)] = m_{ik}$, $k = 1, \dots, k_i$. In the classical case, with power functions $f_{ik}(c_i) = c_i^k$, this reduces to knowing the first k_i moments of c_i . Given this information, the feasible set of marginal distributions for coefficient c_i is represented as

$$(1.5) \quad \mathbb{F}_i = \left\{ F_i \mid E_{F_i}[f_{ik}(c_i)] = m_{ik}, \quad k = 1, \dots, k_i, \quad E_{F_i}[\mathbb{I}_{\Omega_i}] = 1 \right\},$$

where $\mathbb{I}_S(c_i) = 1$ if $c_i \in S$ and 0 otherwise. We assume that the feasible region $\mathbb{F}_i$ is nonempty and $E_{F_i}[|c_i|] < \infty$ to guarantee a finite value for $Z_{\max}^*$. Under the marginal moment information model, problem (1.3) is formulated as

$$(1.6) \quad \begin{aligned} Z_{\max}^* &= \sup_{\theta, F_1, \dots, F_n} E_{\theta}[Z_{\max}^*(\mathbf{c})] \\ \text{s.t.} \quad &F_i \in \mathbb{F}_i, \quad i = 1, \dots, n, \\ &\theta \in \Theta(F_1, \dots, F_n). \end{aligned}$$

Remark. We make no assumptions on cross moment information in Θ . As the size of the problem increases, the number of estimates needed to capture dependent structures grows exponentially in the dimension of the problem. Instead of trying to characterize this exact structure, we optimize over all dependent structures with the given marginal moment information.

1.2. Contributions. In this paper, building on the work of Bertsimas and Popescu [4] connecting moment problems and semidefinite optimization, we generalize the approach by Meilijson and Nadas [21] and develop techniques to compute $Z_{\max}^*$ and $Z_{\min}^*$ for general 0-1 optimization problems. Our main contributions are as follows:

- (a) We provide a general optimization formulation to compute $Z_{\max}^*$ under limited information on the marginal distributions of the objective coefficients. This formulation has an exponential number of constraints in general.
- (b) Given limited marginal moment information for the objective coefficients, we show that an upper bound on $Z_{\max}^*$, first proposed in [21] in the context of the longest path problem in a directed acyclic graph, is tight for general combinatorial problems. Note that the tightness of the bound even in the context of the longest path problem has not been known to hold in general; see the discussion in [5, p. 844].
- (c) For piecewise polynomial functions f_{ik} , we show that $Z_{\max}^*$ can be computed in polynomial time if the nominal 0-1 maximization problem can be solved in polynomial time. We provide an efficiently solvable semidefinite programming formulation with which to compute the $Z_{\max}^*$ in this case.
- (d) We characterize the asymptotic behavior of the proposed bounds for a general class of combinatorial problems under knowledge of complete and identical marginal distributions. No assumptions on the independence of distributions is made. For this model, we show that for the minimum cost $N \times N$ linear assignment problem, the spanning tree, and the traveling salesman problem on N vertices, with the distribution function of the cost coefficients satisfying $F_{\alpha}(c) = \rho_{\alpha} c^{\alpha}$ as $c \downarrow 0$ for $\alpha \geq 0$, the tight lower bound $Z_{\min}^*$ scales as

$$\lim_{N \rightarrow \infty} \left(\frac{Z_{\min}^*}{N^{(\alpha-1)/\alpha}} \right) = C_{\alpha}^*,$$

TABLE 1.1

Scaling constants with distribution function of form $F(c) = \rho_\alpha c^\alpha$ as $c \downarrow 0$.

0-1 Optimization problem	$\lim_{N \rightarrow \infty} \left(\frac{Z_{\min}^*}{N^{(\alpha-1)/\alpha}} \right)$
$N \times N$ Linear assignment	$\left(\frac{\alpha}{\alpha+1} \right) \rho_\alpha^{-1/\alpha}$
Spanning tree on N node complete graph	$\left(\frac{\alpha}{\alpha+1} \right) \rho_\alpha^{-1/\alpha} 2^{1/\alpha}$
Traveling salesman on N node complete graph	$\left(\frac{\alpha}{\alpha+1} \right) \rho_\alpha^{-1/\alpha} 2^{1/\alpha}$

where C_α^* is explicitly computed in Table 1.1. It is interesting that the asymptotic bounds exhibit the same scaling behavior (although with different constants) under the assumption of independence and that the limiting constants for the minimum spanning tree and the traveling salesman problems are the same.

1.3. Structure of the paper. In section 2, we provide a general duality-based approach with which to compute the tight upper bound on the expected optimal objective value of a 0-1 maximization problem under marginal moment information. Unfortunately, this formulation has exponentially many constraints in general. In section 3, we obtain our first set of results based on the reformulation in [21] for the problem. We introduce a semidefinite programming approach for solving a class of such problems. In section 4, we provide numerical comparisons with other methods for a particular application in project management. In section 5, we develop the bound under complete and identical marginal distributions and provide new asymptotic bounds for classical combinatorial problems. In section 6, we summarize our conclusions.

2. General approach. Given limited marginal information for each objective coefficient c_i , we are interested in computing the tight upper bound on the expected optimal objective value of a 0-1 maximization problem. Given the feasible set of marginal distributions in (1.5), we can rewrite problem (1.6) as

$$\begin{aligned}
 (2.1) \quad Z_{\max}^* &= \sup_{\theta} E_{\theta}[Z_{\max}^*(\mathbf{c})] \\
 \text{s.t.} \quad &E_{\theta}[f_{ik}(c_i)] = m_{ik}, \quad k = 1, \dots, k_i, \quad i = 1, \dots, n, \\
 &E_{\theta}[\mathbb{I}_{\Omega}] = 1,
 \end{aligned}$$

where Ω denotes $\Omega_1 \times \dots \times \Omega_n$. We construct the dual of this formulation by introducing a dual variable y_{ik} for each moment equality constraint and y_{00} for the probability mass constraint. Since the primal problem is bounded, the dual problem is formulated as

$$\begin{aligned}
 (2.2) \quad Z_D^* &= \min \left(y_{00} + \sum_{i=1}^n \sum_{k=1}^{k_i} y_{ik} m_{ik} \right) \\
 \text{s.t.} \quad &y_{00} + \sum_{i=1}^n \sum_{k=1}^{k_i} y_{ik} f_{ik}(c_i) \geq Z_{\max}^*(\mathbf{c}) \quad \forall \mathbf{c} \in \Omega.
 \end{aligned}$$

THEOREM 2.1 (see Isii [14]). (a) *The primal and dual formulations are related by weak duality as $Z_{\max}^* \leq Z_D^*$.*

(b) *Furthermore, if the moments lie interior to the set of feasible moment vectors for arbitrary multivariate distributions, then strong duality holds, i.e., $Z_{\max}^* = Z_D^*$.*

The strong duality condition in Theorem 2.1(b) is a Slater-type regularity condition that requires the moment vector to lie in the interior of the moment space. Under marginal moment specification, the multivariate moment space is defined as the product of univariate moment spaces. Checking for the interior point condition hence reduces to checking for the interior point condition for the univariate moment spaces. Such univariate conditions are easier to verify (as positive definiteness conditions on the moments matrix [15]). From this point on, we assume that this regularity condition is satisfied and hence the optimum objective value of problem (2.2), $Z_D^* = Z_{\max}^*$.

We now express the right-hand side of the constraint in formulation (2.2) explicitly. Since the objective function $Z_{\max}^*(\mathbf{c})$ is a piecewise linear convex function, we can rewrite the dual formulation as

$$(2.3) \quad \begin{aligned} Z_{\max}^* = \min & \left(y_{00} + \sum_{i=1}^n \sum_{k=1}^{k_i} y_{ik} m_{ik} \right) \\ \text{s.t. } & y_{00} + \sum_{i=1}^n \sum_{k=1}^{k_i} y_{ik} f_{ik}(c_i) - \sum_{i \in B_p} c_i \geq 0 \quad \forall \mathbf{c} \in \Omega, \quad p = 1, \dots, P. \end{aligned}$$

Each constraint in formulation (2.3) is equivalent to the nonnegativity of a multivariate function over a subset of $\Re^n$. For the simplest case with Ω defined by a discrete set of values, problems (2.1) and (2.3) reduce to the standard linear programming primal and dual problems. In general, there is an exponential number of such constraints due to the exponential number of feasible solutions to the nominal problem (1.1). For the simplest case with $f_{ik}(c_i) = c_i^k$ and $\Omega = \Re^n$, each constraint reduces to the nonnegativity of a multivariate polynomial over $\Re^n$. A simple sufficient condition to ensure this is a sum of squares decomposition for the polynomial. This condition can be expressed by a positive semidefiniteness constraint on matrices of increasing size in the dual variables (cf. [15] and [25]). As the order of the relaxation increases, the size of these matrices increases drastically. Moreover, tractable necessary conditions to ensure the nonnegativity of a multivariate polynomial are in general not known. These reasons seem to suggest that formulation (2.3) is difficult to solve even for simple $f_{ik}(\cdot)$, regardless of the difficulty of the nominal problem.

3. Convex reformulation. In this section, we provide a reformulation for computing $Z_{\max}^*$ that uses the marginal information structure of the objective vector based on a formulation originally proposed in [21] for a specific combinatorial problem under complete marginal distributions. Specifically, we apply the approach to general 0-1 optimization problems.

3.1. Complete marginal distribution information. We first focus on the case with completely known marginal distributions F_i for each objective coefficient, and we outline the approach of Meilijson and Nadas [21]. We denote by $x^+ = \max(0, x)$. For each feasible solution to problem (1.1) indexed by p and for an arbitrary

trary vector $\mathbf{d} \in \mathfrak{R}^n$, we have

$$\begin{aligned} \sum_{i \in B_p} c_i &= \sum_{i \in B_p} d_i + \sum_{i \in B_p} (c_i - d_i), \\ &\leq \max_{p=1, \dots, P} \sum_{i \in B_p} d_i + \sum_{i=1}^n [c_i - d_i]^+, \\ &= Z_{\max}^*(\mathbf{d}) + \sum_{i=1}^n [c_i - d_i]^+. \end{aligned}$$

Since the right-hand side of the inequality is independent of any particular feasible solution of problem (1.1), we have

$$Z_{\max}^*(\mathbf{c}) \leq Z_{\max}^*(\mathbf{d}) + \sum_{i=1}^n [c_i - d_i]^+.$$

Taking expectations with respect to F_i and the minimum over $\mathbf{d} \in \mathfrak{R}^n$ we obtain

$$E_{\theta}[Z_{\max}^*(\mathbf{c})] \leq \min_{\mathbf{d} \in \mathfrak{R}^n} \left(Z_{\max}^*(\mathbf{d}) + \sum_{i=1}^n E_{F_i}[c_i - d_i]^+ \right) \quad \forall \theta \in \Theta(F_1, \dots, F_n).$$

Furthermore, a joint probability distribution is constructed in [21] such that the upper bound with given marginal distributions F_i is tight. This brings them to their central result that solving a convex minimization problem in n variables yields the tight upper bound:

$$(3.1) \quad \sup_{\theta \in \Theta(F_1, \dots, F_n)} E_{\theta}[Z_{\max}^*(\mathbf{c})] = \min_{\mathbf{d} \in \mathfrak{R}^n} \left(Z_{\max}^*(\mathbf{d}) + \sum_{i=1}^n E_{F_i}[c_i - d_i]^+ \right).$$

3.2. Partial marginal distribution information. In this section, we consider the case where the marginal distribution function F_i is not assumed to be known exactly but to lie in the set $\mathbb{F}_i$. Optimizing over this set of feasible marginal distributions, and using (3.1), the tight upper bound $Z_{\max}^*$ is obtained by solving

$$Z_{\max}^* = \sup_{F_i \in \mathbb{F}_i, i=1, \dots, n} \min_{\mathbf{d} \in \mathfrak{R}^n} \left(Z_{\max}^*(\mathbf{d}) + \sum_{i=1}^n E_{F_i}[c_i - d_i]^+ \right).$$

By interchanging the order of sup and min, we obtain an upper bound on $Z_{\max}^*$:

$$(3.2) \quad Z_{\max}^* \leq \min_{\mathbf{d} \in \mathfrak{R}^n} \left(Z_{\max}^*(\mathbf{d}) + \sum_{i=1}^n \sup_{F_i \in \mathbb{F}_i} E_{F_i}[c_i - d_i]^+ \right).$$

Klein Haneveld [13] showed that if the optimal marginal distribution $F_i^* \in \mathbb{F}_i$ for the inner subproblem is the same for all d_i , that is,

$$(3.3) \quad \sup_{F_i \in \mathbb{F}_i} E_{F_i}[c_i - d_i]^+ = E_{F_i^*}[c_i - d_i]^+ \quad \forall d_i \in \mathfrak{R}, \quad i = 1, \dots, n,$$

then bound (3.2) is tight. We next give some examples of marginal moment information under which condition (3.3) is satisfied.

(a) Given bounded range $\Omega_i = [\underline{c}_i, \bar{c}_i]$, since the function $[c_i - d_i]^+$ is a nondecreasing function in c_i , the optimal marginal distribution F_i^* is

$$\mathcal{P}_{F_i^*}(c_i = \bar{c}_i) = 1.$$

Solving formulation (3.2) yields the tight upper bound as

$$(3.4) \quad Z_{\max}^* = Z_{\max}^*(\bar{\mathbf{c}}).$$

(b) Given bounded range $\Omega_i = [\underline{c}_i, \bar{c}_i]$ and first moment μ_i for each coefficient c_i , the optimal marginal distribution F_i^* is known to be [20]

$$\mathcal{P}_{F_i^*}(c_i = \underline{c}_i) = \frac{\bar{c}_i - \mu_i}{\bar{c}_i - \underline{c}_i}, \quad \mathcal{P}_{F_i^*}(c_i = \bar{c}_i) = \frac{\mu_i - \underline{c}_i}{\bar{c}_i - \underline{c}_i}.$$

Since this distribution is independent of d_i , solving formulation (3.2) yields the tight bound $Z_{\max}^*$. The decision vector $\mathbf{d}$ can be restricted to lie in the range $\mathbf{\Omega} = [\underline{\mathbf{c}}, \bar{\mathbf{c}}]$ without affecting the optimal value. Hence the tight upper bound is the solution to the linear program:

$$(3.5) \quad Z_{\max}^* = \min_{\mathbf{d} \in [\underline{\mathbf{c}}, \bar{\mathbf{c}}]} \left(Z_{\max}^*(\mathbf{d}) + \sum_{i=1}^n \left(\frac{\mu_i - \underline{c}_i}{\bar{c}_i - \underline{c}_i} \right) (\bar{c}_i - d_i) \right).$$

(c) Given semi-infinite range $\Omega_i = [\underline{c}_i, \infty)$ and first moment μ_i , Birge and Mad-dox [5] extended the previous result to obtain

$$(3.6) \quad Z_{\max}^* = Z_{\max}^*(\underline{\mathbf{c}}) + \sum_{i=1}^n (\mu_i - \underline{c}_i).$$

However, with additional second moment information, the optimal two-atom distribution F_i^* is dependent on the variable d_i [5]. Hence formulation (3.2) was not known to be tight for higher order moment information. However, we next show that it is in fact tight for any prescribed set of marginal moment information.

3.3. The main results. Let Z_1^* denote the optimum objective value of the right-hand side of formulation (3.2). With prescribed marginal information in (1.5), this bound is computed as

$$(3.7) \quad \begin{aligned} Z_1^* = \min & \left(Z_{\max}^*(\mathbf{d}) + \sum_{i=1}^n \sup_{F_i} E_{F_i}[c_i - d_i]^+ \right) \\ \text{s.t.} & \quad E_{F_i}[f_{ik}(c_i)] = m_{ik}, \quad k = 1, \dots, k_i, \quad i = 1, \dots, n, \\ & \quad E_{F_i}[\mathbb{I}_{\Omega_i}] = 1, \quad i = 1, \dots, n. \end{aligned}$$

Since no cross moment information is known, the i th inner subproblem is written as

$$\begin{aligned} \sup_{F_i} & \quad E_{F_i}[c_i - d_i]^+ \\ \text{s.t.} & \quad E_{F_i}[f_{ik}(c_i)] = m_{ik}, \quad k = 1, \dots, k_i, \\ & \quad E_{F_i}[\mathbb{I}_{\Omega_i}] = 1. \end{aligned}$$

To formulate the dual of the i th subproblem, we introduce variables $\tilde{y}_{i0}$ for the probability mass constraint and $\tilde{y}_{ik}$ for the moment equality constraints. The dual of the

univariate subproblem is then written as

$$\begin{aligned} & \min \left(\tilde{y}_{i0} + \sum_{k=1}^{k_i} \tilde{y}_{ik} m_{ik} \right) \\ & \text{s.t.} \quad \tilde{y}_{i0} + \sum_{k=1}^{k_i} \tilde{y}_{ik} f_{ik}(c_i) \geq c_i - d_i \quad \forall c_i \in \Omega_i, \\ & \quad \tilde{y}_{i0} + \sum_{k=1}^{k_i} \tilde{y}_{ik} f_{ik}(c_i) \geq 0 \quad \forall c_i \in \Omega_i. \end{aligned}$$

Under the strong duality assumption, we substitute the dual of each subproblem into formulation (3.7) to obtain the equivalent formulation:

$$\begin{aligned} Z_1^* &= \min \left(Z_{\max}^*(\mathbf{d}) + \sum_{i=1}^n \tilde{y}_{i0} + \sum_{i=1}^n \sum_{k=1}^{k_i} \tilde{y}_{ik} m_{ik} \right) \\ (3.8) \quad & \text{s.t.} \quad \tilde{p}_{i1}(c_i) := \tilde{y}_{i0} + d_i + \sum_{k=1}^{k_i} \tilde{y}_{ik} f_{ik}(c_i) - c_i \geq 0 \quad \forall c_i \in \Omega_i, \quad i = 1, \dots, n, \\ & \quad \tilde{p}_{i2}(c_i) := \tilde{y}_{i0} + \sum_{k=1}^{k_i} \tilde{y}_{ik} f_{ik}(c_i) \geq 0 \quad \forall c_i \in \Omega_i, \quad i = 1, \dots, n. \end{aligned}$$

We denote by $\tilde{p}_{i1}(c_i)$ and $\tilde{p}_{i2}(c_i)$ the two univariate functions of c_i that are nonnegative over Ω_i . There is a total of $2n$ such functions in formulation (3.8).

We next prove the equivalence of formulation (3.8) with the original dual (2.3), rewritten here for clarity:

$$\begin{aligned} Z_{\max}^* &= \min \left(y_{00} + \sum_{i=1}^n \sum_{k=1}^{k_i} y_{ik} m_{ik} \right) \\ (3.9) \quad & \text{s.t.} \quad y_{00} + \sum_{i=1}^n \sum_{k=1}^{k_i} y_{ik} f_{ik}(c_i) - \sum_{i \in B_p} c_i \geq 0 \quad \forall \mathbf{c} \in \mathbf{\Omega}, \quad p = 1, \dots, P. \end{aligned}$$

THEOREM 3.1. *The upper bound on the expected optimal objective value of the 0-1 maximization problem Z_1^* obtained by solving formulation (3.8) for a prescribed set of marginal moment information is tight, i.e., $Z_1^* = Z_{\max}^*$.*

Proof. Clearly, formulation (3.8) provides an upper bound on the solution obtained by solving formulation (3.9). Hence $Z_{\max}^* \leq Z_1^*$.

To show the reverse inequality, consider an optimal solution to formulation (3.9) represented by variables y_{00}^*, y_{ik}^* . We now generate a feasible solution to formulation (3.8) in the following manner. We set $\tilde{y}_{ik} = y_{ik}^*$ for all k and i with $k, i \geq 1$. Having fixed the variables $\tilde{y}_{ik}$, we solve problem (3.8) to optimality for the remaining variables. Let $\tilde{y}_{i0}^*$ and d_i^* be the corresponding optimal values for the remaining variables.

We first prove a minimality property of the nonnegative univariate functions in formulation (3.8) that will be used later in the proof. We show that we can find an optimal solution to this problem such that the value $\tilde{y}_{i0}^*$ is minimal for $\tilde{p}_{i2}(c_i)$ to be nonnegative over Ω_i . To see why, suppose that there is an $\epsilon > 0$ such that we decrease $\tilde{y}_{i0}^*$ by ϵ and $\tilde{p}_{i2}(c_i)$ remains nonnegative over Ω_i . Then we can increase d_i^* by ϵ such that $\tilde{p}_{i1}(c_i)$ remains unchanged. Since the objective function is an increasing function

in $\tilde{y}_{i0}$, the modification in $\tilde{y}_{i0}$ decreases the objective function by ϵ . Now, since $Z_{\max}^*(\mathbf{d}) = \max_p (\sum_{i \in B_p} d_i)$, the modification in d_i changes the objective function by at most ϵ . Hence the above modification will not increase the objective function.

Let $\tilde{y}_{i0}^*$ be chosen such that it is the minimal value for $\tilde{p}_{i2}(c_i)$ to be nonnegative over the specified support. Similarly, having fixed $\tilde{y}_{i0}^*$, we can decrease d_i^* as long as $\tilde{p}_{i1}(c_i)$ is nonnegative over Ω_i since the objective is an increasing function in d_i . Hence we can restrict our attention to optimal values of $\tilde{y}_{i0}^*$ and d_i^* that are minimal for the nonnegative functions $\tilde{p}_{i1}(c_i)$ and $\tilde{p}_{i2}(c_i)$ over Ω_i .

In general, given n such univariate functions $\tilde{p}_i(c_i) = \sum_{k=1}^{k_i} a_{ik} f_{ik}(c_i) + a_{i0}$ with a_{i0} at the minimal value for $\tilde{p}_i(c_i)$ to be nonnegative over Ω_i , the minimal value of a_{00} for the multivariate function $\tilde{p}(\mathbf{c}) = \sum_{i=1}^n \sum_{k=1}^{k_i} a_{ik} f_{ik}(c_i) + a_{00}$ to be nonnegative over Ω is $\sum_{i=1}^n a_{i0}$. To see this, we start with

$$\tilde{p}_i(c_i) = \sum_{k=1}^{k_i} a_{ik} f_{ik}(c_i) + a_{i0} \geq 0 \quad \forall c_i \in \Omega_i, \quad i = 1, \dots, n,$$

and add the n functions to obtain

$$\sum_{i=1}^n \sum_{k=1}^{k_i} a_{ik} f_{ik}(c_i) + \sum_{i=1}^n a_{i0} \geq 0 \quad \forall \mathbf{c} \in \Omega.$$

From the minimality of a_{i0} for the univariate function $\tilde{p}_i(c_i)$, we know that there exists an $\tilde{c}_i \in \Omega_i$ such that

$$\tilde{p}_i(\tilde{c}_i) = \sum_{k=1}^{k_i} a_{ik} f_{ik}(\tilde{c}_i) + a_{i0} = 0, \quad i = 1, \dots, n.$$

Adding these equalities, we obtain a vector $\tilde{\mathbf{c}} = (\tilde{c}_1, \dots, \tilde{c}_n)$ that lies in Ω , which satisfies

$$\sum_{i=1}^n \sum_{k=1}^{k_i} a_{ik} f_{ik}(\tilde{c}_i) + \sum_{i=1}^n a_{i0} = 0.$$

Thus, the minimal value for a_{00} such that the function $\tilde{p}(\mathbf{c})$ is nonnegative over Ω is $a_{00} = \sum_{i=1}^n a_{i0}$.

Now consider any feasible solution to problem (1.1) indexed by p . Then, from formulation (3.8) we obtain

$$\begin{aligned} \forall i \text{ such that } i \in B_p, \quad \tilde{y}_{i0}^* + d_i^* + \sum_{k=1}^{k_i} y_{ik}^* f_{ik}(c_i) - c_i &\geq 0 \quad \forall c_i \in \Omega_i, \\ \forall i \text{ such that } i \notin B_p, \quad \tilde{y}_{i0}^* + \sum_{k=1}^{k_i} y_{ik}^* f_{ik}(c_i) &\geq 0 \quad \forall c_i \in \Omega_i. \end{aligned}$$

Summing up these n constraints, we obtain

$$\sum_{i=1}^n \tilde{y}_{i0}^* + \sum_{i \in B_p} d_i^* + \sum_{i=1}^n \sum_{k=1}^{k_i} y_{ik}^* f_{ik}(c_i) - \sum_{i \in B_p} c_i \geq 0 \quad \forall \mathbf{c} \in \Omega, \quad p = 1, \dots, P.$$

From the minimality of the univariate functions, we know that $\sum_{i=1}^n \tilde{y}_{i0}^* + \sum_{i \in B_p} d_i^*$ is the minimal value for this function to be nonnegative over Ω . Comparing this with the multivariate function in formulation (3.9), we obtain

$$y_{00}^* \geq \sum_{i=1}^n \tilde{y}_{i0}^* + \sum_{i \in B_p} d_i^*, \quad p = 1, \dots, P,$$

which reduces to

$$(3.10) \quad y_{00}^* \geq \sum_{i=1}^n \tilde{y}_{i0}^* + Z_{\max}^*(\mathbf{d}^*).$$

From formulation (3.8) we obtain

$$\begin{aligned} Z_1^* &\leq Z_{\max}^*(\mathbf{d}^*) + \sum_{i=1}^n \tilde{y}_{i0}^* + \sum_{i=1}^n \sum_{k=1}^{k_i} y_{ik}^* m_{ik} \\ &\leq y_{00}^* + \sum_{i=1}^n \sum_{k=1}^{k_i} y_{ik}^* m_{ik} \quad (\text{from (3.10)}), \\ &= Z_{\max}^*. \quad \square \end{aligned}$$

By setting $\mathbf{c} := -\mathbf{c}$ and $\mathbf{d} := -\mathbf{d}$, an equivalent result is obtained for $Z_{\min}^*$.

COROLLARY 3.2. *The tight lower bound on the expected optimal objective value of a 0-1 minimization problem for a prescribed set of marginal moment information is obtained by solving*

$$(3.11) \quad Z_{\min}^* = \max_{\mathbf{d} \in \mathbb{R}^n} \left(Z_{\min}^*(\mathbf{d}) + \sum_{i=1}^n \inf_{F_i \in \mathbb{F}_i} E_{F_i}[\min(0, c_i - d_i)] \right).$$

3.4. Solution techniques. Since Theorem 3.1 implies that $Z_{\max}^* = Z_1^*$, we focus on solving formulation (3.8). We show that for a fairly general class of piecewise polynomial functions f_{ik} , this formulation can be solved as a semidefinite program. Some examples of such functions are $(c_i - a)^+$ or $\mathbb{I}_{(-\infty, a]}$. The power moment functions $f_{ik}(c_i) = c_i^k$ are included in this class.

For each coefficient c_i , the constraints in formulation (3.8) imply the nonnegativity of two univariate functions, $\tilde{p}_{i1}(c_i)$ and $\tilde{p}_{i2}(c_i)$ over Ω_i . For piecewise polynomial functions f_{ik} , we can decompose the support set Ω_i to intervals Ω_{ij} , $j = 1, \dots, l_i$, with $\cup_j \Omega_{ij} = \Omega_i$ such that $\tilde{p}_{i1}(c_i)$ and $\tilde{p}_{i2}(c_i)$ are polynomials over each of these intervals. We can now rewrite formulation (3.8) as

$$(3.12) \quad \begin{aligned} Z_{\max}^* &= \min \left(Z_{\max}^*(\mathbf{d}) + \sum_{i=1}^n \tilde{y}_{i0} + \sum_{i=1}^n \sum_{k=1}^{k_i} \tilde{y}_{ik} m_{ik} \right) \\ \text{s.t.} \quad &\tilde{p}_{i1}(c_i) \geq 0 \quad \forall c_i \in \Omega_{ij}, \quad j = 1, \dots, l_i, \quad i = 1, \dots, n, \\ &\tilde{p}_{i2}(c_i) \geq 0 \quad \forall c_i \in \Omega_{ij}, \quad j = 1, \dots, l_i, \quad i = 1, \dots, n, \end{aligned}$$

where the constraints now correspond to the nonnegativity of univariate polynomials over Ω_{ij} . If Ω_{ij} is a finite set of atoms, this reduces to linear constraints.

The key observation is that although it is difficult to express exactly the nonnegativity of a multivariate polynomial, we can use positive semidefinite constraints to express the nonnegativity of a univariate polynomial over an interval. This simplification

arises due to the equivalence of the sum of squares representation and nonnegativity of a polynomial in the univariate case; see [23] and [4]. We focus on three specific intervals, namely, the entire real line $\Omega = (-\infty, \infty)$, the positive ray $\Omega = [0, \infty)$, and the segment $\Omega = [0, 1]$. The semidefinite representation for all other intervals of $\Re$ can be obtained from simple affine transformations of these three cases.

PROPOSITION 3.3 (see Nesterov [23] and Bertsimas and Popescu [4]). (a) *A univariate polynomial $f(c) = \sum_{r=0}^{2k} a_r c^r$ is nonnegative over the interval $\Omega = (-\infty, \infty)$ if and only if there exists a positive semidefinite matrix $\mathbf{Y} = [Y_{ij}]_{i,j=0,\dots,k}$ such that*

$$a_r = \sum_{i,j: i+j=r} Y_{ij}, \quad r = 0, \dots, 2k,$$

$$\mathbf{Y} \succeq \mathbf{0}.$$

(b) *A univariate polynomial $f(c) = \sum_{r=0}^k a_r c^r$ is nonnegative over the interval $\Omega = [0, \infty)$ if and only if there exists a positive semidefinite matrix $\mathbf{Y} = [Y_{ij}]_{i,j=0,\dots,k}$ such that*

$$0 = \sum_{i,j: i+j=2r-1} Y_{ij}, \quad r = 1, \dots, k,$$

$$a_r = \sum_{i,j: i+j=2r} Y_{ij}, \quad r = 0, \dots, k,$$

$$\mathbf{Y} \succeq \mathbf{0}.$$

(c) *A univariate polynomial $f(c) = \sum_{r=0}^k a_r c^r$ is nonnegative over the interval $\Omega = [0, 1]$ if and only if there exists a positive semidefinite matrix $\mathbf{Y} = [Y_{ij}]_{i,j=0,\dots,k}$ such that*

$$0 = \sum_{i,j: i+j=2r-1} Y_{ij}, \quad r = 1, \dots, k,$$

$$\sum_{l=0}^r \binom{k-l}{r-l} a_l = \sum_{i,j: i+j=2r} Y_{ij}, \quad r = 0, \dots, k,$$

$$\mathbf{Y} \succeq \mathbf{0}.$$

Clearly, from Proposition 3.3, we can replace the condition that a univariate polynomial is nonnegative over an interval of $\Re$ by a requirement on a matrix to be positive semidefinite and a set of linear equalities that must be satisfied. The specific form depends on the nature of the interval Ω . For example, from Proposition 3.3(a), for a univariate polynomial to be nonnegative over $\Re$, the degree of the polynomial must be even.

From Proposition 3.3, we can rewrite formulation (3.12) as

$$(3.13) \quad \begin{aligned} Z_{\max}^* = \min & \left(Z_{\max}^*(\mathbf{d}) + \sum_{i=1}^n \tilde{y}_{i0} + \sum_{i=1}^n \sum_{k=1}^{k_i} \tilde{y}_{ik} m_{ik} \right) \\ \text{s.t. } & \mathbf{W}_{ij}(\tilde{y}_{i0}, \dots, \tilde{y}_{ik_i}, -1, d_i, \Omega_{ij}) \succeq \mathbf{0}, \quad j = 1, \dots, l_i, \quad i = 1, \dots, n, \\ & \mathbf{X}_{ij}(\tilde{y}_{i0}, \dots, \tilde{y}_{ik_i}, \Omega_{ij}) \succeq \mathbf{0}, \quad j = 1, \dots, l_i, \quad i = 1, \dots, n, \end{aligned}$$

where $\mathbf{W}_{ij}$ and $\mathbf{X}_{ij}$ denote the semidefinite constraint matrix $\mathbf{Y}$ obtained from Proposition 3.3. The entries of the semidefinite matrix depends on the nature of the interval Ω_{ij} and the coefficients of the polynomial.

THEOREM 3.4. *Given marginal distribution information as moments of piecewise polynomial functions, the tight upper (lower) bound on the expected optimal objective value of a 0-1 maximization (minimization) problem can be computed in polynomial time if the nominal problem can be solved in polynomial time.*

Proof. Formulation (3.13) can be reformulated as

$$(3.14) \quad \begin{aligned} Z_{\max}^* = \min & \left(t + \sum_{i=1}^n \tilde{y}_{i0} + \sum_{i=1}^n \sum_{k=1}^{k_i} \tilde{y}_{ik} m_{ik} \right) \\ \text{s.t.} \quad & \mathbf{W}_{ij}(\tilde{y}_{i0}, \dots, \tilde{y}_{ik_i}, -1, d_i, \Omega_{ij}) \succeq \mathbf{0}, \quad j = 1, \dots, l_i, \quad i = 1, \dots, n, \\ & \mathbf{X}_{ij}(\tilde{y}_{i0}, \dots, \tilde{y}_{ik_i}, \Omega_{ij}) \succeq \mathbf{0}, \quad j = 1, \dots, l_i, \quad i = 1, \dots, n, \\ & t \geq Z_{\max}^*(\mathbf{d}), \end{aligned}$$

where t is an additional variable. We have a polynomial number of polynomial-size semidefinite matrices in this formulation. One approach to solve this problem is to disaggregate the term $Z_{\max}^*(\mathbf{d})$ in terms of its feasible solutions. However, the number of linear constraints in such a formulation would be exponentially large. This implies that solving formulation (3.14) is still difficult in general. For the class of polynomially solvable 0-1 maximization problems, we can, however, solve this efficiently by considering the separation version of the problem. Since the semidefinite constraints are polynomial sized, this essentially reduces to the separation problem: Given t and $\mathbf{d}$, verify if $t \geq Z_{\max}^*(\mathbf{d})$, and if not find a violated inequality.

To solve the separation problem, we solve the nominal discrete optimization problem with objective vector $\mathbf{d}$. If this can be solved in polynomial time, then we can find the optimal decision vector indexed by $p^* \in \{1, \dots, P\}$ and $Z_{\max}^*(\mathbf{d})$ in polynomial time. Then we simply test if $t \geq Z_{\max}^*(\mathbf{d})$, and if not an index p^* is found such that $\sum_{i \in B_{p^*}} d_i > t$. The desired result follows from the equivalence of separation and optimization [11]. $\square$

Remarks. (a) Theorem 3.4 implies that the tight lower bounds for the expected optimal objective values of combinatorial problems such as the minimum spanning tree, assignment, matching, and shortest path problems are computable in polynomial time under the marginal moment model.

(b) The previously known results (3.4), (3.5), and (3.6) can be viewed as special cases of our Theorems 3.1 and 3.4.

(c) The results hold even if we only know upper bounds on moments on $E[f_{ik}(c_i)]$. The only difference is that the corresponding dual variables are nonnegative.

(d) The variables d_i in formulation (3.14) can be restricted to lie in $[\underline{c}_i, \bar{c}_i]$, where $\underline{c}_i$ and $\bar{c}_i$ define lower and upper bounds on the range of Ω_i .

The solution method described thus far is a general cutting plane algorithm that solves the nominal problem (1.1) at each step in conjunction with semidefinite constraints. We now provide a compact semidefinite reformulation for a class of 0-1 optimization problems that may be useful computationally.

THEOREM 3.5. *For 0-1 optimization problems with compact linear programming formulations, the tight upper (lower) bound on the expected optimal objective value of a 0-1 maximization (minimization) problem under the marginal moment model can be computed as a compact semidefinite program.*

Proof. Any 0-1 optimization problem can be reformulated as a linear programming

problem over the convex hull of its feasible region:

$$(3.15) \quad \begin{aligned} Z_{\max}^*(\mathbf{d}) = \max \quad & \mathbf{d}'\mathbf{x} \\ \text{s.t.} \quad & \mathbf{A}\mathbf{x} \leq \mathbf{b}, \end{aligned}$$

where $(\mathbf{A}, \mathbf{b})$ provides the convex hull representation. Using linear programming duality we have

$$(3.16) \quad \begin{aligned} Z_{\max}^*(\mathbf{d}) = \min \quad & \mathbf{p}'\mathbf{b} \\ \text{s.t.} \quad & \mathbf{p}'\mathbf{A} = \mathbf{d}', \\ & \mathbf{p} \geq \mathbf{0}. \end{aligned}$$

Substituting this dual into formulation (3.13), we obtain

$$(3.17) \quad \begin{aligned} Z_{\max}^* = \min \quad & \left(\mathbf{p}'\mathbf{b} + \sum_{i=1}^n \tilde{y}_{i0} + \sum_{i=1}^n \sum_{k=1}^{k_i} \tilde{y}_{ik} m_{ik} \right) \\ \text{s.t.} \quad & \mathbf{W}_{ij}(\tilde{y}_{i0}, \dots, \tilde{y}_{ik_i}, -1, d_i, \Omega_{ij}) \succeq \mathbf{0}, \quad j = 1, \dots, l_i, \quad i = 1, \dots, n, \\ & \mathbf{X}_{ij}(\tilde{y}_{i0}, \dots, \tilde{y}_{ik_i}, \Omega_{ij}) \succeq \mathbf{0}, \quad j = 1, \dots, l_i, \quad i = 1, \dots, n, \\ & \mathbf{p}'\mathbf{A} = \mathbf{d}', \\ & \mathbf{p} \geq \mathbf{0}. \end{aligned}$$

The convex hull $(\mathbf{A}, \mathbf{b})$ representation could be exponentially large or not known explicitly. However, for a class of problems this linear representation is compact. For such problems, formulation (3.17) is a compact semidefinite program that can be solved efficiently in polynomial time. $\square$

Examples of 0-1 optimization problems with compact linear programming representation include the assignment and network flow problems such as the shortest path problem. The longest path problem on acyclic graph is also an example of 0-1 optimization problem with compact LP formulation, although the longest path problem is NP-hard in general.

4. Application in project management. In this section, we provide an application of the techniques developed to the problem of computing the longest path in a directed acyclic graph. Such problems arise in project management and scheduling.

4.1. Description of the problem. A project is specified as a set of activities that has to be completed given certain precedence relationships. We represent a project in an activity-on-arc framework with an acyclic directed graph $G(V \cup \{s, t\}, E)$, where $|E| = n$ arcs. An arc in this graph represents an activity, and a node represents the completion of all activities leading to this node. Node s represents the start of the project while node t represents the completion of the project. The precedence constraints in the graph imply that if there exist arcs (i, j) and (j, k) in the graph, then job (i, j) must be performed before job (j, k) . Such a graph constructed is necessarily acyclic, otherwise it would lead to an inconsistent ordering of jobs. We let the nonnegative random arc lengths c_{ij} denote the time required to complete each individual activity. Given this setup, the parameter of importance is the time required to complete the project measured by the longest path from the start node s to the end node t in the graph. The longest path in the graph $Z_{\max}^*(\mathbf{c})$ is computed by

solving

$$\begin{aligned}
 (4.1) \quad Z_{\max}^*(\mathbf{c}) = \max \quad & \sum_{(i,j) \in E} c_{ij} x_{ij} \\
 \text{s.t.} \quad & \sum_{j:(i,j) \in E} x_{ij} - \sum_{j:(j,i) \in E} x_{ji} = \begin{cases} 1 & \text{if } i = s, \\ -1 & \text{if } i = t, \\ 0 & \text{if } i \in V, \end{cases} \\
 & x_{ij} \in \{0, 1\} \quad \forall (i, j) \in E.
 \end{aligned}$$

In fact, we will focus on a more general measure of project performance, called project tardiness. Assume that we are specified a due date T for the project. Project tardiness is defined as a linear cost that is incurred only if the project completion time $Z_{\max}^*(\mathbf{c})$ exceeds the due date. If the project is completed before the deadline, then no cost is incurred. Mathematically, it is expressed as a piecewise convex function of project completion time:

$$\text{Project tardiness } G(T) := [Z_{\max}^*(\mathbf{c}) - T]^+.$$

It is clear that for $T = 0$, $G(0)$ reduces to $Z_{\max}^*(\mathbf{c})$.

Computing the longest path in a general graph is NP-hard, while it is polynomially time solvable for acyclic graphs [16]. However, if the arc lengths are random variables that are distributed independently and restricted to two possible values, Hagstrom [12] has shown that the computation of the expected longest path is #P-hard. Furthermore, the expected longest path cannot be computed in time polynomial in the number of values the completion time takes unless $P = NP$. This suggests that in general it is difficult to compute the exact expected project tardiness under uncertainty in activity duration, hence motivating the interest in bounds and approximations.

Robillard and Trahan [26] studied the expected completion time $E[Z_{\max}^*(\mathbf{c})]$ under assumptions of complete knowledge of distribution of activity durations and independence. Assuming independence, but only limited moment information, Devroye [8] computed upper bounds on the first moment and second moments of the completion time. However, Levy and Wiest [18] argued that activity durations are often dependent due to resource limitations. It is, however, not practical to estimate this complete multivariate joint distribution. Hence there has been effort focused on estimating the expected case tardiness under knowledge of limited moment information of the individual activities.

One approach to this problem deals with approximating the expected project tardiness. Under specific assumptions on the moments, the central limit theorem has been used to approximate the distribution of the project completion time; see [2], [27], and [19]. Under this method, the completion times obtained from different paths in the network are assumed to obey a multivariate normal distribution. Then, evaluating the expected project tardiness involves calculating the maximum of correlated normal distributions, a generally nontrivial calculation. Furthermore, if the various activities are correlated, one needs to impose restrictions on the distribution of arc lengths for the central limit theorem to apply. We do not make any such assumptions in this paper. Finally, even if the central limit theorem can be applied, it will not be a good approximation for smaller networks.

An alternative approach to the problem focuses on computing bounds on the expected project tardiness. A simple lower bound can be computed by applying

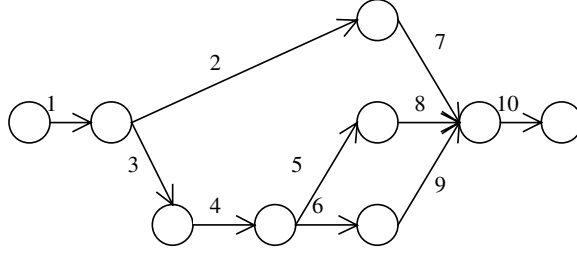


FIG. 4.1. Project network in Example 1.

Jensen's inequality to the convex tardiness function. Computing tight upper bounds is more challenging and more important as it provides an estimate of the worst case performance. Klein Haneveld [13] provided a tight upper bound on $E[G(T)]$ when the first moments of activity durations are given. With additional second moment information, Birge and Maddox [5] used the formulation in [21] to compute the upper bound on expected tardiness. However, as the computational results indicate, an approximation of the original objective function is used to compute this upper bound. The lack of precision from using such a linearization approach for solving a nonlinear problem demonstrates an additional advantage of using the semidefinite optimization approach. It should be noted that while formulation (3.13) can be directly used to compute the worst case expected project completion time, it could be easily extended to the tardiness objective. Specifically, we need to replace the term $Z_{\max}^*(\mathbf{d})$ in the objective function by $[Z_{\max}^*(\mathbf{d}) - T]^+$ and solve the corresponding semidefinite program to compute the worst case expected tardiness denoted as $G^*(T)$:

$$(4.2) \quad \begin{aligned} G^*(T) = \min \quad & \left([Z_{\max}^*(\mathbf{d}) - T]^+ + \sum_{i=1}^n \tilde{y}_{i0} + \sum_{i=1}^n \sum_{k=1}^{k_i} \tilde{y}_{ik} m_{ik} \right) \\ \text{s.t.} \quad & \mathbf{W}_{ij}(\tilde{y}_{i0}, \dots, \tilde{y}_{ik_i}, -1, d_i, \Omega_{ij}) \succeq \mathbf{0}, \quad j = 1, \dots, l_i, \quad i = 1, \dots, n, \\ & \mathbf{X}_{ij}(\tilde{y}_{i0}, \dots, \tilde{y}_{ik_i}, \Omega_{ij}) \succeq \mathbf{0}, \quad j = 1, \dots, l_i, \quad i = 1, \dots, n. \end{aligned}$$

Since $Z_{\max}^*(\mathbf{d})$ can be computed in polynomial time, formulation (4.2) is solvable in polynomial time.

4.2. Computational examples. We consider two sample projects taken from [5]. The specified data are as follows:

- (a) The minimum time required to complete each activity is known. We assume that the maximum time required to complete an activity is infinity.
- (b) For each activity, the first and second moments of the time to complete it are known. In addition for the first project, third moment information is assumed to be known. We add in this information to check the value of the extra information on the upper bound.

Under this marginal moment information, we solve formulation (4.2) with SeDuMi version 1.03 developed by Sturm [29]. An interior point method is used by the software to solve the semidefinite program. The computations were conducted on a Pentium II (550 MHz) Windows 2000 platform and the total computation time was less than a minute.

Example 1. The first project depicted in Figure 4.1 consists of 10 activities that are distributed over 3 paths. The marginal distribution information for the activities

TABLE 4.1
Activity duration data for Example 1.

Arc	Range	First mom.	Second mom.	Third mom.
1	$[1, \infty)$	1.0	1.0	1.0
2	$[4, \infty)$	3.0	9.333	30.0
3	$[3, \infty)$	2.0	4.333	10.0
4	$[3, \infty)$	2.5	6.333	16.25
5	$[7, \infty)$	5.0	26.333	146.0
6	$[5, \infty)$	4.0	16.333	68.0
7	$[5, \infty)$	3.0	10.333	40.0
8	$[5, \infty)$	4.5	20.333	92.25
9	$[2, \infty)$	1.5	2.333	3.75
10	$[6, \infty)$	5.0	25.333	130.0

TABLE 4.2
Upper bounds on expected tardiness for Example 1.

Information	Bound/ T	0	15	18.33	21.67	25
Range, first, second mom.	$G^*(T)$ reported in [5]	20.35	5.35	2.98	1.27	0.73
Range, first, second mom.	$G^*(T)$ from eq. (4.2)	20.29	5.29	2.66	0.94	0.47
Range, first, second, third mom.	$G^*(T)$ from eq. (4.2)	20.10	5.10	2.53	0.75	0.18

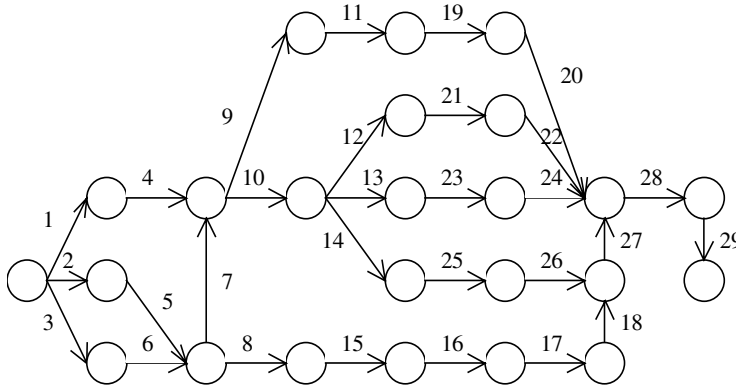


FIG. 4.2. Project network in Example 2.

is provided in Table 4.1. We consider five different deadline times T . First, we compute the worst case expected tardiness with range, first, and second moment information and compare it with the results reported in [5]. The results are provided in Table 4.2. In light of Theorem 3.1, it is clear that in [5] a heuristic method is used to calculate the bound. We next calculate the tight upper bound with additional third moment information. It is noted from Table 4.2 that the bounds are considerably tighter, especially for larger deadline dates T , indicating the value that additional information has on these estimates. It is important to note that our semidefinite programming formulation provides an easy way to incorporate third and higher order moment information, whereas previous techniques could not handle this.

Example 2. The second project is a larger one with 29 activities that are distributed over 14 paths. The activity-on-arc graph representation of the project is provided in Figure 4.2 and the data are provided in Table 4.3. The deadline time T is varied from 45.10 to 85 approximately in steps of 10, and the worst case expected

TABLE 4.3
Activity duration data for Example 2.

Arc	Range	First mom.	Second mom.	Arc	Range	First mom.	Second mom.
1	$[6, \infty)$	8.0	66.778	16	$[3, \infty)$	4.0	16.111
2	$[14, \infty)$	17.0	298.0	17	$[8, \infty)$	10.0	100.444
3	$[14, \infty)$	17.0	298.0	18	$[1, \infty)$	2.0	4.111
4	$[2, \infty)$	4.0	17.0	19	$[6, \infty)$	8.0	65.0
5	$[1, \infty)$	2.0	4.444	20	$[1, \infty)$	2.0	4.111
6	$[1, \infty)$	2.0	4.111	21	$[6, \infty)$	8.0	65.0
7	$[0.4, \infty)$	0.6	0.370	22	$[1, \infty)$	2.0	4.111
8	$[0.4, \infty)$	0.6	0.370	23	$[6, \infty)$	8.0	66.778
9	$[2, \infty)$	3.0	9.250	24	$[1, \infty)$	2.0	4.111
10	$[2, \infty)$	3.0	9.250	25	$[4, \infty)$	6.0	36.444
11	$[3, \infty)$	4.0	16.111	26	$[1, \infty)$	2.0	4.111
12	$[3, \infty)$	4.0	16.111	27	$[0.1, \infty)$	0.5	0.257
13	$[3, \infty)$	4.0	16.111	28	$[2, \infty)$	4.0	16.444
14	$[3, \infty)$	4.0	16.111	29	$[1, \infty)$	2.0	4.111
15	$[2, \infty)$	3.0	9.250				

TABLE 4.4
Upper bounds on expected tardiness for Example 2.

Information	Bound/ T	45.10	55	65	75	85
Range, first, second mom.	$G^*(T)$ reported in [5]	6.77	2.51	1.72	1.04	0.73
Range, first, second mom.	$G^*(T)$ from eq. (4.2)	6.52	2.38	1.38	0.97	0.73

project tardiness is computed for these times. The computed values are provided in Table 4.4 along with the values reported in [5]. Clearly these results indicate that the semidefinite programming method provides a general tractable approach for computing the worst case expected tardiness under moment information.

5. Asymptotic performance. In this section, we study the properties of the bounds asymptotically as the size of the problem increases to infinity. We make the following assumptions:

- (A) Each objective coefficient is a random variable with a common prespecified distribution function F . For simplicity, we assume that F is a continuous distribution. No assumption of independence is made.

The complete knowledge of F is an important one in this setting to obtain interesting scaling results. Under this setting, the tight upper bound on the expected optimal value of a 0-1 maximization problem is computed as

$$(5.1) \quad \begin{aligned} Z_{\max}^* &= \sup_{\theta} E_{\theta}[Z_{\max}^*(\mathbf{c})] \\ \text{s.t. } &\theta \in \Theta(F, \dots, F). \end{aligned}$$

We restrict ourselves to a fairly general class of a 0-1 optimization problems that satisfy the following two properties:

- (B) All feasible solutions to problem (1.1) indexed by $p \in \{1, \dots, P\}$ have the same cardinality. If we denote this common cardinality as K_1 , we have

$$K_1 = |i : i \in B_p| \quad \forall p = 1, \dots, P,$$

where $|S|$ denotes the cardinality of set S .

- (C) Every element of the ground set $\mathbf{X}$ is contained in the same number of feasible solutions. If we denote this common value as K_2 , we have

$$K_2 = |p : i \in B_p| \quad \forall i = 1, \dots, n.$$

TABLE 5.1
Parameters for three classical combinatorial optimization problems.

0-1 Optimization problem	n	K_1	P	K_2
$N \times N$ Linear assignment	N^2	N	$N!$	$(N-1)!$
Spanning tree on N node complete graph	$\frac{N(N-1)}{2}$	$N-1$	N^{N-2}	$2N^{N-3}$
Traveling salesman on N node complete graph	$\frac{N(N-1)}{2}$	N	$\frac{(N-1)!}{2}$	$(N-2)!$

Clearly $K_1 \leq n$ and $K_2 \leq P$. In conjunction, these assumptions imply that K_1 and K_2 are related by

$$K_1 P = K_2 n.$$

Three classical combinatorial problems that satisfy properties (b) and (c) are as follows:

- (a) Linear assignment problem: given a set $\{1, 2, \dots, N\}$ and an $N \times N$ objective coefficient matrix $[c_{ij}]$, find a permutation ϕ of the set that optimizes the sum $\sum_{i=1}^N c_{i\phi(i)}$.
- (b) Spanning tree problem on a complete graph: given a complete undirected graph $G(V, E)$ with N vertices and an $N \times N$ symmetric objective coefficient matrix $[c_{ij}]$, find a spanning tree with the optimal sum of the edge coefficients.
- (c) Symmetric traveling salesman problem on a complete graph: given a complete undirected graph $G(V, E)$ with N vertices and an $N \times N$ symmetric objective coefficient matrix $[c_{ij}]$, find a cyclic permutation ϕ of the set of N vertices that optimizes $\sum_{i=1}^N c_{i\phi(i)}$.

The parameters for these three problems are summarized in Table 5.1.

Remarks. (a) While the first two problems are solvable in polynomial time, the traveling salesman problem is difficult to solve. All three problems, however, share structural properties (B) and (C). The asymptotic results in this section in fact do not depend on the complexity of the nominal optimization problem.

(b) An example of a problem that does not satisfy properties (B) and (C) is the Steiner tree problem.

THEOREM 5.1. *Under assumptions (A)–(C), the tight upper bound on the expected optimal objective value of a 0-1 maximization problem is*

$$(5.2) \quad Z_{\max}^* = K_1 E_F \left[c \mid c \geq F^{-1} \left(1 - \frac{K_1}{n} \right) \right],$$

where F^{-1} is the inverse distribution function for F .

Proof. From the previous sections, we know that tight bound $Z_{\max}^*$ on the expected value of a 0-1 maximization problem given marginal distribution F is obtained by solving formulation (3.1). Problem (5.1) can then be solved as

$$(5.3) \quad Z_{\max}^* = \min_{\mathbf{d} \in \mathbb{R}^n} \left(Z_{\max}^*(\mathbf{d}) + \sum_{i=1}^n E_F [c_i - d_i]^+ \right).$$

Formulation (5.3) is a convex minimization problem in n variables. The Karush–Kuhn–Tucker conditions provide the necessary and sufficient optimality conditions for this problem. These conditions are

- (i) $\lambda_p \geq 0$ for all $p = 1, \dots, P$,

- (ii) $\sum_{p=1}^P \lambda_p = 1$,
- (iii) $Z_{\max}^*(\mathbf{d}) \geq \sum_{i \in B_p} d_i$ and $\lambda_p = 0$ whenever $Z_{\max}^*(\mathbf{d}) > \sum_{i \in B_p} d_i$ for all $p = 1, \dots, P$, and
- (iv) $1 - \sum_{p: i \in B_p} \lambda_p = F(d_i)$ for all $i = 1, \dots, n$.

We now generate a set of primal d_i and dual variables λ_p that satisfy conditions (i)–(iv).

Set the dual multipliers to $\lambda_p = 1/P, p = 1, \dots, P$. Clearly this satisfies conditions (i) and (ii). From (C), condition (iv) then reduces to

$$1 - \frac{K_2}{P} = F(d_i) \quad \forall i = 1, \dots, n.$$

If F^{-1} denotes the inverse of the distribution function, then the primal variables are set to

$$d_i = F^{-1} \left(1 - \frac{K_2}{P} \right) \quad \forall i = 1, \dots, n$$

or equivalently

$$d_i = F^{-1} \left(1 - \frac{K_1}{n} \right) \quad \forall i = 1, \dots, n.$$

Assumption (B) implies that $Z_{\max}^*(\mathbf{d}) = K_1 d$, satisfying condition (iii). Thus we have the optimal primal and dual variables to formulation (5.3). Substituting the optimal value of d_i into the objective function yields

$$\begin{aligned} Z_{\max}^* &= K_1 F^{-1} \left(1 - \frac{K_1}{n} \right) + n E_F \left[c - F^{-1} \left(1 - \frac{K_1}{n} \right) \right]^+, \\ &= K_1 F^{-1} \left(1 - \frac{K_1}{n} \right) + n E_F \left[c - F^{-1} \left(1 - \frac{K_1}{n} \right) \mid c \geq F^{-1} \left(1 - \frac{K_1}{n} \right) \right] \frac{K_1}{n}, \end{aligned}$$

which reduces to the required result. $\square$

COROLLARY 5.2. *Under assumptions (A)–(C), the tight lower bound on the expected optimal objective value of a 0-1 minimization problem is*

$$(5.4) \quad Z_{\min}^* = K_1 E_F \left[c \mid c \leq F^{-1} \left(\frac{K_1}{n} \right) \right].$$

5.1. Applications. We evaluate the explicit expected optimal objective value for a class of distributions next. Note that in the case $K_1/n \downarrow 0$ as $n \uparrow \infty$, $c \geq 0$, and from (5.4), the asymptotic performance of $Z_{\min}^*$ depends only on the distribution function of the cost coefficient $F(c)$ as $c \downarrow 0$. We thus focus on 0-1 minimization problems with the assumption that the distribution function of the cost coefficients F_α satisfies

$$(5.5) \quad F_\alpha(c) = \rho_\alpha c^\alpha \text{ as } c \downarrow 0.$$

The density function f_α is respectively assumed to satisfy

$$(5.6) \quad f_\alpha(c) = \rho_\alpha \alpha c^{\alpha-1} \text{ as } c \downarrow 0.$$

Distributions of this type have been extensively studied in the asymptotic analysis of combinatorial problems [1], [3]. Models with distributions that satisfy (5.5) include the following:

- (a) Uniformly distributed cost coefficients $U[0, 1]$ with $\alpha = 1$ and $\rho_\alpha = 1$;
- (b) Euclidean model with F_α defined as the distribution of the distances between independently and uniformly distributed points in dimension α ; here ρ_α represents the volume of a sphere of unit radius in dimension α ; and
- (c) Independent model with the cost coefficients independently distributed with distribution F .

We generalize these models by allowing the cost coefficients to be dependent. Under this distributional assumption we obtain the following result.

THEOREM 5.3. *Under assumptions (A)–(C) with cost distribution satisfying (5.5), the tight lower bound on the expected optimal objective value of a 0-1 minimization problem that satisfies $K_1 = \Theta(\sqrt{n})$ scales asymptotically to a (positive) constant C_α as*

$$(5.7) \quad \lim_{n \rightarrow \infty} \left(\frac{Z_{\min}^*}{n^{(\alpha-1)/2\alpha}} \right) = C_\alpha,$$

where the asymptotic constant C_α depends on the nominal problem.

Proof. For cost coefficients with distribution function in (5.5), we obtain

$$F_\alpha^{-1}(c) = \left(\frac{c}{\rho_\alpha} \right)^{1/\alpha} \text{ as } c \downarrow 0.$$

The tight lower bound $Z_{\min}^*$ from Corollary 5.2 is then computed as

$$\begin{aligned} Z_{\min}^* &= K_1 E_{F_\alpha} \left[c \mid c \leq \left(\frac{K_1}{\rho_\alpha n} \right)^{1/\alpha} \right], \\ &= K_1 \left[\int_0^{\left(\frac{K_1}{\rho_\alpha n} \right)^{1/\alpha}} \rho_\alpha \alpha c^{\alpha-1} dc \right] \bigg/ \left[\int_0^{\left(\frac{K_1}{\rho_\alpha n} \right)^{1/\alpha}} \rho_\alpha \alpha c^{\alpha-1} dc \right] \text{ as } n \uparrow \infty. \end{aligned}$$

Note that with $K_1 = \Theta(\sqrt{n})$, the term $(K_1/\rho_\alpha n)^{1/\alpha} \downarrow 0$ as $n \uparrow \infty$. Evaluating the integrals explicitly, we obtain

$$(5.8) \quad Z_{\min}^* = \left(\frac{\alpha}{\alpha+1} \right) K_1 \left(\frac{K_1}{n\rho_\alpha} \right)^{1/\alpha} \text{ as } n \uparrow \infty,$$

which brings us to the scaling result. $\square$

COROLLARY 5.4. *The tight lower bound on the expected optimal objective value of the linear assignment, spanning tree, and traveling salesman problem with the marginal distribution satisfying (5.5) scales asymptotically to a (positive) constant C_α^* as*

$$(5.9) \quad \lim_{N \rightarrow \infty} \left(\frac{Z_{\min}^*}{N^{(\alpha-1)/\alpha}} \right) = C_\alpha^*,$$

where the asymptotic constant C_α^* depends on the nominal problem.

Proof. Using (5.8) and $N = \Theta(\sqrt{n})$ for the three combinatorial problems of interest, we obtain the desired scaling result. $\square$

We evaluate this asymptotic behavior for three specific problems in more detail next.

TABLE 5.2
Asymptotic scaling constant for the linear assignment problem.

Parameter α	With independence: $\tilde{C}_\alpha^*$	Without independence: C_α^*
2	0.7535	0.6667
3	0.8953	0.7500
4	0.9474	0.8000
5	0.9602	0.8333
6	0.9742	0.8571

Minimum cost linear assignment problem. (a) Uniform distribution. The random minimum cost linear assignment problem has been extremely well studied under the uniform distribution model on $[0, 1]$. Under the additional assumption of independence, Walkup [30] proved an upper bound of 3 on the expected objective value. The first proven lower bound for this expected value was 1.368 [17]. These bounds have respectively been strengthened to 1.94 in [7] and 1.51 in [24]. Aldous [1] more recently proved a conjecture of Mezard and Parisi [22] that under this model the expected optimal objective value approaches $\pi^2/6 \approx 1.645$ as the size approaches ∞ . In fact, this limiting value holds for any nonnegative continuous distribution such that the density of independent costs is 1 at 0. All of the above results hold under the strict assumption of independence.

Without this assumption, by setting $\alpha = 1$, $\rho_\alpha = 1$, $K_1 = N$, and $n = N^2$ in (5.8), we obtain the tight lower bound as

$$(5.10) \quad \text{For the linear assignment problem, } Z_{\min}^* = 0.5 \quad \forall N.$$

Thus, by dropping the assumption of independence, the lower bound for the linear assignment problem becomes $C_1^* = 0.5$.

(b) Power density. The linear assignment problem has been studied under the assumption that the cost coefficients are distributed on $[0, 1]$ with density $\alpha c^{\alpha-1}$ for $\alpha > 1$. Under the assumption of independence, Donath [9] showed that asymptotically the expected optimal objective value is bounded as

$$\tilde{C}_\alpha^* \leq \lim_{N \rightarrow \infty} \left(\frac{E[Z_{\min}^*(\mathbf{c})]}{N^{(\alpha-1)/\alpha}} \right) \leq \tilde{C}_\alpha^* \left(\frac{\alpha-1}{\alpha} \right)$$

for some constant $\tilde{C}_\alpha^*$. The values of this constant for small values of α are provided by the same author.

By setting $K_1 = N$ and $n = N^2$ in (5.8), we obtain

$$(5.11) \quad \text{For the linear assignment problem, } \left(\frac{Z_{\min}^*}{N^{(\alpha-1)/\alpha}} \right) = \left(\frac{\alpha}{\alpha+1} \right) \quad \forall N.$$

Clearly, the tight lower bound scales with respect to $N^{(\alpha-1)/\alpha}$ under our model, too. This is similar to the scaling behavior observed under the independent model. In Table 5.2, we compare the constant $\tilde{C}_\alpha^*$ reported from [9] with $C_\alpha^* = \alpha/(\alpha+1)$ for five values of α . As should be expected, C_α^* is smaller than $\tilde{C}_\alpha^*$.

For these two particular distributions it can be verified that our results for the linear assignment problem hold for all N and are not just an asymptotic result. This is in contrast to the independent distribution model.

Minimum spanning tree and traveling salesman problems. (a) Uniform distribution. The random minimum spanning tree problem on a complete graph has been studied under the uniform distribution model. Under the additional assumption of independence, Frieze [10] showed that the expected length of the minimum spanning tree asymptotically converges to $\zeta(3) \approx 1.202$.

Without making the assumption of independence on arc length distribution, we obtain with $K_1 = N - 1$ and $n = N(N - 1)/2$ from (5.8):

$$Z_{\min}^* = 1 - \frac{1}{N} \quad \text{as } N \uparrow \infty.$$

As the size of the problem N approaches infinity, the tight lower bound clearly converges to $C_1^* = 1$:

$$(5.12) \quad \text{For the spanning tree problem, } \lim_{N \rightarrow \infty} Z_{\min}^* = 1.$$

This value is surprisingly close to the expected optimal objective value under independence. For the traveling salesman problem, we similarly obtain

$$Z_{\min}^* = 1 + \frac{1}{1 - N} \quad \text{as } N \uparrow \infty$$

and the same asymptotic constant of 1,

$$(5.13) \quad \text{For the traveling salesman problem, } \lim_{N \rightarrow \infty} Z_{\min}^* = 1.$$

(b) Euclidean and independent model. For the independent model with distribution function satisfying (5.5), the expected optimal objective value of the spanning tree asymptotically scales as [3]

$$\lim_{N \rightarrow \infty} \left(\frac{E[Z_{\min}^*(c)]}{N^{(\alpha-1)/\alpha}} \right) = \frac{1}{\alpha \rho_\alpha^{1/\alpha}} \sum_{k=1}^{\infty} \frac{\Gamma(k + \frac{1}{d} - 1)}{k! k^{1/d+1}}.$$

Using the same model but without assuming independence, we obtain the tight lower bound for the spanning tree problem as

$$(5.14) \quad \lim_{N \rightarrow \infty} \left(\frac{Z_{\min}^*}{N^{(\alpha-1)/\alpha}} \right) = \left(\frac{\alpha}{\alpha + 1} \right) \rho_\alpha^{-1/\alpha} 2^{1/\alpha}.$$

Under the Euclidean model we can set the value of $\rho_\alpha = \pi^{d/2}/\Gamma(d/2 + 1)$. The values of the limiting constants are provided in Table 5.3. Interestingly, the traveling salesman problem exhibits the same asymptotic scaling constant but the rates of convergence are different.

5.2. Limited marginal information. The previous section showed that without the assumption of independence, the asymptotic performance for a general class of combinatorial optimization problems exhibits interesting behavior under the model of completely known marginal distributions. We now show that under the marginal moment model such problems in fact may exhibit a different kind of scaling behavior.

Consider probabilistic combinatorial optimization problems that satisfy assumptions (B) and (C), and the following:

- (A') Each objective coefficient is a random variable with common first and second moments μ and $\mu^2 + \sigma^2$.

TABLE 5.3
Asymptotic scaling constant for the spanning tree problem.

Parameter α	With independence: $\tilde{C}_\alpha^*$	Without independence: C_α^*
2	0.5684	0.5319
3	0.6094	0.5862
4	0.6558	0.6383
5	0.7010	0.6867
6	0.7440	0.7317

- (D) The ground set can be partitioned into disjoint collection of feasible solutions, i.e., there is a set $\{0, \dots, \tilde{P} - 1\}$ such that

$$\mathbf{X} = \cup_{p=0, \dots, \tilde{P}-1} B_p, \quad B_i \cap B_j = \emptyset \quad \forall i \neq j, i, j = 0, \dots, \tilde{P} - 1.$$

Note that we change the indexing of the feasible solutions here slightly for ease of developing the result.

Assumption (D) is valid for a wide variety of combinatorial optimization problem. For instance, for the assignment problem, (D) follows from the well-known fact that the edges of a complete bipartite graph $K_{N \times N}$ can be partitioned into a collection of N disjoint perfect matching. For the spanning tree and traveling salesman problems, assumption (D) is obtained for an even and an odd number of vertices N , respectively.

Under these assumptions, we show that the tight lower bound $Z_{\min}^* = 0$ for these problems. To see this, we construct an instance of the problem, and a corresponding distribution satisfying all the conditions, such that the expected minimum cost solution is zero.

Let $\{B_0, B_1, \dots, B_{\tilde{P}-1}\}$ be the feasible solutions stipulated by assumption (D). We define

$$m = \left(\frac{\mu^2}{\mu^2 + \sigma^2} \right) \tilde{P},$$

$$S_k = \cup_{i=k}^{k+m-1} B_i, \quad k = 0, \dots, \tilde{P} - 1,$$

where the addition is taken modulo $\tilde{P}$. Consider the cost function $\mathbf{c}^k$, with

$$c_i^k = \begin{cases} \left(\frac{\mu^2 + \sigma^2}{\mu} \right) & \text{if } i \in S_k, \\ 0 & \text{otherwise.} \end{cases}$$

Note that for a particular objective function $\mathbf{c}^k$, since there is a feasible solution B_l , $l \notin S_k$, with zero optimal objective value, $\mathbf{c}_k(B_l) = 0$. Construct a probability distribution with $\mathbf{c} = \mathbf{c}_k$ with probability $1/\tilde{P}$. For this distribution it can be verified that the first two moments are

$$E[c_i] = \frac{m}{\tilde{P}} \left(\frac{\mu^2 + \sigma^2}{\mu} \right) = \mu, \quad i = 1, \dots, n,$$

$$E[c_i^2] = \frac{m}{\tilde{P}} \left(\frac{\mu^2 + \sigma^2}{\mu} \right)^2 = \mu^2 + \sigma^2, \quad i = 1, \dots, n.$$

Furthermore, for each realization of $\mathbf{c}$, the optimum solution value is 0, so we have $Z_{\min}^* = 0$.

6. Conclusions and extensions. In this paper, we addressed the problem of computing tight bounds on the expected value of a 0-1 optimization problem under uncertainty in the objective coefficients. The feasible multivariate distributions were characterized by known moment information on the marginal distributions of the objective coefficients. We showed that given moment information on piecewise polynomial functions of the objective coefficients, the tight upper (lower) bound on the expected optimal objective value of the 0-1 maximization (minimization) problem can be computed in polynomial time if the deterministic problem is solvable in polynomial time. We provided an efficiently solvable semidefinite program to compute this tight bound. Under an extension of this model with completely known and identical marginal distributions we analyzed the asymptotic bounds for a class of combinatorial problems.

REFERENCES

- [1] D. ALDOUS, *The $\zeta(2)$ limit in the random assignment problem*, Random Structures Algorithms, 18 (2001), pp. 381–418.
- [2] K. P. ANKLESARIA AND Z. DREZNER, *A multivariate approach to estimating the completion time for PERT networks*, J. Oper. Res. Soc., 40 (1986), pp. 811–815.
- [3] F. AVRAM AND D. BERTSIMAS, *The minimum spanning tree constant in geometrical probability and under the independent model: A unified approach*, Ann. Appl. Probab., 2 (1992), pp. 113–130.
- [4] D. BERTSIMAS AND I. POPESCU, *Optimal inequalities in probability: A convex optimization approach*, SIAM J. Optim., to appear.
- [5] J. R. BIRGE AND M. J. MADDOX, *Bounds on expected project tardiness*, Oper. Res., 43 (1995), pp. 838–850.
- [6] R. E. BURKARD AND U. FINCKE, *Probabilistic analysis of some combinatorial optimization problems*, Discrete Appl. Math., 12 (1985), pp. 21–29.
- [7] D. COPPERSMITH AND G. SORKIN, *Constructive bounds and exact expectations for the random assignment problem*, Random Structures Algorithms, 15 (1999), pp. 113–144.
- [8] L. P. DEVROYE, *Inequalities for the completion times of stochastic PERT networks*, Math. Oper. Res., 4 (1979), pp. 441–447.
- [9] W. E. DONATH, *Algorithms and average-value bounds for assignment problems*, IBM J. Res. Dev., 13 (1969), pp. 380–386.
- [10] A. M. FRIEZE, *On the value of a minimum spanning tree problem*, Discrete Appl. Math., 10 (1985), pp. 47–56.
- [11] M. GROTSCHTEL, L. LOVASZ, AND A. SCHRIJVER, *Geometric Algorithms and Combinatorial Optimization*, Springer, Berlin, Heidelberg, 1988.
- [12] J. H. HAGSTROM, *Computational complexity of PERT problems*, Networks, 18 (1988), pp. 139–147.
- [13] W. K. KLEIN HANEVELD, *Robustness against dependence in PERT: An application of duality and distributions with known marginals*, Math. Programming Stud., 27 (1986), pp. 153–182.
- [14] K. ISII, *On the sharpness of Chebyshev-type inequalities*, Ann. Inst. Statist. Math., 14 (1963), pp. 185–197.
- [15] J. B. LASSERRE, *Global optimization with polynomials and the problem of moments*, SIAM J. Optim., 11 (2001), pp. 796–817.
- [16] E. L. LAWLER, *Combinatorial Optimization: Networks and Matroids*, Holt, Rinehart, and Winston, New York, 1976.
- [17] A. J. LAZARUS, *Certain expected values in the random assignment problem*, Oper. Res. Lett., 14 (1993), pp. 207–214.
- [18] F. K. LEVY AND J. D. WIEST, *A Management Guide to PERT/CPM: With GERT/PDM/DPCM and Other Networks*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1977.
- [19] A. LUDWIG, R. H. MOHRING, AND F. STORK, *A computational study on bounding the makespan distribution in stochastic project networks*, Ann. Oper. Res., 102 (2001), pp. 49–64.
- [20] A. MADANSKY, *Inequalities for stochastic linear programming problems*, Management Sci., 6 (1960), pp. 197–204.
- [21] I. MEILIJSON AND A. NADAS, *Convex majorization with an application to the length of critical*

- path*, J. Appl. Probab., 16 (1979), pp. 671–677.
- [22] M. MEZARD AND G. PARISI, *On the solution of the random link matching problems*, J. Phys., 48 (1987), pp. 1451–1459.
 - [23] Y. NESTEROV, *Squared function systems and optimization problems*, in High Performance Optimization, H. Frenk et al., eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000, pp. 405–440.
 - [24] B. OLIN, *Asymptotic Properties of Random Assignment Problems*, Ph.D. thesis, Division of Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology, Stockholm, Sweden, 1992.
 - [25] P. A. PARILLO, *Structured Semidefinite Programs and Semi-algebraic Geometry Methods in Robustness and Optimization*, Ph.D. thesis, California Institute of Technology, 2000.
 - [26] P. ROBILLARD AND M. TRAHAN, *The completion times of PERT networks*, Oper. Res., 25 (1976), pp. 15–29.
 - [27] D. SCULLI AND Y. W. SHUM, *An approximate solution to the PERT problem*, Comput. Math. Appl., 21 (1991), pp. 1–7.
 - [28] J. M. STEELE, *Probability Theory and Combinatorial Optimization*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 69, SIAM, Philadelphia, 1997.
 - [29] J. F. STURM, *SeDuMi v. 1.03*. Available at <http://fewcal.kub.nl/sturm/software/sedumi.html>.
 - [30] D. W. WALKUP, *On the expected value of a random assignment problem*, SIAM J. Comput., 8 (1979), pp. 440–442.
 - [31] G. WEISS, *Stochastic bounds on distributions of optimal value functions with applications to PERT, network flows and reliability*, Oper. Res., 34 (1986), pp. 595–605.

NEWTON-TYPE METHODS FOR OPTIMIZATION PROBLEMS WITHOUT CONSTRAINT QUALIFICATIONS*

A. F. IZMAILOV[†] AND M. V. SOLODOV[‡]

Abstract. We consider equality-constrained optimization problems, where a given solution may not satisfy any constraint qualification but satisfies the standard second-order sufficient condition for optimality. Based on local identification of the rank of the constraints degeneracy via the singular-value decomposition, we derive a modified primal-dual optimality system whose solution is locally unique, nondegenerate, and thus can be found by standard Newton-type techniques. Using identification of active constraints, we further extend our approach to mixed equality- and inequality-constrained problems, and to mathematical programs with complementarity constraints (MPCC). In particular, for MPCC we obtain a local algorithm with quadratic convergence under the second-order sufficient condition only, without any constraint qualifications, not even the special MPCC constraint qualifications.

Key words. constraints degeneracy, regularity, second-order sufficiency, singular-value decomposition, Newton method, mathematical programs with complementarity constraints (MPCC)

AMS subject classifications. 90C30, 90C33, 90C55, 65K05

DOI. 10.1137/S1052623403427264

1. Introduction. We consider the mathematical programming problem

$$(1.1) \quad \begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & F(x) = 0, \quad G(x) \leq 0, \end{array}$$

where $f : \mathbf{R}^n \rightarrow \mathbf{R}$ is a smooth function, and $F : \mathbf{R}^n \rightarrow \mathbf{R}^l$ and $G : \mathbf{R}^n \rightarrow \mathbf{R}^m$ are smooth mappings. Let $\bar{x} \in \mathbf{R}^n$ be a local solution of (1.1). For the superlinear convergence of the method presented below, we shall need the problem data to be twice differentiable near $\bar{x}$, with the second derivatives being continuous at $\bar{x}$. For quadratic convergence, second derivatives should be Lipschitz continuous near $\bar{x}$.

The stationary points of problem (1.1) and the associated (normal) Lagrange multipliers are characterized by the Karush–Kuhn–Tucker (KKT) optimality system

$$(1.2) \quad \frac{\partial L}{\partial x}(x, \lambda, \mu) = 0, \quad F(x) = 0, \quad \mu \geq 0, \quad G(x) \leq 0, \quad \langle \mu, G(x) \rangle = 0,$$

where

$$L : \mathbf{R}^n \times \mathbf{R}^l \times \mathbf{R}^m \rightarrow \mathbf{R}, \quad L(x, \lambda, \mu) = f(x) + \langle \lambda, F(x) \rangle + \langle \mu, G(x) \rangle$$

is the standard Lagrangian function of problem (1.1).

*Received by the editors May 6, 2003; accepted for publication (in revised form) February 11, 2004; published electronically November 17, 2004.

<http://www.siam.org/journals/siopt/15-1/42726.html>

[†]Moscow State University, Faculty of Computational Mathematics and Cybernetics, Department of Operations Research, Leninskiye Gori, GSP-2, 119899 Moscow, Russia (izmaf@ccas.ru). The research of this author was supported by the Russian Foundation for Basic Research under grant 01-01-00810 and the Council for the State Support of Leading Scientific Schools under grant 00-15-96141.

[‡]Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro, RJ 22460-320, Brazil (solodov@impa.br). The research of this author was supported in part by CNPq grants 300734/95-6(RN) and 471780/2003-0, and by FAPERJ.

For a local solution $\bar{x}$ of (1.1), let $\bar{Y}$ be the set of Lagrange multipliers associated with $\bar{x}$; that is,

$$\bar{Y} = \bar{Y}(\bar{x}) = \left\{ (\lambda, \mu) \in \mathbf{R}^l \times \mathbf{R}^m \mid \frac{\partial L}{\partial x}(\bar{x}, \lambda, \mu) = 0, \mu \geq 0, \langle \mu, G(\bar{x}) \rangle = 0 \right\}.$$

If nonempty, $\bar{Y}$ is a convex polyhedral set. As is well known, $\bar{Y}$ is nonempty and bounded if and only if the Mangasarian–Fromovitz constraint qualification (MFCQ) holds:

$$(1.3) \quad \text{rank } F'(\bar{x}) = l$$

and

$$\exists \bar{\xi} \in \ker F'(\bar{x}) \text{ such that } G'_{\bar{I}}(\bar{x})\bar{\xi} < 0,$$

where $\bar{I} = \bar{I}(\bar{x}) = \{i = 1, \dots, m \mid G_i(\bar{x}) = 0\}$ is the set of indices of inequality constraints active at $\bar{x}$. If the stronger linear independence constraint qualification (LICQ) holds, i.e.,

$$\text{the rows of } F'(\bar{x}), G'_{\bar{I}}(\bar{x}) \text{ are linearly independent,}$$

then $\bar{Y}$ is a singleton. At the same time, as is well known, violation of MFCQ can give rise to the situation when $\bar{Y} = \emptyset$.

In this paper, we are concerned with problem (1.1) (or system (1.2)) in the following setting: We do not assume *any* constraint qualification (CQ) to hold at the solution $\bar{x}$; at the same time, we assume the existence of a pair of Lagrange multipliers $(\bar{\lambda}, \bar{\mu}) \in \bar{Y}$ and the standard second-order sufficient condition (SOSC) with this pair of multipliers:

$$(1.4) \quad \left\langle \frac{\partial^2 L}{\partial x^2}(\bar{x}, \bar{\lambda}, \bar{\mu})\xi, \xi \right\rangle > 0 \quad \forall \xi \in \bar{K} \setminus \{0\},$$

where $\bar{K}$ is the critical cone of problem (1.1) at $\bar{x}$; that is,

$$\bar{K} = \bar{K}(\bar{x}) = \{\xi \in \ker F'(\bar{x}) \mid \langle f'(\bar{x}), \xi \rangle \leq 0, G'_{\bar{I}}(\bar{x})\xi \leq 0\}.$$

The case of possible violation of standard CQs has recently received much attention in the literature [40, 20, 21, 16, 35, 1, 2, 3, 23, 17, 5, 4, 24, 41, 42, 43]. As discussed in some of those references, the motivation for considering such *irregular* cases comes from various problems, where either standard CQs are inherently violated (e.g., mathematical programs with equilibrium constraints (MPEC) [34, 37, 3, 36, 38, 18]) or constraints tend to be degenerate or “nearly” (numerically) degenerate (e.g., some classes of large-scale problems). Of interest are both theoretical properties of irregular problems as well as convergence of optimization algorithms applied to such problems (and possible modification of the algorithms to improve robustness). We note that most of the cited non-MPEC literature (actually, all except for [21, 20, 2, 17]) deal with the case when LICQ can be violated (or, more generally, the multiplier can be nonunique), but MFCQ still holds. We comment on those issues in more detail in section 5, where some comparisons are also given.

We emphasize that in this paper, we do not assume *any* CQ. Hence, the existence of multipliers is not automatic. At the same time, our development relies on the

existence of the pair of multipliers $(\bar{\lambda}, \bar{\mu})$. We point out that the situation when CQs are violated but the multipliers exist is not unusual and deserves special consideration. For instance, it is known that for purely equality-constrained problems, multipliers exist generically, provided the number of variables is large enough with respect to the number of constraints [6]. Also, much of the MPEC literature (see, for example, [36, 18]) is concerned with the case of normal Lagrange multipliers. Even though their existence cannot be guaranteed, it turns out that in many cases of MPEC interest the multipliers do exist.

The rest of the paper is organized as follows. In section 2, we introduce the main ideas for the purely equality-constrained problem. The local algorithm we propose is well defined and attains superlinear/quadratic convergence under the sole assumption of SOSC. It is based on the local analysis of the constraints degeneracy via the singular-value decomposition, and on using this information to construct a modified (optimality) system of nonlinear equations whose solution is locally unique and non-degenerate. In section 3, we extend the method to the case of mixed equality and inequality constraints via the identification of constraints active at a solution. We note that SOSC is again sufficient for this construction. An application to mathematical programs with complementarity constraints (MPCC) is discussed in section 4, together with two examples to illustrate this application. In particular, we show that our approach leads to a local quadratically convergent algorithm even if the special MPCC constraint qualifications do not hold (classical CQs are automatically violated in MPCC). In section 5, we compare our approach with some other algorithms, which either do not require CQs or use some relaxed CQs.

2. Equality-constrained problems. We start our discussion with the special case of the purely equality-constrained problem

$$(2.1) \quad \begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & F(x) = 0. \end{array}$$

In this case,

$$L(x, \lambda) = f(x) + \langle \lambda, F(x) \rangle, \quad x \in \mathbf{R}^n, \lambda \in \mathbf{R}^l,$$

and the KKT system (1.2) reduces to the Lagrange system of nonlinear equations

$$(2.2) \quad \Phi(x, \lambda) = 0,$$

where

$$\Phi : \mathbf{R}^n \times \mathbf{R}^l \rightarrow \mathbf{R}^n \times \mathbf{R}^l, \quad \Phi(x, \lambda) = \left(\frac{\partial L}{\partial x}(x, \lambda), F(x) \right).$$

The set of Lagrange multipliers associated with a local solution $\bar{x}$ of (2.1) is the solution set of the system of linear equations:

$$(2.3) \quad \bar{Y} = \{ \lambda \in \mathbf{R}^l \mid (F'(\bar{x}))^T \lambda = -f'(\bar{x}) \}.$$

For (2.1), MFCQ and LICQ are the same: They both reduce to (1.3). Furthermore, we have that $\bar{Y}$ is a singleton if and only if (1.3) holds. Thus, without (1.3), $\bar{Y}$ cannot be a singleton (in which case $\bar{Y}$ is a plane parallel to $\ker(F'(\bar{x}))^T = (\text{im } F'(\bar{x}))^\perp$). It follows that without (1.3), a solution $(\bar{x}, \bar{\lambda})$ of (2.2) cannot be isolated, $\Phi'(\bar{x}, \bar{\lambda})$ is necessarily singular, and one cannot use standard Newton-type methods for solving (2.2) (or

(2.1)). For future discussion, recall that a Newton method for (2.2) is equivalent to the sequential quadratic programming (SQP) algorithm applied to (2.1); see, e.g., [8].

Thus, we face the situation where any solution of (2.2) of the form $(\bar{x}, \lambda)$, $\lambda \in \mathbf{R}^l$, is expected to be degenerate. It is important to comment that a class of methods developed for computing degenerate solutions of general systems of nonlinear equations (see [27, 9]; see also [26, section 2] and the references therein) is not applicable in this context for some principal reasons. Specifically, it can be shown that the so-called nonsingularity of the second differential [9] is automatically violated for the Lagrange optimality system (2.2) in the situation where SOSC holds but (1.3) does not. Thus, while this class of methods for solving degenerate equations can be useful elsewhere, it does not help in solving (2.2) (or (2.1)) when LICQ is violated. Another possibility is to use the regularization methods, e.g., of the Levenberg–Marquardt type, such as those in [17]. This class of methods can attain the superlinear convergence under the stated assumptions. However, the regularization parameter has to be driven to zero, thus making the subproblems locally numerically ill-conditioned whenever standard regularity conditions do not hold. In this paper, we propose a different approach, one not based on the Levenberg–Marquardt type of regularization.

Suppose, for a moment, that $\bar{x}$ and $\bar{\lambda} \in \bar{Y}$ are known. Let $\bar{\Pi}$ be the orthogonal projector in $\mathbf{R}^l$ onto $(\text{im } F'(\bar{x}))^\perp = \ker(F'(\bar{x}))^T$, i.e.,

$$(2.4) \quad \bar{\Pi}^2 = \bar{\Pi}, \quad \text{im } \bar{\Pi} = \ker(F'(\bar{x}))^T, \quad \ker \bar{\Pi} = \text{im } F'(\bar{x}),$$

and consider the following modification of system (2.2):

$$(2.5) \quad \begin{aligned} (F'(x))^T \lambda &= -f'(x), \\ F(x) + \bar{\Pi}(\lambda - \bar{\lambda}) &= 0. \end{aligned}$$

Obviously, $(\bar{x}, \bar{\lambda})$ is a solution of (2.5). Let $(\bar{x}, \lambda)$ be any other solution. Then it holds that

$$\left. \begin{aligned} \lambda - \bar{\lambda} &\in \ker(F'(x))^T = (\text{im } F'(\bar{x}))^\perp \\ \lambda - \bar{\lambda} &\in \ker \bar{\Pi} = \text{im } F'(\bar{x}) \end{aligned} \right\} \Rightarrow \lambda - \bar{\lambda} = 0.$$

Hence, $\bar{\lambda}$ is the unique solution of (2.5) associated with $\bar{x}$.

Of course, $\bar{x}$ and $\bar{Y}$ are unknown. The idea is to construct a suitable approximation of the system (2.5) using the information available at some $(\tilde{x}, \tilde{\lambda}) \in \mathbf{R}^n \times \mathbf{R}^l$, which is close enough to $\bar{x}$ and some $\bar{\lambda} \in \bar{Y}$. We shall modify the system (2.2) in such a way that the new system will have an isolated solution of the form $(\bar{x}, \hat{\lambda})$ with some $\hat{\lambda} \in \bar{Y}$ close to $\bar{\lambda}$. This general idea goes back to [22, 27]. However, the proposals in [22, 27] have some drawbacks when compared to the approach developed here. First, the regularized system proposed in the cited work is different and, in particular, it is overdetermined. Second, the initial guess close enough to the specific multiplier is required. And even more importantly, no procedure for approximating $\bar{\Pi}$ was suggested in [22, 27], nor were conditions that are necessary for a suitable approximation discussed. Here, we give these questions a thorough consideration. Furthermore, we extend the approach to mixed constraints, including the important case of complementarity constraints.

Let a pair $(\tilde{x}, \tilde{\lambda})$ close enough to $(\bar{x}, \bar{\lambda})$ be given, and consider the “regularized” system

$$(2.6) \quad \Phi_{\tilde{x}, \tilde{\lambda}}(x, \lambda) = 0,$$

where

$$\Phi_{\tilde{x}, \tilde{\lambda}} : \mathbf{R}^n \times \mathbf{R}^l \rightarrow \mathbf{R}^n \times \mathbf{R}^l, \quad \Phi_{\tilde{x}, \tilde{\lambda}}(x, \lambda) = \left(\frac{\partial L}{\partial x}(x, \lambda), F(x) + \Pi(\tilde{x}, \tilde{\lambda})(\lambda - \tilde{\lambda}) \right),$$

and $\Pi(\tilde{x}, \tilde{\lambda})$ is a projector in $\mathbf{R}^l$; that is,

$$(2.7) \quad (\Pi(\tilde{x}, \tilde{\lambda}))^2 = \Pi(\tilde{x}, \tilde{\lambda}),$$

satisfying

$$(2.8) \quad \Pi(\tilde{x}, \tilde{\lambda}) \rightarrow \bar{\Pi} \quad \text{as } \tilde{x} \rightarrow \bar{x}, \tilde{\lambda} \rightarrow \bar{\lambda}.$$

In what follows, we shall show how an appropriate $\Pi(\tilde{x}, \tilde{\lambda})$ can be constructed using the singular-value decomposition of $F'(\tilde{x})$. But first, we show that the general idea outlined above indeed resolves the problem of nonisolated multipliers.

PROPOSITION 2.1. *Let $(\bar{x}, \bar{\lambda}) \in \mathbf{R}^n \times \mathbf{R}^l$ be a solution of (2.2).*

Then for every $(\tilde{x}, \tilde{\lambda}) \in \mathbf{R}^n \times \mathbf{R}^l$ close enough to $(\bar{x}, \bar{\lambda})$, there exists a unique element $\hat{\lambda} = \hat{\lambda}(\tilde{x}, \tilde{\lambda}) \in \mathbf{R}^l$ such that $(\tilde{x}, \hat{\lambda})$ is a solution of (2.6) (that is, $\hat{\lambda}$ is the unique element in $\bar{Y}$ satisfying $\Pi(\tilde{x}, \tilde{\lambda})(\hat{\lambda} - \tilde{\lambda}) = 0$). Moreover,

$$(2.9) \quad \hat{\lambda} \rightarrow \bar{\lambda} \quad \text{as } \tilde{x} \rightarrow \bar{x}, \tilde{\lambda} \rightarrow \bar{\lambda}.$$

Proof. Consider the linear system of equations

$$(2.10) \quad (F'(\bar{x}))^T \lambda = -f'(\bar{x}), \quad \bar{\Pi} \tilde{\Pi} \lambda = \bar{\Pi} \tilde{\Pi} \tilde{\lambda},$$

where $\lambda \in \mathbf{R}^l$ is the unknown, and $\tilde{\lambda} \in \mathbf{R}^l$ and the projector $\tilde{\Pi} : \mathbf{R}^l \rightarrow \mathbf{R}^l$ are parameters. One can think of the image space of the operator associated with (2.10) as the l -dimensional space $\text{im}(F'(\bar{x}))^T \times (\text{im } F'(\bar{x}))^\perp$ (recall that $(\text{im } F'(\bar{x}))^\perp = \ker(F'(\bar{x}))^T = \text{im } \bar{\Pi}$). Note also that the right-hand side of (2.10) certainly belongs to $\text{im}(F'(\bar{x}))^T \times (\text{im } F'(\bar{x}))^\perp$, since $-f'(\bar{x}) = (F'(\bar{x}))^T \bar{\lambda}$.

We first show that for the choice of $\tilde{\Pi} = \bar{\Pi}$, the operator of the system (2.10) is nonsingular. Take an arbitrary $\eta \in \mathbf{R}^l$ satisfying

$$(F'(\bar{x}))^T \eta = 0, \quad \bar{\Pi}^2 \eta = 0.$$

Taking into account the definition of $\bar{\Pi}$ (and, in particular, the equality $\bar{\Pi}^2 = \bar{\Pi}$), from the second equation it follows that $\eta \in \ker \bar{\Pi} = \text{im } F'(\bar{x})$. On the other hand, the first equation can be written in the form $\eta \in \ker(F'(\bar{x}))^T = (\text{im } F'(\bar{x}))^\perp$. We conclude that $\eta = 0$, which establishes the assertion.

Next, note that the system (2.10) with $\tilde{\Pi} = \bar{\Pi}$ and $\tilde{\lambda} = \bar{\lambda}$ has a solution $\bar{\lambda}$, and according to what we have already proved, this solution is unique. Now, according to the standard stability theorem for linear systems with nonsingular matrices, we conclude that for every $(\tilde{\Pi}, \tilde{\lambda})$ close enough to $(\bar{\Pi}, \bar{\lambda})$ (in the appropriate space), (2.10) has a unique solution $\hat{\lambda}$ and, moreover,

$$(2.11) \quad \hat{\lambda} \rightarrow \bar{\lambda} \quad \text{as } \tilde{\Pi} \rightarrow \bar{\Pi}, \tilde{\lambda} \rightarrow \bar{\lambda}.$$

Furthermore, taking into account the equality $\tilde{\Pi}^2 = \tilde{\Pi}$, from the second equation in (2.10) we have

$$\begin{aligned} 0 &= \bar{\Pi} \tilde{\Pi} (\hat{\lambda} - \tilde{\lambda}) \\ &= \tilde{\Pi}^2 (\hat{\lambda} - \tilde{\lambda}) + (\bar{\Pi} - \tilde{\Pi}) \tilde{\Pi} (\hat{\lambda} - \tilde{\lambda}) \\ &= (E + (\bar{\Pi} - \tilde{\Pi})) \tilde{\Pi} (\hat{\lambda} - \tilde{\lambda}), \end{aligned}$$

where $E + (\bar{\Pi} - \tilde{\Pi})$ is nonsingular provided that $\tilde{\Pi}$ is close enough to $\bar{\Pi}$ (E denotes the identity operator). Hence, for such $\tilde{\Pi}$, it must hold that $\tilde{\Pi}(\hat{\lambda} - \tilde{\lambda}) = 0$. Combined with the first equation in (2.10), this equality means that $(\bar{x}, \hat{\lambda})$ is a solution of the system

$$(2.12) \quad \frac{\partial L}{\partial x}(x, \lambda) = 0, \quad F(x) + \tilde{\Pi}(\lambda - \tilde{\lambda}) = 0.$$

Suppose now that there exists some other $\lambda \in \mathbf{R}^l$ such that $(\bar{x}, \lambda)$ also solves (2.12). Then $\lambda \in \bar{Y}$ and $\tilde{\Pi}(\lambda - \tilde{\lambda}) = 0$. As is easy to see, it follows that $(\bar{x}, \lambda)$ satisfies (2.10). But the solution of the latter system is unique, as established above. Hence, $(\bar{x}, \hat{\lambda})$ is the unique solution of (2.12).

To complete the proof, it remains to recall (2.8). In particular, the relation (2.9) follows from (2.11). $\square$

We next show how the linear operator satisfying (2.7), (2.8) can be constructed using the singular-value decomposition of $F'(\tilde{x})$ for any $\tilde{x}$ close enough to $\bar{x}$. The construction is based on the identification of the rank of $F'(\tilde{x})$, using information at $(\tilde{x}, \tilde{\lambda})$ only, and a certain *error bound*.

Recall that the singular-value decomposition of $F'(x)$ is

$$F'(x) = U(x)\Sigma(x)(V(x))^T,$$

where $\Sigma(x)$ is the diagonal $l \times n$ matrix with singular values $\sigma_j(x)$ as the diagonal entries, $j = 1, \dots, s$, $s = \min\{n, l\}$, $\sigma_1(x) \geq \sigma_2(x) \geq \dots \geq \sigma_s(x) \geq 0$, while $U(x)$ and $V(x)$ are the orthogonal $l \times l$ and $n \times n$ matrices, respectively. Evidently, $\text{rank } F'(x)$ coincides with the number of positive singular values, and $\text{im } F'(x)$ is spanned by the columns of $U(x)$ corresponding to these values (the columns of $U(x)$ are called left singular vectors).

The error bound to be used below holds under the SOSC. For (2.1), the SOSC (1.4) takes the form

$$(2.13) \quad \left\langle \frac{\partial^2 L}{\partial x^2}(\bar{x}, \bar{\lambda})\xi, \xi \right\rangle > 0 \quad \forall \xi \in \ker F'(\bar{x}) \setminus \{0\}.$$

By [21, Lemma 2], (2.13) implies that there exists a constant $M > 0$ such that

$$(2.14) \quad \|x - \bar{x}\| + \text{dist}(\lambda, \bar{Y}) \leq M\|\Phi(x, \lambda)\|$$

for all $(x, \lambda) \in \mathbf{R}^n \times \mathbf{R}^l$ close enough to $(\bar{x}, \bar{\lambda})$.

PROPOSITION 2.2. *Let $(\bar{x}, \bar{\lambda}) \in \mathbf{R}^n \times \mathbf{R}^l$ be a solution of (2.2), and assume that the error bound (2.14) holds with some $M > 0$ for all $(x, \lambda) \in \mathbf{R}^n \times \mathbf{R}^l$ close enough to $(\bar{x}, \bar{\lambda})$. Let $\theta \in (0, 1)$.*

Then for every $(\tilde{x}, \tilde{\lambda}) \in \mathbf{R}^n \times \mathbf{R}^l$ close enough to $(\bar{x}, \bar{\lambda})$, the integer

$$(2.15) \quad r(\tilde{x}, \tilde{\lambda}) = \max\{0, \max\{j = 1, \dots, s \mid \sigma_j(\tilde{x}) > \|\Phi(\tilde{x}, \tilde{\lambda})\|^\theta\}\}$$

coincides with $\text{rank } F'(\tilde{x})$.

Proof. Let $\theta \in (0, 1)$ be fixed, and let $\bar{r} = \bar{r}(\bar{x}) = \text{rank } F'(\bar{x})$. By the well-known results of the perturbation theory for linear operators [30, 31], if F is sufficiently smooth, then it holds that

$$|\sigma_j(\tilde{x}) - \sigma_j(\bar{x})| \leq \|F'(\tilde{x}) - F'(\bar{x})\|$$

for all $j = 1, \dots, s$. Then for $(\tilde{x}, \tilde{\lambda})$ close enough to $(\bar{x}, \bar{\lambda})$, and for every $j = \bar{r} + 1, \dots, s$, we have that

$$\begin{aligned}\sigma_j(\tilde{x}) &= |\sigma_j(\tilde{x}) - \sigma_j(\bar{x})| \\ &\leq \|F'(\tilde{x}) - F'(\bar{x})\| \\ &\leq N\|\tilde{x} - \bar{x}\| \\ &\leq MN\|\Phi(\tilde{x}, \tilde{\lambda})\| \\ &\leq \|\Phi(\tilde{x}, \tilde{\lambda})\|^\theta,\end{aligned}$$

where the second inequality follows from the local Lipschitz-continuity (with some modulus $N > 0$) of F' , and the third inequality is by (2.14).

The right-hand side of the last inequality tends to zero as $(\tilde{x}, \tilde{\lambda}) \rightarrow (\bar{x}, \bar{\lambda})$ and hence, separates the first $\bar{r}$ singular values of $F'(\tilde{x})$ from the other singular values. In particular, it is easily seen that $r(\tilde{x}, \tilde{\lambda}) = \bar{r}$, where $r(\tilde{x}, \tilde{\lambda})$ is defined by (2.15). $\square$

Proposition 2.2 provides all the information necessary for defining the operator $\Pi(\tilde{x}, \tilde{\lambda})$.

PROPOSITION 2.3. *Suppose that the assumptions of Proposition 2.2 hold. Let $\Pi(\tilde{x}, \tilde{\lambda})$ be defined as the orthogonal projector onto the linear subspace in $\mathbf{R}^l$ spanned by the last $l - r(\tilde{x}, \tilde{\lambda})$ left singular vectors of $F'(\tilde{x})$.*

Then conditions (2.7) and (2.8) hold.

Proof. Condition (2.7) follows immediately by the definition of $\Pi(\tilde{x}, \tilde{\lambda})$.

Condition (2.8) follows from the fact that for each $\tilde{\lambda}$ close enough to $\bar{\lambda}$, $\Pi(\cdot, \tilde{\lambda})$ is the so-called total projector of the 0-group of the symmetric operator $F'(\cdot)(F'(\cdot))^T$ near the point $\bar{x}$, and from the continuity properties of total projectors [30]. (Singular vectors can behave quite chaotically, while total projectors are continuous, and even smooth, under natural assumptions.) $\square$

Throughout the rest of the paper, we assume that $\Pi(\tilde{x}, \tilde{\lambda})$ is chosen according to the presented procedure. A linear operator $\Pi(\tilde{x}, \tilde{\lambda})$ possessing the required properties can be defined differently from the choice presented above [9]. We do not consider those other possibilities here in order to simplify the presentation. We also note that other identification functions could be used in (2.15), similar to techniques for identification of active constraints in [13].

We are now ready to state our local algorithm. The starting point $(\tilde{x}, \tilde{\lambda})$ is supplied as the input information for the algorithm, together with the problem data.

ALGORITHM 2.1. *Preliminary step.* Fix $\theta \in (0, 1)$.

Initialization step. Compute $r(\tilde{x}, \tilde{\lambda})$ according to (2.15). Define $\Pi(\tilde{x}, \tilde{\lambda})$ as the orthogonal projector onto the linear subspace in $\mathbf{R}^l$ spanned by the last $l - r(\tilde{x}, \tilde{\lambda})$ left singular vectors of $F'(\tilde{x})$.

Main step. Generate the sequence $\{(x^k, \lambda^k)\}$ by the Newton method applied to the system (2.6), starting from the initial point $(x^0, \lambda^0) = (\tilde{x}, \tilde{\lambda})$:

$$(x^{k+1}, \lambda^{k+1}) = (x^k, \lambda^k) - (\Phi'_{\tilde{x}, \tilde{\lambda}}(x^k, \lambda^k))^{-1} \Phi_{\tilde{x}, \tilde{\lambda}}(x^k, \lambda^k), \quad k = 0, 1, \dots$$

We note that although Algorithm 2.1 is a completely implementable local scheme, some issues still need to be resolved before its successful numerical realization can be possible. The most important question is how to couple it with some globally convergent method so that the switch to the fast locally convergent iterations can be done efficiently (for example, it is important to avoid an excessive number of singular-value decompositions when we are far from a solution, and so the local Newton steps are not

yet acceptable). Also, it should be verified in practice how large or small is the neighborhood of a solution where (2.15) gives correct identification, and how this depends on the choice of the identification function. We next provide an example showing that the region where (2.15) gives correct identification and where the resulting Newton method applied to (2.6) attains quadratic convergence can be quite large. Of course, this example is very simple and does not warrant any general conclusions.

Example 2.1. Let $n = 1$, $l = 1$, $f(x) = x^2/2$, $F(x) = x^2/2$.

Obviously, $\bar{x} = 0$ is a solution of (2.1) with $\bar{Y} = \mathbf{R}$. LICQ is violated, while SOSC (2.13) holds for any $\bar{\lambda} > -1$.

If we apply the standard Newton method to the Lagrangian system (2.2) starting at some $(x^0, \lambda^0) \in \mathbf{R} \times \mathbf{R}$, then the resulting sequence is given by

$$x^{k+1} = x^k/2, \quad \lambda^{k+1} = (\lambda^k - 1)/2,$$

which evidently converges to $(\bar{x}, \bar{\lambda}) = (0, -1)$, but only linearly. Note also that $\{\lambda^k\}$ is attracted by the multiplier for which SOSC does not hold.

Take any $\tilde{x} \in (0, 1)$ and, for example, $\tilde{\lambda} = 0$. Applying Algorithm 2.1, singular-value decomposition of $F'(\tilde{x})$ gives $\Sigma(\tilde{x}) = \tilde{x}$, $U(\tilde{x}) = 1$, $V(\tilde{x}) = 1$. Furthermore, for any $\theta \in (0, 1)$,

$$\|\Phi(\tilde{x}, \tilde{\lambda})\|^\theta = \left((1 + \tilde{\lambda})^2 \tilde{x}^2 + \tilde{x}^4/4 \right)^{\theta/2} > \tilde{x}^\theta > \tilde{x} = \sigma_1(\tilde{x}).$$

Hence, (2.15) implies that

$$r(\tilde{x}, \tilde{\lambda}) = 0, \quad \Pi(\tilde{x}, \tilde{\lambda}) = 1,$$

which results in the modified optimality system

$$0 = \Phi_{\tilde{x}, \tilde{\lambda}}(x, \lambda) = (x + \lambda x, x^2/2 + \lambda).$$

This system has the unique solution $\hat{\lambda} = 0$ associated with $\bar{x}$, and the Jacobian of this system at $(\bar{x}, \hat{\lambda})$ is the identity matrix. Thus, Algorithm 2.1 generates a quadratically convergent sequence.

Note that in this example, (2.15) defines the correct modified optimality system for any $\tilde{x} \in (0, 1)$, which is the largest region for which it makes sense to speak about quadratic convergence to $(\bar{x}, \hat{\lambda})$.

We proceed to establish the convergence properties of Algorithm 2.1.

THEOREM 2.4. *Let $\bar{x} \in \mathbf{R}^n$ be a local solution of (2.1), and assume that $\bar{\lambda} \in \bar{Y}$ is such that the SOSC (2.13) is satisfied.*

Then for every neighborhood $\mathcal{U}$ of $(\bar{x}, \bar{\lambda})$, and for every $(\tilde{x}, \tilde{\lambda}) \in \mathcal{U}$ close enough to $(\bar{x}, \bar{\lambda})$, Algorithm 2.1 is well defined and generates the sequence $\{(x^k, \lambda^k)\}$ such that

- (a) $\{(x^k, \lambda^k)\}$ remains in $\mathcal{U}$;
- (b) $\{(x^k, \lambda^k)\}$ converges superlinearly/quadratically (depending on the smoothness of f and F) to $(\bar{x}, \hat{\lambda})$, where $\hat{\lambda}$ is defined in Proposition 2.1.

Proof. Essentially, we have to show that the matrix

$$\Phi'_{\tilde{x}, \tilde{\lambda}}(\tilde{x}, \tilde{\lambda}) = \begin{pmatrix} \frac{\partial^2 L}{\partial x^2}(\tilde{x}, \tilde{\lambda}) & (F'(\tilde{x}))^T \\ F'(\tilde{x}) & \bar{\Pi} \end{pmatrix}$$

is nonsingular. Once this fact is established, the assertions follow from the theorem on small perturbations of nonsingular matrices and from the standard argument for

proving local convergence of the Newton method (taking into account Propositions 2.1 and 2.2, and [21, Lemma 2]).

Take an arbitrary pair $(\xi, \eta) \in \ker \Phi'_{\bar{x}, \bar{\lambda}}(\bar{x}, \bar{\lambda})$, i.e.,

$$(2.16) \quad \frac{\partial^2 L}{\partial x^2}(\bar{x}, \bar{\lambda})\xi + (F'(\bar{x}))^T \eta = 0, \quad F'(\bar{x})\xi + \bar{\Pi}\eta = 0.$$

Recall that $F'(\bar{x})\xi$ and $\bar{\Pi}\eta$ belong to the complementary subspaces. Hence, the second equality can be decoupled into two equations:

$$(2.17) \quad F'(\bar{x})\xi = 0, \quad \bar{\Pi}\eta = 0.$$

By the first equalities in (2.16) and (2.17),

$$\begin{aligned} \left\langle \frac{\partial^2 L}{\partial x^2}(\bar{x}, \bar{\lambda})\xi, \xi \right\rangle &= -\langle (F'(\bar{x}))^T \eta, \xi \rangle \\ &= -\langle \eta, F'(\bar{x})\xi \rangle \\ &= 0. \end{aligned}$$

Taking into account again the first equality in (2.17), the SOSC (2.13) implies that $\xi = 0$. Now, from the first equality in (2.16), it follows that $\eta \in \ker(F'(\bar{x}))^T = (\text{im } F'(\bar{x}))^\perp$. On the other hand, the second equality in (2.17) implies the inclusion $\eta \in \ker \bar{\Pi} = \text{im } F'(\bar{x})$. We conclude that $\eta = 0$. We thus established that $\Phi'_{\bar{x}, \bar{\lambda}}(\bar{x}, \bar{\lambda})$ is nonsingular. $\square$

Remark 2.1. A more accurate investigation of the iteration of Algorithm 2.1 leads to the following estimate of the rate of convergence (under the appropriate smoothness assumptions):

$$\|(x^{k+1} - \bar{x}, \lambda^{k+1} - \hat{\lambda})\| = O(\|x^k - \bar{x}\| \|\lambda^k - \hat{\lambda}\| + \|x^k - \bar{x}\|^2).$$

3. Mixed constraints. We now turn our attention to the general problem (1.1). As before, we assume that we have good initial approximations to a local solution $\bar{x}$ of (1.1) and some $(\bar{\lambda}, \bar{\mu}) \in \bar{Y}$ satisfying the SOSC (1.4).

3.1. Using identification of active constraints. As is well known, the KKT system (1.2) can be written in the form

$$(3.1) \quad \Psi(x, \lambda, \mu) = 0,$$

where $\Psi : \mathbf{R}^n \times \mathbf{R}^l \times \mathbf{R}^m \rightarrow \mathbf{R}^n \times \mathbf{R}^l \times \mathbf{R}^m$,

$$\Psi(x, \lambda, \mu) = \left(\frac{\partial L}{\partial x}(x, \lambda, \mu), F(x), \min\{\mu, -G(x)\} \right),$$

with the minimum operation taken componentwise. According to [21, Lemma 2] (see also [17, Lemma 5]), the residual of (3.1) can be used to estimate the distance from a given $(\tilde{x}, \tilde{\lambda}, \tilde{\mu}) \in \mathbf{R}^n \times \mathbf{R}^l \times \mathbf{R}^m$ to the set $\{\bar{x}\} \times \bar{Y}$. Specifically, under our assumptions there exists a constant $M > 0$ such that

$$(3.2) \quad \|x - \bar{x}\| + \text{dist}((\lambda, \mu), \bar{Y}) \leq M \|\Psi(x, \lambda, \mu)\|$$

for all $(x, \lambda, \mu) \in \mathbf{R}^n \times \mathbf{R}^l \times \mathbf{R}^m$ close enough to $(\bar{x}, \bar{\lambda}, \bar{\mu})$. This error bound can be further used to identify the constraints active at $\bar{x}$, following the approach of [13].

PROPOSITION 3.1. *Let $(\bar{x}, \bar{\lambda}, \bar{\mu}) \in \mathbf{R}^n \times \mathbf{R}^l \times \mathbf{R}^m$ be a solution of (1.2), and assume that the error bound (3.2) holds with some $M > 0$ for all $(x, \lambda, \mu) \in \mathbf{R}^n \times \mathbf{R}^l \times \mathbf{R}^m$ close enough to $(\bar{x}, \bar{\lambda}, \bar{\mu})$. Let $\tau \in (0, 1)$.*

Then for every $(\tilde{x}, \tilde{\lambda}, \tilde{\mu}) \in \mathbf{R}^n \times \mathbf{R}^l \times \mathbf{R}^m$ close enough to $(\bar{x}, \bar{\lambda}, \bar{\mu})$, the index set

$$(3.3) \quad I(\tilde{x}, \tilde{\lambda}, \tilde{\mu}) = \{i = 1, \dots, m \mid -G_i(\tilde{x}) \leq \|\Psi(\tilde{x}, \tilde{\lambda}, \tilde{\mu})\|^\tau\}$$

coincides with $\bar{I}$.

Remark 3.1. As remarked by a referee, there exist some conflicting opinions about the value of the identification technique proposed in [13] (the possible criticism is that the neighborhood of correct identification is difficult to characterize, and in some cases can be rather small). That said, satisfactory computational experience with some algorithms which employ this technique is reported in [29, 10], and it is also used, for example, in [14, 43] (although there are no numerical results for the latter two references).

Once the set $I = \bar{I}$ is identified, we can consider instead of (1.1) the purely equality-constrained problem

$$(3.4) \quad \begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & F(x) = 0, \quad G_I(x) = 0. \end{array}$$

Obviously, violation at $\bar{x}$ of LICQ for (1.1) means that (3.4) also does not satisfy LICQ. To apply the approach developed in section 2, we need to make sure that SOSC for (3.4) holds whenever we have SOSC for the original problem (1.1). It is easy to see that this is indeed the case.

PROPOSITION 3.2. *Let $\bar{x} \in \mathbf{R}^n$ be a local solution of (1.1), and assume that $(\bar{\lambda}, \bar{\mu}) \in \bar{Y}$ is such that the SOSC (1.4) is satisfied.*

Then $\bar{x}$ is a local solution of problem (3.4) with $I = \bar{I}$, and $(\bar{\lambda}, \bar{\mu}_{\bar{I}})$ is a Lagrange multiplier associated with this solution, i.e.,

$$(3.5) \quad f'(\bar{x}) + (F'(\bar{x}))^T \bar{\lambda} + (G'_{\bar{I}}(\bar{x}))^T \bar{\mu}_{\bar{I}} = 0.$$

Furthermore, SOSC for (3.4) holds with this multiplier:

$$(3.6) \quad \begin{aligned} & \langle f''(\bar{x})\xi, \xi \rangle + \langle \bar{\lambda}, F''(\bar{x})[\xi, \xi] \rangle + \langle \bar{\mu}_{\bar{I}}, G''_{\bar{I}}(\bar{x})[\xi, \xi] \rangle > 0 \\ & \forall \xi \in (\ker F'(\bar{x}) \cap \ker G'_{\bar{I}}(\bar{x})) \setminus \{0\}. \end{aligned}$$

Proof. The equality (3.5) obviously follows from the definition of $\bar{Y}$. Furthermore, since $\bar{\mu}_i = 0$ for all $i \notin \bar{I}$, the left-most expression in (3.6) equals $\langle \frac{\partial^2 L}{\partial x^2}(\bar{x}, \bar{\lambda}, \bar{\mu})\xi, \xi \rangle$. Since we also have that $\ker F'(\bar{x}) \cap \ker G'_{\bar{I}}(\bar{x}) \subset \bar{K}$, it follows that (1.4) implies (3.6). $\square$

We complete this section with a formal statement of the local algorithm for mixed-constrained problems. The starting point $(\tilde{x}, \tilde{\lambda}, \tilde{\mu})$ is supplied as the input information for the algorithm.

ALGORITHM 3.1. *Preliminary step.* Fix $\tau \in (0, 1)$.

Initialization step. Define the index set $I(\tilde{x}, \tilde{\lambda}, \tilde{\mu})$ according to (3.3).

Main step. Apply Algorithm 2.1 to problem (3.4) with $I = I(\tilde{x}, \tilde{\lambda}, \tilde{\mu})$, using $(\tilde{x}, \tilde{\lambda}, \tilde{\mu}_I)$ as the starting point.

The following theorem is a direct consequence of Theorem 2.4 and Propositions 3.1 and 3.2.

THEOREM 3.3. *Let $\bar{x} \in \mathbf{R}^n$ be a local solution of (1.1), and assume that $(\bar{\lambda}, \bar{\mu}) \in \bar{Y}$ is such that the SOSC (1.4) is satisfied.*

Then for every $(\tilde{x}, \tilde{\lambda}, \tilde{\mu}) \in \mathbf{R}^n \times \mathbf{R}^l \times \mathbf{R}^m$ close enough to $(\bar{x}, \bar{\lambda}, \bar{\mu})$, Algorithm 3.1 is well defined and generates a sequence $\{(x^k, \lambda^k, \mu^k)\}$, which converges superlinearly/quadratically (depending on the smoothness of f , F , and G) to $(\bar{x}, \hat{\lambda}, \hat{\mu}_{\bar{I}})$, with some $\hat{\lambda} \in \mathbf{R}^l$ and $\hat{\mu}_{\bar{I}} \in \mathbf{R}^{|\bar{I}|}$ such that

$$(\hat{\lambda}, \hat{\mu}_{\bar{I}}) \rightarrow (\bar{\lambda}, \bar{\mu}_{\bar{I}}) \text{ as } \tilde{x} \rightarrow \bar{x}, \tilde{\lambda} \rightarrow \bar{\lambda}, \tilde{\mu} \rightarrow \bar{\mu}.$$

3.2. Reformulation with slack variables. Instead of identifying active constraints, one could transform (1.1) into the purely equality-constrained problem using slack variables:

$$(3.7) \quad \begin{aligned} & \text{minimize} && f(x) \\ & \text{subject to} && F(x) = 0, \quad g_i(x) + \sigma_i^2 = 0, \quad i = 1, \dots, m, \end{aligned}$$

where $g_1, \dots, g_m$ are the components of G . As is well known, local solutions of (1.1) and (3.7) are in correspondence with each other: $\bar{x}$ is a local solution of (1.1) if and only if $(\bar{x}, \bar{\sigma})$ with $\bar{\sigma} = (\sqrt{-g_1(\bar{x})}, \dots, \sqrt{-g_m(\bar{x})})$ is a well-defined local solution of (3.7).

This approach is quite attractive because of its universality and simplicity. However, it requires stronger assumptions. Specifically, it is known that to guarantee SOSC for (3.7) at $(\bar{x}, \bar{\sigma})$ with multiplier $(\bar{\lambda}, \bar{\mu})$, one needs not only SOSC for the original problem (1.1), but also the strict complementarity condition (SCC), i.e., $\bar{\mu}_i > 0$ for all $i \in \bar{I}$ (see [7, Theorem 1.32]). We note that the SCC is regarded as a somewhat restrictive (and hence, undesirable) assumption.

4. Application to MPCC. An interesting application of the approach developed in this paper is concerned with MPCC, the mathematical programs with complementarity constraints. Here, we consider the problem of the form

$$(4.1) \quad \begin{aligned} & \text{minimize} && f(x) \\ & \text{subject to} && F(x) = 0, \quad G(x) \leq 0, \\ & && H_1(x) \geq 0, \quad H_2(x) \geq 0, \quad \langle H_1(x), H_2(x) \rangle = 0, \end{aligned}$$

where $H_1, H_2 : \mathbf{R}^n \rightarrow \mathbf{R}^s$ are smooth mappings. Of course, (4.1) can be considered as a usual mixed-constrained optimization problem. Recall, however, that MFCQ is violated at every feasible point of this problem (since there exists no feasible point strictly satisfying the inequality constraints). Therefore, this class of problems is of particular interest in the context of our development.

The Lagrangian for (4.1) is given by

$$\begin{aligned} L(x, \lambda, \mu, \mu^1, \mu^2, \nu) &= f(x) + \langle \lambda, F(x) \rangle + \langle \mu, G(x) \rangle \\ &\quad - \langle \mu^1, H_1(x) \rangle - \langle \mu^2, H_2(x) \rangle + \nu \langle H_1(x), H_2(x) \rangle, \\ x \in \mathbf{R}^n, \lambda \in \mathbf{R}^l, \mu \in \mathbf{R}^m, \mu^1, \mu^2 \in \mathbf{R}^s, \nu \in \mathbf{R}, \end{aligned}$$

and the set $\bar{Y}$ associated with a local solution $\bar{x}$ of (4.1) consists of tuples $(\lambda, \mu, \mu^1, \mu^2, \nu) \in \mathbf{R}^l \times \mathbf{R}^m \times \mathbf{R}^s \times \mathbf{R}^s \times \mathbf{R}$ such that

$$\begin{aligned} (F'(\bar{x}))^T \lambda + (G'(\bar{x}))^T \mu - (H'_1(\bar{x}))^T (\mu^1 - \nu H_2(\bar{x})) - (H'_2(\bar{x}))^T (\mu^2 - \nu H_1(\bar{x})) &= -f'(\bar{x}), \\ \mu &\geq 0, \quad \langle \mu, G(\bar{x}) \rangle = 0, \\ \mu^1 &\geq 0, \quad \langle \mu^1, H_1(\bar{x}) \rangle = 0, \quad \mu^2 \geq 0, \quad \langle \mu^2, H_2(\bar{x}) \rangle = 0. \end{aligned}$$

It is easy to see that $\bar{x}$ is a stationary point in (4.1) in the sense that $\bar{Y} \neq \emptyset$ if and only if it is strongly stationary in the terminology of [36]. Moreover, the multipliers corresponding to these stationarity concepts are related by explicit formulas (see [18, Proposition 4.1]). We note that strong stationarity is regarded as a reasonable stationarity concept in the context of Newton-type methods for MPCC.

As is easy to see,

$$(4.2) \quad \bar{K} = \left\{ \xi \in \ker F'(\bar{x}) \mid \begin{array}{l} \langle f'(\bar{x}), \xi \rangle \leq 0, \quad G'_I(\bar{x})\xi \leq 0, \\ (H_1)'_{I_1 \setminus I_2}(\bar{x})\xi = 0, \quad (H_2)'_{I_2 \setminus I_1}(\bar{x})\xi = 0, \\ (H_1)'_{I_0}(\bar{x})\xi \geq 0, \quad (H_2)'_{I_0}(\bar{x})\xi \geq 0 \end{array} \right\},$$

where

$$I = \{i = 1, \dots, m \mid G_i(\bar{x}) = 0\},$$

$$I_1 = \{i = 1, \dots, s \mid (H_1)_i(\bar{x}) = 0\}, \quad I_2 = \{i = 1, \dots, s \mid (H_2)_i(\bar{x}) = 0\}, \quad I_0 = I_1 \cap I_2.$$

Note that necessarily $I_1 \cup I_2 = \{1, \dots, s\}$. For (4.1), the SOSC (1.4) takes the form

$$(4.3) \quad \left\langle \frac{\partial^2 L}{\partial x^2}(\bar{x}, \bar{\lambda}, \bar{\mu}, \bar{\mu}^1, \bar{\mu}^2, \bar{\nu})\xi, \xi \right\rangle > 0 \quad \forall \xi \in \bar{K} \setminus \{0\},$$

with $\bar{K}$ defined in (4.2) and a tuple of multipliers $(\bar{\lambda}, \bar{\mu}, \bar{\mu}^1, \bar{\mu}^2, \bar{\nu}) \in \bar{Y}$. This condition is exactly the same as SOSC for MPCC in [18, Definition 3.4] if we take into account the relation between multipliers corresponding to the two different (but equivalent) stationarity concepts. Recall also that the MPCC-LICQ [36, 38, 18] consists of saying that

$$(4.4) \quad \text{the rows of } F'(\bar{x}), G'_I(\bar{x}), (H_1)'_{I_1}(\bar{x}), (H_2)'_{I_2}(\bar{x}) \text{ are linearly independent.}$$

If we have a good approximation to $\bar{x}$ and $(\bar{\lambda}, \bar{\mu}, \bar{\mu}^1, \bar{\mu}^2, \bar{\nu}) \in \bar{Y}$ satisfying (4.3), then Algorithm 3.1 can be directly applied to (4.1), and its local convergence is characterized by Theorem 3.3. Note that no CQ is needed, as compared to other methods for MPECs which typically need special CQs, such as MPCC-LICQ (4.4). We comment more on this issue in section 5.

We complete this section with two examples that illustrate Algorithm 3.1 when applied to MPCC. The first example is rather simple, and, in particular, it satisfies the MPCC-LICQ (4.4). Together with the SOSC (4.3), this means that other methods (such as slightly modified SQP) can also be used in that case. We include this example mainly because it is standard in the MPEC literature, e.g., [36], and because it is easier to follow its development. The second example is a modification of [25, Example 4.2]. Here, MPCC-LICQ is not satisfied, and so this better serves to illustrate the power of the approach presented in this paper.

Example 4.1. Let $n = 2, l = m = 0, s = 1, f(x) = a_1x_1 + a_2x_2$, where $a_1, a_2 > 0$, $H_1(x) = x_1, H_2(x) = x_2$. Obviously, $\bar{x} = 0$ is a solution of (4.1), and

$$\bar{Y} = \{(\bar{\mu}^1, \bar{\mu}^2, \nu) \mid \bar{\mu}^1 = a_1, \bar{\mu}^2 = a_2, \nu \in \mathbf{R}\}.$$

We further have that $I_0 = I_1 = I_2 = \{1\}$, and

$$\bar{K} = \{\xi \in \mathbf{R}^2 \mid a_1\xi_1 + a_2\xi_2 \leq 0, \xi_1 \geq 0, \xi_2 \geq 0\} = \{0\}.$$

Hence, the SOSC (4.3) is satisfied automatically.

On the initialization step of Algorithm 3.1, for $(\tilde{x}, \tilde{\mu}^1, \tilde{\mu}^2, \tilde{\nu})$ close enough to $(\bar{x}, a_1, a_2, \bar{\nu})$ with some $\bar{\nu} \in \mathbf{R}$, it is correctly identified that both inequality constraints are active at the solution; that is, $\bar{I} = I(\tilde{x}, \tilde{\mu}^1, \tilde{\mu}^2, \tilde{\nu}) = \{1, 2\}$. Hence, problem (3.4) takes the form

$$\begin{aligned} & \text{minimize} && a_1 x_1 + a_2 x_2 \\ & \text{subject to} && \tilde{F}(x) = (x_1, x_2, x_1 x_2) = 0. \end{aligned}$$

As is easy to see, $\bar{\Pi}$ is the orthogonal projector onto

$$\ker(\tilde{F}'(\bar{x}))^T = \ker \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} = \{y \in \mathbf{R}^3 \mid y_1 = y_2 = 0\},$$

i.e.,

$$\bar{\Pi}y = (0, 0, y_3), \quad y \in \mathbf{R}^3.$$

On the initialization step of Algorithm 2.1, it is correctly identified that $r(\tilde{x}, \tilde{\mu}^1, \tilde{\mu}^2, \tilde{\nu}) = 2$. Then $\Pi(\tilde{x}, \tilde{\mu}^1, \tilde{\mu}^2, \tilde{\nu})$ is taken as some operator approximating $\bar{\Pi}$.

Suppose, for a moment, that $\Pi(\tilde{x}, \tilde{\mu}^1, \tilde{\mu}^2, \tilde{\nu}) = \bar{\Pi}$. Then the corresponding regularized system (2.6) takes the form

$$\begin{aligned} a_1 - \mu^1 + \nu x_2 &= 0, \\ a_2 - \mu^2 + \nu x_1 &= 0, \\ x_1 &= 0, \quad x_2 = 0, \\ x_1 x_2 + \nu - \tilde{\nu} &= 0. \end{aligned}$$

Evidently, this system has a solution of the form $(\bar{x}, a_1, a_2, \tilde{\nu})$. The Jacobian of this system at this solution is the matrix

$$\begin{pmatrix} 0 & \tilde{\nu} & -1 & 0 & 0 \\ \tilde{\nu} & 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

which is nonsingular for any choice of $\tilde{\nu}$. Hence, the Newton method applied to this system is locally quadratically convergent to this solution.

Obviously, if $\Pi(\tilde{x}, \tilde{\mu}^1, \tilde{\mu}^2, \tilde{\nu})$ is close enough to $\bar{\Pi}$, then we obtain a small perturbation of the above system of equations, and thus the same conclusions hold.

We next present an example where MPCC-LICQ (4.4) does not hold. The authors are not aware of any approaches which lead to a local algorithm with superlinear convergence for this example, except for the one presented here and in [43].

Example 4.2. Let $n = 3$, $l = m = 0$, $s = 2$, $f(x) = x_1 + x_2 + x_1^2 + x_2^2 + (x_3 - 1)^2$, $H_1(x) = (x_1 + x_2, x_1^2 - (x_3 - 1)^2)$, $H_2(x) = (x_2, x_3)$.

As is easy to see, $\bar{x} = (0, 0, 1)$ is the solution of (4.1). We also have that $I_1 = \{1, 2\}$, $I_2 = \{1\}$, $I_0 = \{1\}$.

Note that MPCC-LICQ (4.4) is clearly violated, since

$$H'_1(\bar{x}) = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

In fact, any other “standard” MPEC CQ is violated in this example. On the other hand, the constraints of this problem satisfy the special 2-regularity condition [25], and thus the corresponding (directional) optimality conditions hold. For some critical directions, those conditions can be reduced to the form considered in this section. (Of course, one can just verify the latter directly.) In particular, we have that

$$\bar{Y} = \{(\mu^1, \bar{\mu}^2, \nu) \mid \mu_1^1 = 1, \mu_2^1 \geq 0, \bar{\mu}^2 = 0, \nu \in \mathbf{R}\}.$$

Take $\bar{\mu}^1 = (1, 0)$, $\bar{\nu} = 0$. It can be verified that

$$\frac{\partial^2 L}{\partial x^2}(\bar{x}, \bar{\mu}^1, \bar{\mu}^2, \bar{\nu}) = 2E,$$

where E is the identity matrix. It follows that the SOSC (4.3) holds for this solution.

Suppose we have a point $(\tilde{x}, \tilde{\mu}^1, \tilde{\mu}^2, \tilde{\nu})$ close enough to this $(\bar{x}, \bar{\mu}^1, \bar{\mu}^2, \bar{\nu})$. Then on the initialization step of Algorithm 3.1, it is correctly identified that $\bar{I} = I(\tilde{x}, \tilde{\mu}^1, \tilde{\mu}^2, \tilde{\nu}) = \{1, 2, 3\}$. Therefore, problem (3.4) takes the form

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \tilde{F}(x) = 0, \end{array}$$

where

$$\tilde{F}(x) = \begin{pmatrix} x_1 + x_2 \\ x_1^2 - (x_3 - 1)^2 \\ x_2 \\ x_2(x_1 + x_2) + x_3(x_1^2 - (x_3 - 1)^2) \end{pmatrix}.$$

We further have that $\bar{\Pi}$ is the orthogonal projector onto

$$\ker(\tilde{F}'(\bar{x}))^T = \ker \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \{y \in \mathbf{R}^4 \mid y_1 = y_3 = 0\},$$

i.e.,

$$\bar{\Pi}y = (0, y_2, 0, y_4), \quad y \in \mathbf{R}^4.$$

On the initialization step of Algorithm 2.1, it is correctly identified that $r(\tilde{x}, \tilde{\mu}^1, \tilde{\mu}_1^2, \tilde{\nu}) = 2$. Then $\Pi(\tilde{x}, \tilde{\mu}^1, \tilde{\mu}_1^2, \tilde{\nu})$ is taken as some operator approximating $\bar{\Pi}$. Suppose for a moment that $\Pi(\tilde{x}, \tilde{\mu}^1, \tilde{\mu}_1^2, \tilde{\nu}) = \bar{\Pi}$. Then the corresponding regularized system (2.6) takes the form

$$\begin{aligned} 1 + 2x_1 - \mu_1^1 - 2\mu_2^1 x_1 + \nu(x_2 + 2x_1 x_3) &= 0, \\ 1 + 2x_2 - \mu_1^1 - \mu_1^2 + \nu(x_1 + 2x_2) &= 0, \\ 2(x_3 - 1) + 2\mu_2^1(x_3 - 1) + \nu(x_1^2 - 3x_3^2 + 4x_3 - 1) &= 0, \\ x_1 + x_2 &= 0, \\ x_1^2 - (x_3 - 1)^2 + \mu_2^1 - \tilde{\mu}_2^1 &= 0, \\ x_2 &= 0, \\ x_2(x_1 + x_2) + x_3(x_1^2 - (x_3 - 1)^2) + \nu - \tilde{\nu} &= 0. \end{aligned}$$

It can be easily seen that this system has a solution $(\bar{x}, (1, \tilde{\mu}_2^1), 0, \tilde{\nu})$, and the Jacobian of this system at this solution is

$$\begin{pmatrix} 2(1 - \tilde{\mu}_2^1 + \tilde{\nu}) & \tilde{\nu} & 0 & -1 & 0 & 0 & 0 \\ \tilde{\nu} & 2(1 + \tilde{\nu}) & 0 & -1 & 0 & -1 & 0 \\ 0 & 0 & 2(1 + \tilde{\mu}_2^1 - \tilde{\nu}) & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

which is nonsingular (in fact, even well conditioned) for all $\tilde{\mu}_2^1$ and $\tilde{\nu}$ close enough to $\bar{\mu}_2^1$ and $\bar{\nu}$, respectively. Hence, the Newton method applied to this system is locally quadratically convergent to this solution.

If $\Pi(\tilde{x}, \tilde{\mu}_1^1, \tilde{\mu}_2^1, \tilde{\nu})$ is close enough to $\bar{\Pi}$, then we obtain a small perturbation of the above system of equations, and thus the same conclusions hold.

5. Comparisons with other approaches. In this section, we briefly review some recent work on Newton-type methods for optimization problems, where CQs are relaxed or removed. We note that a direct complete comparison of different approaches is in most cases impossible. Below, we discuss some advantages and disadvantages, making comparisons where possible.

First, we comment on the standard SQP algorithm. As already mentioned above, in the case of equality constraints, the superlinear convergence of SQP requires LICQ and SOSC. By contrast, our approach needs SOSC only. Of course, this does not come without a computational price—singular-value decomposition of the Jacobian of the constraints is used to set up the modified optimality system. Similarly, for mixed constraints, the SQP method needs the strict MFCQ (which is equivalent to the uniqueness of the multiplier) and SOSC, while for our proposal based on the identification of active constraints, SOSC is sufficient. Additionally, subproblems in our method are systems of linear equations, which are simpler than quadratic programs in SQP. On the other hand, it should be kept in mind that SQP is certainly easier to globalize, as some well-developed globalization strategies are readily available.

SQP with nonunique multipliers has been analyzed in [1, 4], which assumes MFCQ and some second-order sufficiency conditions. Global convergence based on linesearch using a nonsmooth penalty function is established. The algorithm is primal: No approximation of the Hessian of the Lagrangian function is used. Thus, the rate of convergence is only linear, of course.

The so-called stabilized SQP and related algorithms [40, 20, 16, 17, 41, 42, 33] employ regularization techniques, with smart choices of regularization parameters based on error bounds for the primal-dual solution of the problem. Most of the cited literature deals with the case of inequality constraints only, but perhaps the analysis could be extended to mixed constraints. In [42, Theorem 8], local superlinear convergence of an algorithm of this kind to the set $\{\bar{x}\} \times \bar{Y}$ is established under MFCQ and the SOSC (2.13) (the latter is supposed to be satisfied for all $\bar{\mu} \in \bar{Y}$). In [17, Theorem 8], local superlinear convergence to $(\bar{x}, \hat{\mu})$ for some $\hat{\mu} \in \bar{Y}$ close to $\bar{\mu}$ is established under the strong SOSC, that is, (2.13) with

$$\bar{K} = \{\xi \in \mathbf{R}^n \mid G'_{\bar{I}_+}(\bar{x})\xi = 0\}, \quad \bar{I}_+ = \{i \in I \mid \bar{\mu}_i > 0\}.$$

In [33], the stabilized version of SQP [40] is combined with active-set strategies, and convergence is established under a certain relaxed CQ assumption and the strong

SOSC. To our knowledge, [20, 17, 33] are the only papers on stabilized SQP, where MFCQ is not assumed to be satisfied. Compared, e.g., to [42], we do not assume any CQ. Compared to [20, 17, 33], we need a weaker SOSC. Thus, conditions for local superlinear convergence in our approach are weaker than those for stabilized versions of SQP. Additionally, each iteration of a stabilized SQP method typically requires more than just solving a QP (in [33] linear equations are solved instead of QPs, but the number of equations to be solved per iteration can be more than one). It should also be noted that the stabilized SQP shares the common (at least conceptual) drawback of regularization methods: Its subproblems become more ill-conditioned (and more difficult to solve) as the method approaches an irregular solution and the regularization parameter tends to zero. Our approach is free from this drawback. On the other hand, comments above concerning globalization still apply.

Another approach which has recently received some attention is the so-called sequential quadratically constrained quadratic programming (SQCQP), which employs subproblems based on the quadratic approximation of both the objective function and the constraints. The overall behavior may be quite favorable, provided an effective method to solve subproblems is available. In [5], for purely inequality-constrained problems, it is shown that the SQCQP algorithm is locally superlinearly convergent to a local solution satisfying MFCQ and the quadratic growth condition. (Under MFCQ, the latter is equivalent to a weaker form of SOSC, that is, (2.13) with $\bar{\mu}$ replaced by some $\mu(\xi) \in \bar{Y}$.) Globalization issues are considered in [19, 39]. Of course, SQCQP subproblems are more complex than quadratic programs (as in SQP) or linear systems of equations (as in our approach). In the convex case, SQCQP subproblems can be solved rather efficiently (after formulating them as second-order cone programs) by the interior points techniques. But in the nonconvex case, the situation is not so clear. Also, we note that an extension to the nonlinear equality-constrained problems seems problematic, as the subproblems would very likely be infeasible (even locally).

We next mention [2], which suggests an approach for transforming a problem with violated MFCQ into a problem satisfying MFCQ. It should be noted, though, that this approach relies on some strong “quadratic growth condition,” i.e., an estimate of the distance to the given solution from *all* (not only feasible) points close enough to this solution. Under MFCQ, this is equivalent to the standard quadratic growth condition, but without MFCQ, it is much stronger, and some regularity of constraints is implicitly subsumed in this condition.

Very recently, another method for solving degenerate problems has been proposed in [43]. In the case of inequality or mixed constraints, it reduces the problem to the purely equality-constrained case using the same technique to identify active constraints that is employed in our approach. Local superlinear rate of convergence is obtained under essentially the same assumptions as ours (but unlike in our method, the quadratic rate is not achieved). One system of linear equations is solved at each iteration, but the system itself is different from ours. The advantage of [43] when compared to our approach is that singular-value decomposition is not needed and the linear system to be solved is closer in structure to systems that arise in the SQP or primal-dual interior point methods (in particular, it preserves sparsity). The disadvantage is that it involves a regularization parameter which tends to zero with convergence, which means that close to the solution subproblems are ill-conditioned (in the absence of CQs).

We now turn our attention to Newton-type methods for MPEC. Inevitable violation of MFCQ in (4.1) leads to numerous difficulties in theoretical justification of those methods. One of the most significant difficulties is the following: SQP subproblems

can be simply infeasible arbitrarily close to a solution. In [3], the special elastic mode is assumed to be used by a QP solver in order to overcome the infeasibility problem. On the other hand, it was recognized recently that even the behavior of rather basic SQP algorithms (without complex modifications) turns out to be quite favorable in many examples of MPCC. In [18], the authors give a theoretical explanation for this phenomenon. They prove local superlinear convergence under the MPCC-LICQ (4.4), the SOSC (4.3) with $\bar{K}$ defined in (4.2), and some additional assumptions on the multipliers and on the properties of a QP solver. The problem is supposed to be reformulated using slacks, and the authors *assume* that SQP subproblems are feasible (unless the starting point satisfies an additional complementarity condition). It should be pointed out, however, that under the assumptions of [18], SQP subproblems are consistent for some important special cases. For example, this is the case when there are only complementarity constraints in (4.1), i.e., there are no additional equality and inequality constraints (see [18, Lemma 6.1]). In the approach described in section 4, we need only strong stationarity and the SOSC (4.3). We do not assume MPCC-LICQ, and our approach is QP-free. That said, comments on globalization again apply.

To complete this discussion, we briefly examine the examples presented in [18]. All these examples are taken from *MacMPEC*, a collection of MPECs [32]. In **jr*** and **s14**, the SOSC (4.3) holds with any choice of the multipliers. In **scholtes4**, strong stationarity does not hold, i.e., $\bar{Y} = \emptyset$. Finally, in **ralph2**, the SOSC (4.3) holds with the appropriate choice of multipliers. Similarly to section 4, it can be shown that our approach is applicable to all those examples, except for **scholtes4** (which also cannot be efficiently handled by any of the other known methods).

Summarizing, our approach requires weaker assumptions for local superlinear convergence than the alternatives, *where any exist* (the exception to this comment is the subsequent paper [43], which needs the same assumptions). Of course, this does not come for free. The computational price to pay is the singular-value decomposition of the Jacobian of the constraints. Arguably, this is not a very serious drawback, because (in theory) this computation has to be done only once, at the initialization stage of the local method. And it should be kept in mind that once it is done, the subproblems to be solved are just linear equations, which are simpler than in (most) alternatives. A more serious drawback is that the presented local algorithm does not come with a *natural* globalization procedure. The subject of globalization is an issue of current research. At this time, we can prove that the local algorithm can be globalized using the so-called hybrid strategy, in the spirit of methods in the complementarity literature, e.g., [11, 15, 12, 28]. Of course, within the hybrid strategy in general, the singular-value decomposition would need to be computed more than once, until the identification in (2.15) and (3.3) becomes correct. This is a clear numerical disadvantage. We are looking for other (better) alternatives. As already mentioned above, a number of questions still need to be resolved before the proposed approach can be implemented and its numerical potential evaluated.

Acknowledgment. Author Izmailov thanks the Instituto de Matemática Pura e Aplicada, where he was a visiting professor when this work was initiated.

REFERENCES

- [1] M. ANITESCU, *Degenerate nonlinear programming with a quadratic growth condition*, SIAM J. Optim., 10 (2000), pp. 1116–1135.
- [2] M. ANITESCU, *Nonlinear Programs with Unbounded Lagrange Multiplier Sets*, Preprint ANL/MCS-P796-0200, Mathematics and Computer Science Division, Argonne National

- Laboratory, Argonne, IL, 2000.
- [3] M. ANITESCU, *On Solving Mathematical Programs with Complementarity Constraints as Nonlinear Programs*, Preprint ANL/MCS-P864-1200, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, 2000.
- [4] M. ANITESCU, *On the rate of convergence of sequential quadratic programming with nondifferentiable exact penalty function in the presence of constraint degeneracy*, Math. Program., 92 (2002), pp. 359–386.
- [5] M. ANITESCU, *A superlinearly convergent sequential quadratically constrained quadratic programming algorithm for degenerate nonlinear programming*, SIAM J. Optim., 12 (2002), pp. 949–978.
- [6] A. V. ARUTYUNOV, *Optimality Conditions: Abnormal and Degenerate Problems*, Kluwer Academic Publishers, Dordrecht, 2000.
- [7] D. P. BERTSEKAS, *Constraint Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.
- [8] D. P. BERTSEKAS, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, MA, 1999.
- [9] O. A. BREZHNEVA AND A. F. IZMAILOV, *Construction of defining systems for finding singular solutions of nonlinear equations*, Comput. Math. Math. Phys., 42 (2002) pp. 8–19.
- [10] A. N. DARYINA, A. F. IZMAILOV, AND M. V. SOLODOV, *A class of active-set Newton methods for mixed complementarity problems*, SIAM J. Optim., to appear.
- [11] T. DE LUCA, F. FACCHINEI, AND C. KANZOW, *A semismooth equation approach to the solution of nonlinear complementarity problems*, Math. Program., 75 (1996), pp. 407–439.
- [12] T. DE LUCA, F. FACCHINEI, AND C. KANZOW, *A theoretical and numerical comparison of some semismooth algorithms for complementarity problems*, Comput. Optim. Appl., 16 (2000), pp. 173–205.
- [13] F. FACCHINEI, A. FISCHER, AND C. KANZOW, *On the accurate identification of active constraints*, SIAM J. Optim., 9 (1999), pp. 14–32.
- [14] F. FACCHINEI AND C. LAZZARI, *Local feasible QP-free algorithms for the constrained minimization of SC^1 functions*, J. Optim. Theory Appl., 119 (2003), pp. 218–316.
- [15] M. C. FERRIS, C. KANZOW, AND T. S. MUNSON, *Feasible descent algorithms for mixed complementarity problems*, Math. Program., 86 (1999), pp. 475–497.
- [16] A. FISCHER, *Modified Wilson's method for nonlinear programs with nonunique multipliers*, Math. Oper. Res., 24 (1999), pp. 699–727.
- [17] A. FISCHER, *Local behaviour of an iterative framework for generalized equations with nonisolated solutions*, Math. Program., 94 (2002), pp. 91–124.
- [18] R. FLETCHER, S. LEYFFER, D. RALPH, AND S. SCHOLTES, *Local Convergence of SQP Methods for Mathematical Programs with Equilibrium Constraints*, Numerical Analysis Report NA/209, Department of Mathematics, University of Dundee, Scotland, UK, 2002.
- [19] M. FUKUSHIMA, Z.-Q. LUO, AND P. TSENG, *A sequential quadratically constrained quadratic programming method for differentiable convex minimization*, SIAM J. Optim., 13 (2003), pp. 1098–1119.
- [20] W. W. HAGER, *Stabilized sequential quadratic programming*, Comput. Optim. Appl., 12 (1999), pp. 253–273.
- [21] W. W. HAGER AND M. S. GOWDA, *Stability in the presence of degeneracy and error estimation*, Math. Program., 85 (1999), pp. 181–192.
- [22] A. F. IZMAILOV, *Lagrange methods for finding degenerate solutions of conditional extremum problems*, Comput. Math. Math. Phys., 36 (1996), pp. 423–429.
- [23] A. F. IZMAILOV AND M. V. SOLODOV, *Optimality conditions for irregular inequality-constrained problems*, SIAM J. Control Optim., 40 (2001), pp. 1280–1295.
- [24] A. F. IZMAILOV AND M. V. SOLODOV, *The theory of 2-regularity for mappings with Lipschitzian derivatives and its applications to optimality conditions*, Math. Oper. Res., 27 (2002), pp. 614–635.
- [25] A. F. IZMAILOV AND M. V. SOLODOV, *Complementarity constraint qualification via the theory of 2-regularity*, SIAM J. Optim., 13 (2002), pp. 368–385.
- [26] A. F. IZMAILOV AND M. V. SOLODOV, *Superlinearly convergent algorithms for solving singular equations and smooth reformulations of complementarity problems*, SIAM J. Optim., 13 (2002), pp. 386–405.
- [27] A. F. IZMAILOV AND A. A. TRETYAKOV, *2-Regular Solutions of Nonlinear Problems, Theory and Numerical Methods*, Fizmatlit, Moscow, 1999 (in Russian).
- [28] C. KANZOW, *Strictly feasible equation-based methods for mixed complementarity problems*, Numer. Math., 89 (2001), pp. 135–160.
- [29] C. KANZOW AND H. D. QI, *A QP-free constrained Newton-type method for variational inequality problems*, Math. Program., 85 (1999), pp. 81–106.

- [30] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1984.
- [31] C. L. LAWSON AND R. J. HANSON, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ, 1974.
- [32] S. LEYFFER, *MacMPEC: AMPL Collection of MPECs*, Technical report, 2000.
- [33] D.-H. LI AND L. QI, *A Stabilized SQP Method via Linear Equations*, Applied mathematics technical report AMR00/5, The University of New South Wales, 2000.
- [34] Z.-Q. LUO, J.-S. PANG, AND D. RALPH, *Mathematical programs with equilibrium constraints*, Cambridge University Press, Cambridge, UK, 1996.
- [35] D. RALPH AND S. J. WRIGHT, *Superlinear convergence of an interior-point method despite dependent constraints*, Math. Oper. Res., 25 (2000), pp. 179–194.
- [36] H. SCHEEL AND S. SCHOLTES, *Mathematical programs with complementarity constraints: Stationarity, optimality and sensitivity*, Math. Oper. Res., 25 (2000), pp. 1–22.
- [37] S. SCHOLTES AND M. STÖHR, *Exact penalization of mathematical programs with equilibrium constraints*, SIAM J. Control Optim., 37 (1999), pp. 617–652.
- [38] S. SCHOLTES AND M. STÖHR, *How stringent is the linear independence assumption for mathematical programs with complementarity constraints?*, Math. Oper. Res., 26 (2001), pp. 851–863.
- [39] M. V. SOLODOV, *On the sequential quadratically constrained quadratic programming methods*, Math. Oper. Res., 29 (2004), pp. 64–79.
- [40] S. J. WRIGHT, *Superlinear convergence of a stabilized SQP method to a degenerate solution*, Comput. Optim. Appl., 11 (1998), pp. 253–275.
- [41] S. J. WRIGHT, *Modifying SQP for degenerate problems*, SIAM J. Optim., 13 (2002), pp. 470–497.
- [42] S. J. WRIGHT, *Constraint identification and algorithm stabilization for degenerate nonlinear programs*, Math. Program., 95 (2003), pp. 137–160.
- [43] S. J. WRIGHT, *An Algorithm for Degenerate Nonlinear Programming with Rapid Local Convergence*, Optimization technical report 03-02, Computer Sciences Department, University of Wisconsin, Madison, 2003.

PROX-METHOD WITH RATE OF CONVERGENCE $O(1/t)$ FOR VARIATIONAL INEQUALITIES WITH LIPSCHITZ CONTINUOUS MONOTONE OPERATORS AND SMOOTH CONVEX-CONCAVE SADDLE POINT PROBLEMS*

ARKADI NEMIROVSKI†

Abstract. We propose a prox-type method with efficiency estimate $O(\epsilon^{-1})$ for approximating saddle points of convex-concave $C^{1,1}$ functions and solutions of variational inequalities with monotone Lipschitz continuous operators. Application examples include matrix games, eigenvalue minimization, and computing the Lovasz capacity number of a graph, and these are illustrated by numerical experiments with large-scale matrix games and Lovasz capacity problems.

Key words. saddle point problem, variational inequality, extragradient method, prox-method, ergodic convergence

AMS subject classifications. 90C25, 90C47

DOI. 10.1137/S1052623403425629

1. Introduction. This paper is inspired by a recent paper of Nesterov [13] in which a new method for minimizing a nonsmooth Lipschitz continuous function f over a convex compact finite-dimensional set X is proposed. The characteristic feature of Nesterov’s method is that under favorable circumstances it exhibits a (*nearly*) *dimension-independent* $O(1/t)$ -rate of convergence: $f(x_t) - \min_X f \leq O(1/t)$, where x_t is the approximate solution built after t iterations. This is in sharp contrast with the results of information-based complexity theory, which state in particular (see [11]) that for a “black-box-oriented” method (one which operates with the values and subgradients of f only, without access to the “structure” of the objective) the number of function evaluations required to build an ϵ -solution when minimizing a Lipschitz continuous, with constant 1, function over an n -dimensional unit Euclidean ball cannot be less than $O(1/\epsilon^2)$, provided that $n \geq 1/\epsilon^2$. The explanation of the apparent “contradiction” between these approaches is that Nesterov’s method is *not* black-box-oriented; specifically, it is assumed that the objective function f is given as a cost function of the first player in a specific convex-concave game:

$$(1.1) \quad f(x) = \max_{y \in Y} \phi(x, y), \quad \phi(x, y) = g(x) + x^T A y + h^T y,$$

where Y is a convex compact set and g is a $C^{1,1}$ (i.e., with Lipschitz continuous gradient) convex function on X .¹ When solving the problem, we are given the structure of the objective, specifically, know X and Y , and are able (a) to compute the value and the gradient of g at a point and (b) to multiply a vector by A and A^T . The result of Nesterov states that if X and Y are simple enough (e.g., are unit Euclidean balls), then it is possible to minimize the objective (1.1) with accuracy ϵ in $O(1) \frac{L\|A\|}{\epsilon}$ steps,

*Received by the editors March 31, 2003; accepted for publication (in revised form) March 1, 2004; published electronically December 9, 2004.

<http://www.siam.org/journals/siopt/15-1/42562.html>

†Faculty of Industrial Engineering and Management, Technion–Israel Institute of Technology, Technion City, Haifa 32000, Israel (nemirovs@ie.technion.ac.il).

¹In fact, Nesterov allows the replacement of the linear-in- y component $h^T y$ with an arbitrary concave function $h(y)$; this, however, makes no difference, since the redefinition $y \leftarrow y^+ = (y, t)$, $Y \leftarrow Y^+ = \{(y, t) : \min_{y' \in Y} h(y') \leq t \leq h(y)\}$ allows us to make $h(\cdot)$ linear.

with a single computation of g, g' , two matrix-vector multiplications (one by A and one by A^T), and $O(\dim X + \dim Y)$ additional arithmetic operations per step; here L is the Lipschitz constant of g' with respect to the standard Euclidean norm, and $\|A\|$ is the standard matrix norm of A . Nesterov's method is based on *smoothing*—approximating f by a function with Lipschitz continuous gradient (the possibility of “computationally cheap” smoothing is given by representation (1.1)) and minimizing the approximation by a method with the rate of convergence $O(1/t^2)$.

In this paper, we propose another method for convex optimization with a nearly dimension-independent, under favorable circumstances, $O(1/t)$ -rate of convergence. Our “favorable circumstances” are essentially the same as in [13]; specifically, we assume that the objective function is given as the cost function of the first player in a game,

$$(1.2) \quad f(x) = \max_{y \in Y} \phi(x, y),$$

with $C^{1,1}$ convex-concave function ϕ ; additionally, both X and Y , as in Nesterov's construction, should be “simple enough” convex compact sets. In spite of similarity in the setup and in the resulting complexity bound, our method is completely different from that of Nesterov: instead of using smoothing and advanced techniques for minimizing $C^{1,1}$ functions, we directly apply to the underlying saddle point problem a simple prox-type method. Specifically, we reduce the problem of minimizing objective (1.2), i.e., the problem of approximating a saddle point of ϕ on $X \times Y$, to solving the associated variational inequality (v.i.):

$$(1.3) \quad \begin{aligned} &\text{find } z_* = (x_*, y_*) \in X \times Y : \langle \Phi(z_*), z - z_* \rangle \geq 0 \quad \forall z \in X \times Y, \\ &\Phi(x, y) = \begin{bmatrix} \frac{\partial}{\partial x} \phi(x, y) \\ -\frac{\partial}{\partial y} \phi(x, y) \end{bmatrix}. \end{aligned}$$

Note that since ϕ is convex-concave and $C^{1,1}$, the operator Φ is monotone and Lipschitz continuous:

$$\|\Phi(z) - \Phi(z')\|_* \leq L_{\|\cdot\|} \|\Phi\| \|z - z'\| \quad \forall z, z' \in Z \equiv X \times Y,$$

where $\|\cdot\|$ is a norm and $\|\cdot\|_*$ is the conjugate norm. We solve (1.3) by a prox-method:

$$(1.4) \quad \begin{aligned} &z_t = \text{solution to v.i. on } Z \\ &\text{with the operator } \Phi_t(z) = \gamma \Phi(z) + \omega'(z) - \omega'(z_{t-1}), \end{aligned}$$

where the “stepsize” γ is positive and $\omega(z)$ is a C^1 strongly convex function on Z :

$$\langle \omega'(z) - \omega'(z'), z - z' \rangle \geq \alpha \|z - z'\|^2 \quad \forall z, z' \in Z \quad [\alpha > 0].$$

It is easily seen that the prox-method converges ergodically to the set of solutions of (1.3): setting $(x^t, y^t) \equiv z^t = \frac{1}{t} \sum_{\tau=1}^t z_{\tau-1}$, one has

$$(1.5) \quad \begin{aligned} &\max_{y \in Y} \phi(x^t, y) - \min_{x \in X} \phi(x, y^t) \leq O(1) \frac{\Theta(z_0)}{\gamma \alpha t}, \\ &\Theta(z_0) = \max_{z \in Z} [\omega(z) - \omega(z_0) - \langle z - z_0, \omega'(z_0) \rangle]; \end{aligned}$$

note that (1.5) even does not exploit the fact that Φ is Lipschitz continuous. The rate of convergence in (1.5) is exactly what we are looking for; the difficulty, however, is

that method (1.4) is “conceptual” rather than implementable: a step requires solving a nontrivial variational inequality. Our central observation is that *in the case of Lipschitz continuous Φ , a step of the prox-method is easily implementable, provided that the stepsize γ is chosen properly*. Specifically, when $\gamma \leq \frac{\alpha}{\sqrt{2}L_{\|\cdot\|}[\Phi]}$, the *prox-mapping*

$$z \mapsto \mathcal{P}_t(z) = \operatorname{argmin}_{z' \in Z} [\omega(z') + \langle \gamma \Phi(z) - \omega'(z_{t-1}), z' \rangle] : Z \rightarrow Z$$

turns out to be a $\frac{1}{\sqrt{2}}$ -contraction: $\|\mathcal{P}_t(z) - \mathcal{P}_t(z')\| \leq \frac{1}{\sqrt{2}}\|z - z'\|$ for all $z, z' \in Z$, and z_t as given by (1.4) is exactly the fixed point of this contraction. It follows that *with a properly chosen stepsize, the prox-method becomes implementable: for all computational purposes, its step requires a small number of fixed point iterations* $z \mapsto \mathcal{P}_t(z)$, i.e., a small number (in fact, just two!) of computations of $\Phi(\cdot)$ and solving auxiliary problems of the form

$$(1.6) \quad \min_{z \in Z} [\omega(z) + \langle \xi, z \rangle].$$

As a result, *one can approximate within accuracy ϵ a saddle point of a $C^{1,1}$ convex-concave function* (or, more generally, a solution to a v.i. with monotone Lipschitz continuous operator) *at the cost* (the number of computations of the associated monotone operator and solving auxiliary problems (1.6)) *which is inversely proportional to ϵ* .

Note that both “building blocks” of our construction—the ergodic convergence properties of a prox-method (see [3, 10, 12, 4, 6, 16] and references therein) and the contraction properties of the prox-mapping—seem to be well studied. Surprisingly, the aforementioned complexity result, to the best of our knowledge, is new even in the standard Euclidean case (the one where $\omega(z) = \frac{1}{2}\langle z, z \rangle$). In this case, our scheme results in the *extragradient* method

$$z_{t-1}^+ = \pi_Z(z_{t-1} - \gamma \Phi(z_{t-1})), \quad z_t = \pi_Z(z_{t-1} - \gamma \Phi(z_{t-1}^+))$$

$$\left[\pi_Z(z) = \operatorname{argmin}_{z' \in Z} \langle z - z', z - z' \rangle \right]$$

proposed, from a different perspective, by Korpelevich as early as 1976 (see [7, 8]); here again the efficiency estimate $O(1/t)$ for the ergodic version $z^t = \frac{1}{t} \sum_{\tau=1}^t z_{\tau-1}^+$ seems to be new.

The rest of the paper is organized as follows. In section 2 we formulate the problem of interest—a variational inequality with Lipschitz continuous monotone operator—and investigate the possibility of solving the problem by a “conceptual” prox-method. In section 3 we present “implementable” versions of the method and carry out the complexity analysis of the resulting algorithms. In section 4, we present some extensions and modifications of the method. In section 5, we discuss a number of generic applications of our algorithms, specifically, solving matrix games, eigenvalue minimization, and computing the Lovasz capacity number of a graph. In particular, we demonstrate that our techniques allow us to compute the Lovasz capacity number ϑ of an n -node graph within accuracy $\epsilon\vartheta$, $\epsilon \in (0, 1]$, in no more than $O(n^4\epsilon^{-1}\sqrt{\ln(n)}\ln(n/\epsilon))$ arithmetic operations.² In concluding section 6, we present encouraging numerical results for matrix games with sparse matrices of sizes ranging up to $20,000 \times 20,000$ and for Lovasz capacity problems on random and Hamming graphs. (The largest graph we consider has 1024 nodes and 197120 arcs.)

²Note that if both the graph and its complement have $O(n^2)$ arcs, then the arithmetic cost of a *single* iteration of an interior point method as applied to computing Lovasz capacity is $O(n^6)$.

2. The problem and the conceptual prox-method.

2.1. The problem. Let Z be a convex compact set in Euclidean space E with inner product $\langle \cdot, \cdot \rangle$, let $\| \cdot \|$ be a norm on E (not necessarily the one associated with the inner product), and let $F : Z \rightarrow E$ be a Lipschitz continuous monotone mapping,

$$(2.1) \quad \begin{aligned} (a) \quad & \forall(z, z' \in Z) : \quad \|F(z) - F(z')\|_* \leq L\|z - z'\|, \\ (b) \quad & \forall(z, z' \in Z) : \quad \langle F(z) - F(z'), z - z' \rangle \geq 0, \end{aligned}$$

where $\| \cdot \|_*$ is the norm conjugate to $\| \cdot \|$:

$$\|\xi\|_* = \max_{z: \|z\| \leq 1} \langle \xi, z \rangle.$$

We are interested in approximating a solution to the v.i. associated with Z, F , i.e., a point $z_* \in Z$ such that

$$\langle F(z), z_* - z \rangle \leq 0 \quad \forall z \in Z.$$

(Note that since F is single-valued and continuous on Z , the latter relation is equivalent to $\langle F(z_*), z - z_* \rangle \geq 0$ for all $z \in Z$, which is the standard definition of a (strong) solution to the v.i. associated with Z, F .)

2.2. Prox-mapping—Preliminaries. Let us fix a continuously differentiable function $\omega(z) : Z \rightarrow \mathbf{R}$ which is strongly convex,

$$(2.2) \quad \langle \omega'(z) - \omega'(w), z - w \rangle \geq \alpha \|z - w\|^2 \quad \forall z, w \in Z \quad [\alpha > 0],$$

and is such that one can easily solve problems of the form

$$\min_{z \in Z} [\omega(z) + \langle e, z \rangle], \quad e \in E.$$

Let also

$$\Omega(\xi) = \max_{z \in Z} [\langle \xi, z \rangle - \omega(z)]$$

be the Legendre transformation of $\omega|_Z$, and let

$$(2.3) \quad \begin{aligned} \Omega_u(\xi) &= \Omega(\xi) - \langle \xi, u \rangle : E \rightarrow \mathbf{R}, \\ H_u(z) &= \Omega_u(\omega'(z)) : Z \rightarrow \mathbf{R}. \end{aligned}$$

Let us fix $\gamma > 0$. Given γ and $z \in Z$, let us define the *prox-mapping*

$$P_z : E \rightarrow Z : \quad P_z(\xi) = \operatorname{argmin}_{w \in Z} [\omega(w) + \langle w, \xi - \omega'(z) \rangle].$$

At least the first statement of the following lemma is well known.

LEMMA 2.1. *The mapping P_z is Lipschitz continuous; specifically,*

$$(2.4) \quad \|P_z(\xi) - P_z(\eta)\| \leq \alpha^{-1} \|\xi - \eta\|_* \quad \forall \xi, \eta \in E.$$

Besides this,

$$(2.5) \quad \forall(u \in Z) : H_u(P_z(\xi)) - H_u(z) \leq \langle \xi, u - P_z(\xi) \rangle + \underbrace{[\omega(z) + \langle \omega'(z), P_z(\xi) - z \rangle - \omega(P_z(\xi))]}_{\leq 0}.$$

Proof. Setting $v = P_z(\xi)$, $w = P_z(\eta)$, we have

$$(2.6) \quad \langle \omega'(v) - \omega'(z) + \xi, v - u \rangle \leq 0 \quad \forall u \in Z,$$

$$(2.7) \quad \langle \omega'(w) - \omega'(z) + \eta, w - u \rangle \leq 0 \quad \forall u \in Z.$$

Setting $u = w$ in (2.6) and $u = v$ in (2.7), we get

$$\langle \omega'(v) - \omega'(z) + \xi, v - w \rangle \leq 0, \quad \langle \omega'(w) - \omega'(z) + \eta, v - w \rangle \geq 0,$$

whence $\langle \omega'(w) - \omega'(v) + [\eta - \xi], v - w \rangle \geq 0$ or

$$\|\eta - \xi\|_* \|v - w\| \geq \langle \eta - \xi, v - w \rangle \geq \langle \omega'(v) - \omega'(w), v - w \rangle \geq \alpha \|v - w\|^2,$$

and (2.4) follows.

To prove (2.5), let $v = P_z(\xi)$. We have

$$\begin{aligned} & H_u(P_z(\xi)) - H_u(z) \\ &= \Omega_u(\omega'(v)) - \Omega_u(\omega'(z)) \\ &= \Omega(\omega'(v)) - \langle u, \omega'(v) \rangle - \Omega(\omega'(z)) + \langle u, \omega'(z) \rangle \\ &= \langle \omega'(v), v \rangle - \omega(v) - \langle u, \omega'(v) \rangle - \langle \omega'(z), z \rangle + \omega(z) + \langle u, \omega'(z) \rangle \\ &= \langle \omega'(v) - \omega'(z) + \xi, v - u \rangle + [\omega(z) + \langle \omega'(z), v - z \rangle - \omega(v)] + \langle \xi, u - v \rangle \\ &\leq \langle \xi, u - v \rangle + [\omega(z) + \langle \omega'(z), v - z \rangle - \omega(v)] \end{aligned}$$

(we have used (2.6)), as required in (2.5). $\square$

2.3. The prox-method. The *prox-method* with general prox-term (“with Bregman distances”) for convex minimization and solving variational inequalities has been investigated in many papers (see [4, 16, 6, 1] and references therein). As applied to the v.i. associated with Z, F , the (slightly modified; see below) method in its conceptual form is as follows.

CPM [conceptual prox-method].

Initialization. Choose starting point $z_0 \in Z$.

Step t , $t = 1, 2, \dots$: Given z_{t-1} , check whether

$$(2.8) \quad P_{z_{t-1}}(F(z_{t-1})) = z_{t-1}.$$

If it is the case, claim that z_{t-1} is the solution to the variational inequality associated with Z, F and terminate. Otherwise choose $\gamma_t > 0$ and a point $w_t \in Z$ such that

$$(2.9) \quad \begin{aligned} & \langle w_t - P_{z_{t-1}}(\gamma_t F(w_t)), \gamma_t F(w_t) \rangle \\ & + [\omega(z_{t-1}) + \langle \omega'(z_{t-1}), P_{z_{t-1}}(\gamma_t F(w_t)) - z_{t-1} \rangle - \omega(P_{z_{t-1}}(\gamma_t F(w_t)))] \leq 0. \end{aligned}$$

Set

$$\begin{aligned} z_t &= P_{z_{t-1}}(\gamma_t F(w_t)), \\ z^t &= \left(\sum_{\tau=1}^t \gamma_\tau \right)^{-1} \sum_{\tau=1}^t \gamma_\tau w_\tau, \end{aligned}$$

and pass to step $t + 1$.

Note: z^t is the approximate solution built in course of t steps.

Note that the standard way to ensure (2.9) is to define w_t as the solution to the v.i. with the monotone operator $F_{t-1}(z) = \omega'(z) - \omega'(z_{t-1}) + \gamma_t F(z)$, which, as it is immediately seen, results in $w_t = P_{z_{t-1}}(\gamma_t F(w_t))$ and thus makes the left-hand side of (2.9) nonpositive. The resulting procedure takes the form

$$(2.10) \quad z_t = \text{solution to the v.i. given by } Z, F_{t-1},$$

and this is exactly what is called the “prox-method with Bregman distances” and what is considered in the aforementioned references (primarily in the case when F is the (sub)gradient mapping generated by a convex function). Note that rule (2.10) is not “directly implementable,” since it requires solving a nontrivial v.i., and thus the “standard” prox-method is conceptual only, as for the above “algorithm.”

The convergence properties of CPM are summarized in the following statement (which is very similar to numerous results on ergodic convergence of gradient-type algorithms for convex minimization and v.i.’s with monotone operators; see, e.g., [3, 10, 11, 12, 16]).

PROPOSITION 2.2. *Consider a “relaxed” version of CPM, where condition (2.9) is weakened to*

$$(2.11) \quad \begin{aligned} & \langle w_t - P_{z_{t-1}}(\gamma_t F(w_t)), \gamma_t F(w_t) \rangle + \omega(z_{t-1}) \\ & + \langle \omega'(z_{t-1}), P_{z_{t-1}}(\gamma_t F(w_t)) - z_{t-1} \rangle - \omega(P_{z_{t-1}}(\gamma_t F(w_t))) \leq \epsilon_t. \end{aligned}$$

(i) *If the relaxed algorithm terminates at a certain step N according to rule (2.8), then z_{N-1} is a solution to the v.i.*

$$(2.12) \quad \langle F(z_{N-1}), u - z_{N-1} \rangle \geq 0 \quad \forall u \in Z.$$

(ii) *If the relaxed algorithm does not terminate in the course of N steps, then*

$$(2.13) \quad \begin{aligned} \epsilon(z^N) & \equiv \max_{u \in Z} \langle F(u), z^N - u \rangle \leq \frac{\Theta(z_0) + \sum_{t=1}^N \epsilon_t}{\sum_{t=1}^N \gamma_t}, \\ \Theta(z_0) & = \max_{z \in Z} [\omega(z) - \omega(z_0) - \langle \omega'(z_0), z - z_0 \rangle]. \end{aligned}$$

If, in addition, the v.i. in question is associated with a convex-concave saddle point problem, i.e.,

- $E = E_x \times E_y$ for Euclidean spaces E_x, E_y ,
- $Z = X \times Y$ for convex compacts $X \subset E_x, Y \subset E_y$,
- $F(z) \equiv F(x, y) = \begin{bmatrix} f'_x(x, y) \\ -f'_y(x, y) \end{bmatrix}$ for a continuously differentiable function $f(x, y) : X \times Y \rightarrow \mathbf{R}$, which is convex in $x \in X$ and is concave in $y \in Y$,

then

$$(2.14) \quad \begin{aligned} \epsilon_f(z^N) & \equiv \left[\max_{y \in Y} f(x^N, y) - \min_{x \in X} \max_{y \in Y} f(x, y) \right] + \left[\max_{y \in Y} \min_{x \in X} f(x, y) - \min_{x \in X} f(x, y^N) \right] \\ & \leq \frac{\Theta(z_0) + \sum_{t=1}^N \epsilon_t}{\sum_{t=1}^N \gamma_t}, \end{aligned}$$

with $\Theta(z_0)$ given by (2.13).

Proof. (i) is evident, since $P_z(\xi) = z$ means exactly that $\langle \xi, u - z \rangle \geq 0$ for all $u \in Z$, by definition of $P_z(\cdot)$.

To prove (ii), observe that $z_t = P_{z_{t-1}}(\gamma_t F(w_t))$, whence for every $u \in Z$ and every $t \leq N$ one has

$$\begin{aligned} H_u(z_t) - H_u(z_{t-1}) &\leq \langle \gamma_t F(w_t), u - z_t \rangle + [\omega(z_{t-1}) + \langle \omega'(z_{t-1}), z_t - z_{t-1} \rangle - \omega(z_t)] \\ &\leq \langle \gamma_t F(w_t), u - w_t \rangle + \epsilon_t \end{aligned} \quad \begin{array}{l} \text{[by (2.5)]} \\ \text{[by (2.11)].} \end{array}$$

Thus,

$$(2.15) \quad \langle \gamma_t F(w_t), w_t - u \rangle \leq H_u(z_{t-1}) - H_u(z_t) + \epsilon_t.$$

Summing up the resulting inequalities over $t = 1, \dots, N$, we get

$$\begin{aligned} (2.16) \quad \forall (u \in Z) : \sum_{t=1}^N \gamma_t \langle F(w_t), w_t - u \rangle &\leq H_u(z_0) - H_u(z_N) + \sum_{t=1}^N \epsilon_t \\ &= [\langle z_0 - u, \omega'(z_0) \rangle - \omega(z_0)] + \underbrace{[\langle u - z_N, \omega'(z_N) \rangle + \omega(z_N)]}_{\leq \omega(u)} \\ &\leq \omega(u) - \langle \omega'(z_0), u - z_0 \rangle - \omega(z_0) + \sum_{t=1}^N \epsilon_t \leq \Theta(z_0) + \sum_{t=1}^N \epsilon_t. \end{aligned}$$

Now, to get (2.13), note that by monotonicity of F one has

$$\langle F(w_t), w_t - u \rangle \geq \langle F(u), w_t - u \rangle,$$

and thus (2.16) implies that

$$\left(\sum_{t=1}^N \gamma_t \right) \langle F(u), z^N - u \rangle = \sum_{t=1}^N \gamma_t \langle F(u), w_t - u \rangle \leq \Theta(z_0) + \sum_{t=1}^N \epsilon_t \quad \forall u \in Z,$$

whence

$$\epsilon(z^N) = \max_{u \in Z} \langle F(u), z^N - u \rangle \leq \frac{\Theta(z_0) + \sum_{t=1}^N \epsilon_t}{\sum_{t=1}^N \gamma_t},$$

as required in (2.13).

Now assume that the variational inequality in question comes from a convex-concave saddle point problem. Setting $\lambda_t = \gamma_t / \sum_{\tau=1}^N \gamma_\tau$, (2.16) gives the first inequality in the following computation (where $z_t = (x_t, y_t)$ is the trajectory of the method and $u = (x, y) \in Z = X \times Y$):

$$\begin{aligned}
\frac{\Theta(z_0) + \sum_{t=1}^N \epsilon_t}{\sum_{t=1}^N \gamma_t} &\geq \sum_{t=1}^N \lambda_t \langle F(z_t), z_t - u \rangle \\
&= \sum_{t=1}^N \lambda_t [\langle f'_x(x_t, y_t), x_t - x \rangle + \langle f'_y(x_t, y_t), y - y_t \rangle] \\
&\geq \sum_{t=1}^N \lambda_t [[f(x_t, y_t) - f(x, y_t)] + [f(x_t, y) - f(x_t, y_t)]] \\
&\quad [\text{since } f \text{ is convex in } x \text{ and concave in } y] \\
&= \sum_{t=1}^N \lambda_t [f(x_t, y) - f(x, y_t)] \\
&\geq f\left(\sum_{t=1}^N \lambda_t x_t, y\right) - f\left(x, \sum_{t=1}^N \lambda_t y_t\right) \\
&= f(x^N, y) - f(x, y^N).
\end{aligned}$$

The resulting inequality is valid for all $x \in X$, $y \in Y$, whence

$$\begin{aligned}
\max_{x \in X, y \in Y} [f(x^N, y) - f(x, y^N)] &= \left[\max_{y \in Y} f(x^N, y) - \min_{x \in X} \max_{y \in Y} f(x, y) \right] \\
&\quad + \left[\max_{y \in Y} \min_{x \in X} f(x, y) - \min_{x \in X} f(x, y^N) \right] \\
&\leq \frac{\Theta(z_0) + \sum_{t=1}^N \epsilon_t}{\sum_{t=1}^N \gamma_t},
\end{aligned}$$

as claimed. (Note that $\max_{y \in Y} \min_{x \in X} f(x, y) = \min_{x \in X} \max_{y \in Y} f(x, y)$, since X, Y are convex compact and f is convex-concave.) $\square$

Remark 2.3. From now on, we assume that the starting point z_0 for the conceptual prox-method is chosen as the minimizer of $\omega(\cdot)$ on Z . With this choice of z_0 , one has

$$\Theta(z_0) \leq V[\omega] \equiv \max_{z \in Z} \omega(z) - \min_{z \in Z} \omega(z).$$

3. From conceptual method to implementable algorithm. Our main observation is extremely simple, as follows.

(*) Let F satisfy (2.1), and let $\gamma \leq \frac{\alpha}{\sqrt{2}L}$. Then, for every $z \in Z$, the mapping

$$\mathcal{P}_z^\gamma : Z \rightarrow Z : \mathcal{P}_z^\gamma(w) = P_z(\gamma F(w))$$

is a contraction:

$$(3.1) \quad \|\mathcal{P}_z^\gamma(w) - \mathcal{P}_z^\gamma(w')\| \leq \frac{1}{\sqrt{2}} \|w - w'\| \quad \forall w, w' \in Z.$$

Indeed, the mapping in question is the superposition of the Lipschitz continuous, with constant γL , mapping $w \mapsto \gamma F(w) : (Z, \|\cdot\|) \rightarrow (E, \|\cdot\|_*)$ (see (2.1)) and Lipschitz continuous, with constant α^{-1} (by Lemma 2.1) mapping $\xi \mapsto P_z(\xi) : (E, \|\cdot\|_*) \rightarrow (Z, \|\cdot\|)$.

Observation (*) suggests the following implementation of step t in the conceptual prox-method.

[“Basic implementation”] *Given a point $z_{t-1} \in Z$ which does not satisfy (2.8), set*

$$(3.2) \quad \gamma_t = \gamma \equiv \frac{\alpha}{\sqrt{2}L};$$

build, starting with $w_{t,0} \equiv z_{t-1}$, the iterates

$$(3.3) \quad w_{t,s} = P_{z_{t-1}}(\gamma F(w_{t,s-1})), \quad s = 1, 2, \dots,$$

until the condition

$$(3.4) \quad \begin{aligned} &\langle \gamma F(w_{t,s-1}), w_{t,s-1} - w_{t,s} \rangle + \omega(z_{t-1}) \\ &+ \langle \omega'(z_{t-1}), w_{t,s} - z_{t-1} \rangle - \omega(w_{t,s}) \leq 0 \end{aligned}$$

is met. Denote by s_t the corresponding value of s and set

$$w_t = w_{t,s_t-1} \quad [\Rightarrow z_t = w_{t,s_t}].$$

Note that with the basic implementation we do ensure (2.11) with $\epsilon_t = 0$ (this requirement is nothing but (3.4)). On the other hand, the sequence $\{w_{t,s}\}_{s=0}^\infty$ rapidly converges to the fixed point of the mapping $\mathcal{P}_{z_{t-1}}^\gamma$:

$$\|w_{t,s+1} - w_{t,s}\| \leq \frac{1}{\sqrt{2}} \|w_{t,s} - w_{t,s-1}\|, \quad s = 1, 2, \dots$$

We are in the situation when $z_{t-1} \equiv w_{t,0}$ is not a fixed point of $\mathcal{P}_{z_{t-1}}^1$, whence, as it is immediately seen, $w_{t,0}$ is not a fixed point of the mapping $\mathcal{P}_{z_{t-1}}^\gamma$; it follows that the first term on the left-hand side of (3.4) rapidly converges to zero (as a geometric progression with the ratio $\sqrt{1/2}$), while the second term, as it is immediately seen, is at least $\frac{\alpha}{16} \|w_{t,1} - w_{t,0}\|^2$. Thus, we have the following.

(!) *For all computational purposes, the number of “inner iterations” (3.3) at a step t in the basic implementation of the prox-method can be treated as a moderate constant C , while the number $N(\epsilon)$ of steps sufficient to make the right-hand side of (2.13), (2.14) $\leq \epsilon$ can be bounded from above as*

$$(3.5) \quad N(\epsilon) \leq \frac{2LV[\omega]}{\alpha\epsilon}.$$

We are about to prove that in fact the number of “inner iterations” can be made as small as two. Note that in the first draft of this paper, this possibility of a “two-step implementation” was established only for the case of $C^{1,1}$ function $\omega(\cdot)$; since then, inspired by the latest related results of Nesterov [14], we have succeeded in extending this result to the general case.

LEMMA 3.1. *Let a nonempty set $U \subset Z$ be convex and closed, and let $z \in Z$. Let ξ, η be two points from E , and let $\gamma > 0$. Consider the points*

$$(3.6) \quad \begin{aligned} w &= \operatorname{argmin}_{y \in U} [\langle \gamma \xi - \omega'(z), y \rangle + \omega(y)], \\ z_+ &= \operatorname{argmin}_{y \in U} [\langle \gamma \eta - \omega'(z), y \rangle + \omega(y)]. \end{aligned}$$

Then for all $u \in U$ one has

$$\begin{aligned}
 (3.7) \quad & (a) \quad \|w - z_+\| \leq \alpha^{-1} \gamma \|\xi - \eta\|_*, \\
 & (b) \quad \langle \gamma \eta, w - u \rangle \leq H_u(z) - H_u(z_+) \\
 & \quad \quad \quad + \underbrace{\langle \gamma \eta, w - z_+ \rangle + [\omega(z) + \langle \omega'(z), z_+ - z \rangle - \omega(z_+)]}_{\delta}, \\
 & (c) \quad \delta \leq \underbrace{\langle \gamma(\eta - \xi), w - z_+ \rangle + [\omega(z) + \langle \omega'(z), w - z \rangle + \langle \omega'(w), z_+ - w \rangle - \omega(z_+)]}_{\epsilon}, \\
 & (d) \quad \epsilon \leq \alpha^{-1} \gamma^2 \|\xi - \eta\|_*^2 - \frac{\alpha}{2} [\|w - z\|^2 + \|w - z_+\|^2].
 \end{aligned}$$

Proof. (3.7.a): The verification is completely similar to that for (2.4) (which is nothing but the particular case of (3.7.a) corresponding to the case $U = Z$). For all $u \in U$ one has

$$\begin{aligned}
 (3.8) \quad & (a) \quad \langle \gamma \xi + \omega'(w) - \omega'(z), w - u \rangle \leq 0, \\
 & (b) \quad \langle \gamma \eta + \omega'(z_+) - \omega'(z), z_+ - u \rangle \leq 0.
 \end{aligned}$$

Applying (3.8.a) with $u = z_+$ and (3.8.b) with $u = w$, we get

$$\begin{aligned}
 \gamma \langle \xi, w - z_+ \rangle &\leq \langle \omega'(z) - \omega'(w), w - z_+ \rangle, \\
 \gamma \langle \eta, z_+ - w \rangle &\leq \langle \omega'(z) - \omega'(z_+), z_+ - w \rangle.
 \end{aligned}$$

Summing up these inequalities, we get

$$\gamma \langle \xi - \eta, w - z_+ \rangle \leq \langle \omega'(z_+) - \omega'(w), w - z_+ \rangle,$$

whence

$$-\gamma \|\xi - \eta\|_* \|w - z_+\| \leq \gamma \langle \xi - \eta, w - z_+ \rangle \leq \langle \omega'(z_+) - \omega'(w), w - z_+ \rangle \leq -\alpha \|w - z_+\|^2,$$

and we arrive at (3.7.a).

(3.7.b): Let $\phi(u) = [\langle \gamma \eta, w - u \rangle + H_u(z_+) - H_u(z)]$; an immediate computation demonstrates that

$$\phi(z_+) = \langle \gamma \eta, w - z_+ \rangle + [\omega(z) + \langle \omega'(z), z_+ - z \rangle - \omega(z_+)],$$

so that (3.7.b) is exactly the claim that $\phi(u) \leq \phi(z_+)$ for all $u \in U$. To see that this claim is true, note that $\phi(u) = -\langle \gamma \eta + \omega'(z_+) - \omega'(z), u \rangle + \text{const}$, so that (3.8.b) is a sufficient condition for $z_+ \in U$ to maximize $\phi(\cdot)$ on U .

(3.7.c): For $u \in U$ we have

$$\begin{aligned}
 \delta &\equiv \langle \gamma \eta, w - z_+ \rangle + [\omega(z) + \langle \omega'(z), z_+ - z \rangle - \omega(z_+)] \\
 &= \langle \gamma(\eta - \xi), w - z_+ \rangle + \langle \gamma \xi, w - z_+ \rangle + [\omega(z) + \langle \omega'(z), z_+ - z \rangle - \omega(z_+)] \\
 &\leq \langle \gamma(\eta - \xi), w - z_+ \rangle + \langle \omega'(z) - \omega'(w), w - z_+ \rangle + [\omega(z) + \langle \omega'(z), z_+ - z \rangle - \omega(z_+)] \\
 &\quad \quad \quad \text{[by (3.8.a) with } u = z_+ \text{]} \\
 &= \langle \gamma(\eta - \xi), w - z_+ \rangle + [\omega(z) + \langle \omega'(z), w - z \rangle + \langle \omega'(w), z_+ - w \rangle - \omega(z_+)],
 \end{aligned}$$

as required in (3.7.c).

(3.7.d): For $u \in U$ we have

$$\begin{aligned}
 \epsilon &\equiv \langle \gamma(\eta - \xi), w - z_+ \rangle + [\omega(z) + \langle \omega'(z), w - z \rangle + \langle \omega'(w), z_+ - w \rangle - \omega(z_+)] \\
 &\leq \gamma \|\eta - \xi\|_* \|w - z_+\| + [\omega(z) + \langle \omega'(z), w - z \rangle - \omega(w)] \\
 &\quad + [\omega(w) + \langle \omega'(w), z_+ - w \rangle - \omega(z_+)] \leq \alpha^{-1} \gamma^2 \|\eta - \xi\|_*^2 - \frac{\alpha}{2} [\|w - z\|^2 + \|w - z_+\|^2],
 \end{aligned}$$

where the concluding inequality is given by (3.7.a) and by the strong convexity of $\omega(\cdot)$. $\square$

THEOREM 3.2. *Assume that F satisfies (2.1). Then the basic implementation ensures relation (3.4) (and thus (2.11) with $\epsilon_t = 0$) in no more than two inner iterations. Thus, with the basic implementation, the error bounds (2.13) and (2.14) become, respectively,*

$$(3.9) \quad \epsilon(z^N) \leq \frac{\sqrt{2}\Theta(z_0)L}{\alpha N}$$

and

$$(3.10) \quad \epsilon_f(z^N) \leq \frac{\sqrt{2}\Theta(z_0)L}{\alpha N},$$

while the “computational price” of z^N does not exceed $2N$ computations of $F(\cdot)$ and solving $2N$ auxiliary problems (1.6).

Proof. Let (2.1) be the case. All we need to verify is that if, for a given t , (3.4) is not met with $s_t = 1$, it definitely is met with $s_t = 2$. Indeed, let us set $z = z_{t-1}$, $\xi = F(z)$, $w = w_{t,1}$, $\eta = F(w)$, $z_+ = w_{t,2}$, and $\gamma = \gamma_t$; let also $U = Z$. Comparing the description of the basic implementation and (3.6), we see that our (z, w, z_+, γ, U) are exactly as required by the premise of Lemma 3.1, so that by that lemma we have

$$\begin{aligned}
 &\langle \gamma F(w_{t,1}), w_{t,1} - w_{t,2} \rangle + [\omega(z_{t-1}) + \langle \omega'(z_{t-1}), w_{t,2} - z_{t-1} \rangle - \omega(w_{t,2})] \\
 &\equiv \langle \gamma F(w), w - z_+ \rangle + [\omega(z) + \langle \omega'(z), z_+ - z \rangle - \omega(z_+)] \\
 &\leq \alpha^{-1} \gamma^2 \|F(z) - F(w)\|_*^2 - \frac{\alpha}{2} [\|w - z\|^2 + \|w - z_+\|^2].
 \end{aligned}$$

The concluding quantity in this chain, by (2.1), is $\leq [\alpha^{-1} \gamma^2 L^2 - \frac{\alpha}{2}] \|z - w\|^2$, which is just zero by (3.2). Thus, condition (3.4) is indeed met after at most two inner iterations. $\square$

4. Extensions. *The case of a non-Lipschitzian monotone operator.* In fact, we can handle monotone operators of various degrees of continuity, as follows.

THEOREM 4.1. *Assume that F is monotone and satisfies the relation*

$$(4.1) \quad \|F(z) - F(z')\|_* \leq L \|z - z'\|^\sigma \quad \forall z, z' \in Z,$$

where $L < \infty$ and $\sigma \in [0, 1]$. Note that the case of $\sigma = 0$ corresponds to a bounded monotone operator.

Consider the algorithm

$$\begin{aligned}
 z_{t-1} &\mapsto w_t = P_{z_{t-1}}(\gamma_t F(z_{t-1})) \equiv \operatorname{argmin}_{u \in Z} [\langle \gamma_t F(z_{t-1}) - \omega'(z_{t-1}), u \rangle + \omega(u)], \\
 w_t &\mapsto z_t = P_{z_{t-1}}(\gamma_t F(w_t)) \equiv \operatorname{argmin}_{u \in Z} [\langle \gamma_t F(w_t) - \omega'(z_{t-1}), u \rangle + \omega(u)],
 \end{aligned}
 \tag{4.2}$$

with the stepsizes

$$(4.3) \quad \gamma_t = 0.7 L^{-1} \left(\frac{\Theta[\omega]}{t} \right)^{\frac{1-\sigma}{2}} \alpha^{\frac{1+\sigma}{2}}, \quad \Theta[\omega] = \max_{z_0 \in Z} \Theta(z_0),$$

and the following rule for generating approximate solutions:

$$z^t = \left(\sum_{\tau=\lfloor t/2 \rfloor}^t \gamma_\tau \right)^{-1} \sum_{\tau=\lfloor t/2 \rfloor}^t \gamma_\tau w_\tau.$$

Then for every N one has

$$(4.4) \quad \epsilon(z^N) \leq O(1)L \left(\frac{\Theta[\omega]}{\alpha N} \right)^{\frac{1+\sigma}{2}}$$

with an absolute constant $O(1)$. In the saddle point case, $\epsilon(z^N)$ in this bound can be replaced with $\epsilon_f(z^N)$.

Proof. Applying Lemma 3.1 with $U = Z$, we conclude that algorithm (4.2) guarantees (2.11) with

$$\begin{aligned} \epsilon_t &= \alpha^{-1} \gamma_t^2 \|F(z_{t-1}) - F(w_t)\|_*^2 - \frac{\alpha}{2} [\|w_t - z_{t-1}\|^2 + \|w_t - z_t\|^2] \\ &\leq \alpha^{-1} \gamma_t^2 L^2 \|z_{t-1} - w_t\|^{2\sigma} - \frac{\alpha}{2} \|w_t - z_{t-1}\|^2 \\ &\leq \sup_{d \geq 0} \left[\alpha^{-1} \gamma_t^2 L^2 d^\sigma - \frac{\alpha}{2} d \right], \end{aligned}$$

and the concluding quantity is $\leq \frac{\Theta[\omega]}{t}$ in view of (4.3). Applying (2.13) to the segment of iterations from $\lfloor N/2 \rfloor$ to N , we get

$$\epsilon(z^N) \leq \frac{\Theta[\omega] \left(1 + \sum_{t=\lfloor N/2 \rfloor}^N t^{-1} \right)}{\sum_{t=\lfloor N/2 \rfloor}^N \gamma_t} \leq O(1)L \left(\frac{\Theta[\omega]}{\alpha N} \right)^{\frac{1+\sigma}{2}}. \quad \square$$

“*Bundle*” version of algorithm (4.2). In this version, one keeps in memory and utilizes not only the latest value of F , but past values as well. The algorithm is as follows.

(A) At the beginning of step t , we have at our disposal previous iterates $z_0, z_1, \dots, z_{t-1}$, $w_1, w_2, \dots, w_{t-1}$ (the set of all these iterates is denoted by W_{t-1}) along with a set

$$U_{t-1} = \{u \in Z : h_j^{t-1}(u) \geq 0, i \in J_{t-1}\},$$

which intersects the relative interior of Z . Here $h_j^{t-1}(\cdot)$ are combinations, with non-negative coefficients, of the affine functions

$$\psi_v(u) \equiv \langle F(v), v - u \rangle, \quad v \in W_{t-1}.$$

(B) At step t , we set

$$\begin{aligned} w_t &= \operatorname{argmin}_{y \in U_{t-1}} [\langle \gamma_t F(z_{t-1}) - \omega'(z_{t-1}), y \rangle + \omega(y)], \\ z_t &= \operatorname{argmin}_{y \in U_{t-1}} [\langle \gamma_t F(w_t) - \omega'(z_{t-1}), y \rangle + \omega(y)], \\ \epsilon_t &= \langle \gamma_t (F(w_t) - F(z_{t-1})), w_t - z_t \rangle \\ &\quad + [\omega(z_{t-1}) + \langle \omega'(z_{t-1}), w_t - z_{t-1} \rangle + \langle \omega'(w_t), z_t - w_t \rangle - \omega(z_t)]. \end{aligned} \quad (4.5)$$

(C) Applying Lemma 3.1 to $U = U_{t-1}$, $z = z_{t-1}$, $\xi = F(z_{t-1})$, $\eta = F(w_t)$, $\gamma = \gamma_t$, we get

$$(4.6) \quad \begin{aligned} \max_{u \in U_{t-1}} [\langle \gamma_t F(w_t), w_t - u \rangle - H_u(z_{t-1}) + H_u(z_t)] &\leq \epsilon_t \\ &\leq \alpha^{-1} \gamma_t^2 \|F(z_{t-1}) - F(w_t)\|_*^2 - \frac{\alpha}{2} [\|w_t - z_{t-1}\|^2 + \|w_t - z_t\|^2]. \end{aligned}$$

Recalling the structure of U_{t-1} , there exist nonnegative weights $\{\lambda_v^t\}_{v \in W_{t-1}}$ such that

$$\max_{u \in Z} \left[\langle \gamma_t F(w_t), w_t - u \rangle - H_u(z_{t-1}) + H_u(z_t) + \sum_{v \in W_{t-1}} \lambda_v^t \langle F(v), v - u \rangle \right] \leq \epsilon_t,$$

whence

$$(4.7) \quad \begin{aligned} \forall u \in Z : \langle \gamma_t F(w_t), w_t - u \rangle + \sum_{v \in W_{t-1}} \lambda_v^t \langle F(v), v - u \rangle &\leq H_u(z_{t-1}) - H_u(z_t) + \epsilon_t, \\ \epsilon_t &\leq \alpha^{-1} \gamma_t^2 \|F(z_{t-1}) - F(w_t)\|_*^2 - \frac{\alpha}{2} [\|w_t - z_{t-1}\|^2 + \|w_t - z_t\|^2]. \end{aligned}$$

We compute the weights $\{\lambda_v^t\}_{v \in W_{t-1}}$, set

$$(4.8) \quad \Gamma_t = \gamma_t + \sum_{v \in W_{t-1}} \lambda_v^t,$$

choose U_t in accordance with the above restrictions, and pass to step $t + 1$.

(D) The approximate solution z^t obtained in the course of t steps is

$$z^t = \left(\sum_{\tau=\lfloor t/2 \rfloor}^t \Gamma_\tau \right)^{-1} \sum_{\tau=\lfloor t/2 \rfloor}^t \left[\gamma_\tau w_\tau + \sum_{v \in W_{\tau-1}} \lambda_v^\tau v \right].$$

THEOREM 4.2. *Let F be monotone. For the outlined algorithm, one has*

$$(4.9) \quad \epsilon(z^N) \leq \underbrace{\left(\frac{\sum_{t=\lfloor N/2 \rfloor}^N \gamma_t}{\sum_{t=\lfloor N/2 \rfloor}^N \Gamma_t} \right)}_{\chi \leq 1} \frac{\Theta[\omega] + \sum_{\tau=\lfloor N/2 \rfloor}^N \epsilon_\tau}{\sum_{t=\lfloor N/2 \rfloor}^N \gamma_t}.$$

In particular, if F satisfies (4.1) and γ_t are chosen according to (4.3), bound (4.9) becomes

$$(4.10) \quad \epsilon(z^t) \leq O(1) \chi L \left(\frac{\Theta[\omega]}{\alpha t} \right)^{\frac{1+\sigma}{2}}.$$

In the case when F comes from a game with convex-concave cost function $f(x, y)$, the error measure $\epsilon(\cdot)$ in the above bounds can be replaced with $\epsilon_f(\cdot)$.

Proof. Error bound (4.9) can be derived from (4.7) exactly in the same fashion as the error bounds in Proposition 2.2 were derived from (2.15). Relation $\chi \leq 1$ follows from $\Gamma_t \geq \gamma_t$ (see (4.8) and note that $\lambda_v^t \geq 0$). As in the proof of Theorem 4.1, in the case of (4.1), (4.3) we have $\epsilon_t \leq t^{-1} \Theta[\omega]$, and therefore (4.9) implies (4.10). $\square$

Note that the bundle version of algorithm (4.2) is more computationally demanding than its prototype; as a compensation, it has a better theoretical efficiency estimate (since $\chi \leq 1$).

5. Examples. We are about to list a number of interesting particular cases of the above construction.

Euclidean setup. In this case, $\omega(z) = \frac{1}{2}\langle z, z \rangle \equiv \frac{1}{2}\|z\|_2^2$ and $\|\cdot\| = \|\cdot\|_2$, which results in $\alpha = 1$, $\Theta[\omega] = \frac{1}{2} \max_{z', z'' \in Z} \|z' - z''\|_2^2$. The basic implementation here becomes

$$(5.1) \quad \begin{aligned} (a) \quad & w_t = \Pi_Z(z_{t-1} - \gamma F(z_{t-1})), \\ (b) \quad & z_t = \Pi_Z(z_{t-1} - \gamma F(w_t)), \\ (c) \quad & z^t = \frac{1}{t} \sum_{\tau=1}^t w_\tau, \end{aligned}$$

where

- $\Pi_Z(u) = \operatorname{argmin}_{v \in Z} \|u - v\|_2$ is the projector onto Z ,
- $\gamma = \frac{1}{\sqrt{2}L}$, $L \geq \sup_{z, z' \in Z, z \neq z'} \frac{\|F(z) - F(z')\|_2}{\|z - z'\|_2}$.

The efficiency estimates (3.9) and (3.10) become, respectively,

$$(5.2) \quad \epsilon(z^N) \leq \frac{\sqrt{2}LD^2[Z]}{N}, \quad \epsilon_f(z^N) \leq \frac{\sqrt{2}LD^2[Z]}{N},$$

where $D[Z] = \max_{z, z' \in Z} \|z - z'\|_2$ is the $\|\cdot\|_2$ -diameter of Z .

Note that in order for the method to be implementable, the projector $\Pi_Z(\cdot)$ onto Z should be easy to compute; that is, Z should be “simple,” like a Euclidean ball, a box, or the standard simplex $\Delta_n = \{x \in \mathbf{R}^n : x \geq 0, \sum_x = 1\}$. A less trivial example is the “matrix box,” defined as follows:

- E is the space $\mathbf{S}^{(d_1, \dots, d_m)}$ of block-diagonal symmetric matrices x with $d_i \times d_i$ diagonal blocks $x[i]$, $i = 1, \dots, m$, equipped with the Frobenius inner product $\langle x, y \rangle = \operatorname{Tr}(xy)$;
- $Z = \{x : |x[i]|_\infty \leq a_i, i = 1, \dots, m\}$, where $|u|_\infty$ is the usual spectral norm (i.e., the maximal singular value) of a matrix.

Method (5.1.a–b) is nothing but the *extragradient method* for variational inequalities proposed by Korpelevich in 1976 (see [7, 8]; for recent results, see [15] and references therein); in particular, she proved the convergence of z_t to a solution of the v.i. The bounds (5.2) for the ergodic version of this method are, to the best of our knowledge, new.

Note that the rate of convergence exhibited by the prox-method is, in a sense, optimal. Specifically, given $D, L > 0$, consider the case when Z is the direct product of two n -dimensional Euclidean balls X, Y of radii D and $f(x, y) = y^T(Ax - b)$ for a symmetric operator A with operator norm not exceeding L , and b is such that the operator equation $Ax = b$ has a solution in X . The corresponding saddle point problem $\max_{y \in Y} f(x, y) = D\|Ax - b\|_2$ is associated with an affine Lipschitz continuous monotone operator with constant L with respect to $\|\cdot\| = \|\cdot\|_2$. Let $\mathcal{B}$ be an arbitrary “first order” method for solving saddle point problems of the outlined type, i.e., a method which is given b, X, Y in advance while having no direct access to A ; however, $\mathcal{B}$ has access to an “oracle” which, given an input vector x , returns the vector Ax . It turns out [9] that whenever $N \leq \frac{1}{4}n$ and $\mathcal{B}$ is a first order method which is allowed to make no more than N calls to the oracle, there is a problem in the family such that the result z obtained by $\mathcal{B}$ as applied to the problem satisfies the relations $\epsilon(z) \geq O(1)\frac{LD^2}{N}$, $\epsilon_f(z) \geq O(1)\frac{LD^2}{N}$. Thus, when Z is the direct product of two n -dimensional Euclidean balls and the dimension is large ($n > N$), bounds (5.2) cannot be improved by more than an absolute constant factor.

Spectahedron setup. In this case,

- $E = \mathbf{S}^{(d_1, \dots, d_m)}$;
- $Z = \{z \in E : z \succeq 0, \text{Tr}(z) = 1, z[i] \preceq a_i I_{d_i}, i = 1, \dots, m\}$;
- $\omega(\cdot)$ is the “regularized matrix entropy”

$$(5.3) \quad \begin{aligned} \omega(z) &= \sum_{i=1}^n (\lambda_i(z) + n^{-1}\delta) \ln(\lambda_i(z) + n^{-1}\delta), \\ n &= \sum_{\ell=1}^m k_\ell, \end{aligned}$$

where $\delta \in (0, 1]$ is a permanently fixed regularization parameter (say, $\delta = 10^{-16}$) and $\lambda_j(z)$, $j = 1, \dots, n \equiv d_1 + \dots + d_m$, are the eigenvalues of $z \in E$;

- $\|\cdot\|$ is the norm $|z|_1 = \sum_j |\lambda_j(z)|$ (so that $\|\cdot\|_*$ is the usual matrix norm $|z|_\infty$).

Note that in the case of $d_i = 1$, $i = 1, \dots, m$, Z is just the truncated simplex $\{x \in \mathbf{R}^m : 0 \leq x_i \leq b_i, \sum_i x_i = 1\}$.

It is known (see, e.g., [2]) that for the spectahedron setup one has

$$(5.4) \quad \alpha \geq \frac{1}{2}, \quad \Theta[\omega] \leq 4 \ln \left(\frac{n}{\delta} \right) = O(1) \ln n,$$

so that the efficiency estimates (2.13), (2.14) for the basic implementation become, respectively,

$$(5.5) \quad \epsilon(z^N) \leq O(1) \frac{L \ln(n)}{N}, \quad \epsilon_f(z^N) \leq O(1) \frac{L \ln(n)}{N},$$

where L is an a priori upper bound on the “1, ∞ ” Lipschitz constant

$$L_* = \sup_{z, z' \in Z, z \neq z'} \frac{|F(z) - F(z')|_\infty}{|z - z'|_1}$$

of the operator F .

Note that basically the only computation at a step of our prox-method, aside from computing F at a point, is computing $P_z(\xi)$ for given z and ξ , that is, solving the optimization problem of the form

$$(5.6) \quad \min_{z \in Z} \{\omega(z) + \langle p, z \rangle\}.$$

It is easily seen that solving the latter problem reduces to

- (1) computing the eigenvalue decompositions $p[i] = V_i \text{Diag}\{\pi_{i1}, \dots, \pi_{id_i}\} V_i^T$ of the diagonal blocks $p[i]$ of p ;
- (2) finding the solution ξ^* to the problem

$$(5.7) \quad \min_{\xi} \left\{ \sum_{i=1}^m \sum_{j=1}^{d_i} [(\xi_{ij} + n\delta^{-1}) \ln(\xi_{ij} + n\delta^{-1}) + \pi_{id_i} \xi_{ij}] : 0 \leq \xi_{ij} \leq a_i, \sum_{i=1}^m \sum_{j=1}^{d_i} \xi_{ij} = 1 \right\};$$

(3) recovering the optimal solution z^* to (5.6) as $z^*[i] = V_i \text{Diag} \{ \xi_{i1}^*, \dots, \xi_{id_i}^* \} V_i^T$, $i = 1, \dots, m$.

Now, (3) is easy. Step (2) also is easy: $\xi_{ij}^* = f_{ij}(\lambda_*)$, where

$$f_{ij}(\lambda) = \Pi_{[0, a_i]}(\underbrace{\exp\{\lambda - \pi_{ij}\} - n^{-1}\delta}_{s_{ij}(\lambda)})$$

is the point closest to $s_{ij}(\lambda)$ in the segment $[0, a_i]$, and λ_* is the root of the equation

$$(5.8) \quad \sum_i \sum_j f_{ij}(\lambda) = 1.$$

This equation can easily be solved by bisection; as a result, it takes just $O(n)$ operations to implement (2) within machine precision. Note also that when δ is less than machine zero, from the computer's viewpoint $f_{ij}(\lambda) = \min[a_i, \exp\{\lambda - \pi_{ij}\}]$; if, in addition, there is no truncation (i.e., $a_i = 1$ for all i), computing ξ_{ij}^* requires no bisection, since within machine precision one has

$$\xi_{ij}^* = \frac{\exp\{-\pi_{ij}\}}{\sum_r \sum_s \exp\{-\pi_{rs}\}}.^3$$

We see that the only “nontrivial” operation in solving (5.6) is (1). This operation is “numerically tractable” when the sizes d_i of blocks in $z \in Z$ are within a few hundreds, and is easy, when these sizes are small integers; in particular, with $d_i = O(1)$, solving (5.6) within machine precision costs just $O(d_1 + \dots + d_m)$ operations.

Mixed setups. In some applications (e.g., in those we are about to consider), E is the direct product of Euclidean spaces E_k , $k = 1, \dots, K$; Z is the direct product of “standard” sets $Z_k \subset E_k$ (simplexes, boxes, spectahedrons, etc.); and we know what are “good setup parameters” $\omega_k(\cdot), \|\cdot\|_k$ for every one of the factors. The question is how to “assemble” these entities into a “reasonably good” setup for a variational inequality on $Z = Z_1 \times \dots \times Z_K$. When answering this question, we assume that we know the quantities $\Theta_k = \Theta[\omega_k]$, the parameters α_k of strong convexity of $\omega_k(\cdot)|_{Z_k}$ with respect to $\|\cdot\|_k$, and upper bounds L_{pq} on “partial Lipschitz constants,”

$$L_{pq}^* = \max_{z, z' \in Z} \left\{ \frac{\|F_p(z) - F_p(z')\|_p^*}{\|z_q - z'_q\|_q} : z_\ell = z'_\ell, \ell \neq q, z_q \neq z'_q \right\},$$

of the monotone operator F in question. Here u_p is the natural projection of u on E_p , and $\|\cdot\|_k^*$ is the norm on E_k which is conjugate to $\|\cdot\|_k$. To simplify our considerations, assume that $L_{pq} = L_{qp}$.⁴ Let us look at “assemblings” of the form

(5.9)

$$\omega(z) = \sum_{k=1}^K \gamma_k \omega_k(z_k), \quad \|z\| = \sqrt{\sum_{k=1}^K \mu_k^2 \|z_k\|_k^2} \quad \left[\Leftrightarrow \|z\|_* = \sqrt{\sum_{k=1}^K \mu_k^{-2} [\|z_k\|_k^*]^2} \right],$$

³This explains the computational advantages of regularized entropy with δ of the order of machine zero as compared to other functions ensuring (5.4), like $\sum_i \lambda_i^p(x)$ with appropriately chosen p , specifically, $p = 1 + O(1)/\ln(n)$; with the latter choice of ω , computing $P_z(\cdot)$ would always require a bisection to solve the corresponding version of (5.8).

⁴Note that in the saddle point case ($Z = X \times Y$, Z_k are direct factors in either X or Y , and $F_k(z) = \epsilon_k(\nabla f(z))_k$ with $\epsilon_k = \pm 1$), which is the case of primary interest here, one has $L_{k\ell}^* = L_{\ell k}^*$, so that the assumption $L_{k\ell} = L_{\ell k}$ is fully justified.

where $\mu_k > 0$, $\gamma_k > 0$ are parameters of the construction. Note that auxiliary problems (1.6) associated with $Z = Z_1 \times \cdots \times Z_K$ and $\omega(\cdot)$ given by (5.9) are easy, provided that problems (1.6) associated with every pair $Z_k, \omega_k(\cdot)$ are easy. Further, we can easily express the (natural bounds on the) quantities $\Theta[\omega]$, α , and L (the Lipschitz constant L of F with respect to $\|\cdot\|$) in terms of μ, γ ; a straightforward computation demonstrates that

$$(5.10) \quad \Theta[\omega] \leq \tilde{\Theta} = \sum_k \gamma_k \Theta[\omega_k], \quad \alpha \geq \tilde{\alpha} = \min_k \frac{\gamma_k \alpha_k}{\mu_k^2}, \quad L \leq \tilde{L} = \|\mu_k^{-1} \mu_\ell^{-1} L_{k\ell}\|_\infty.$$

Now, what matters for the complexity bound of our method, the setup being given by $\omega(\cdot)$, $\|\cdot\|$, is the quantity $\alpha^{-1} \Theta L$ (the smaller its value, the better), and it is natural to look for the assembling which results in the smallest possible upper bound $\tilde{\alpha}^{-1} \tilde{\Theta} \tilde{L}$ on this quantity. This problem can be easily solved; an optimal solution is given by (5.9) with the parameters γ_k, μ_k given by

$$(5.11) \quad M_{k\ell} := L_{k\ell} \sqrt{\frac{\Theta_k \Theta_\ell}{\alpha_k \alpha_\ell}}, \quad \sigma_k := \frac{\sum_\ell M_{k\ell}}{\sum_{p,\ell} M_{p\ell}}, \quad \gamma_k = \frac{\sigma_k}{\Theta_k}, \quad \mu_k = \sqrt{\gamma_k \alpha_k}.$$

For the resulting assembling, one has

$$(5.12) \quad \tilde{\alpha} = \tilde{\Theta} = 1, \quad \tilde{L} = \sum_{k,\ell} L_{k\ell} \sqrt{\frac{\Theta_k \Theta_\ell}{\alpha_k \alpha_\ell}}.^5$$

By Theorem 3.2, the efficiency estimate for the resulting algorithm is

$$(5.13) \quad \epsilon(z^N) \leq \frac{\sqrt{2} \sum_{k,\ell} L_{k\ell} \sqrt{\frac{\Theta_k \Theta_\ell}{\alpha_k \alpha_\ell}}}{N}.$$

(In the saddle point case, $\epsilon(\cdot)$ can be replaced with $\epsilon_f(z^N)$.)

Example 1: Matrix game. Assume that we are interested in finding a saddle point of a bilinear function $x^T A y$, $x \in \mathbf{R}^p$, $y \in \mathbf{R}^q$, on the product Z of two standard simplexes $\Delta_p = \{x \geq 0 : \sum_i x_i = 1\}$ and $\Delta_q = \{y \geq 0 : \sum_j y_j = 1\}$. Consider the prox-method with mixed setup corresponding to spectahedron setups for the factors. The monotone operator in question is $F(x, y) = \begin{bmatrix} A y \\ -A^T x \end{bmatrix}$, so that $L_{11} = L_{22} = 0$, while $L_{12} = L_{21}$ is the Lipschitz constant of the mapping $x \mapsto A^T x$ considered as a mapping from $(\mathbf{R}^p, \|\cdot\|_1)$ to $(\mathbf{R}^q, \|\cdot\|_\infty \equiv \|\cdot\|_1^*)$, so that $L_{12} = L_{21} = \max_{i,j} |A_{ij}|$. Applying (5.4) and (5.13), we see that the algorithm yielded by the mixed setup obeys the efficiency estimate

$$(5.14) \quad \epsilon_f(z^N) \leq O(1) \frac{\max_{i,j} |A_{ij}| \sqrt{\ln(p+1) \ln(q+1)}}{N}.$$

(Recall that we treat the entropy regularization parameter δ as an absolute constant.) In the basic implementation of the resulting method, effort per step is dominated by

⁵The fact that (5.11) implies that $\tilde{\alpha} = \tilde{\Theta} = 1$ follows immediately from (5.10). To verify that $\lambda \equiv \sum_{k,\ell} L_{k\ell} \sqrt{\frac{\Theta_k \Theta_\ell}{\alpha_k \alpha_\ell}} \equiv \sum_{k,\ell} M_{k\ell} \geq \tilde{L} \equiv \|\mu_k^{-1} \mu_\ell^{-1} L_{k\ell}\|_\infty \equiv \|M_{k\ell} / \sqrt{\sigma_k \sigma_\ell}\|_\infty$, note that the matrix $\lambda \text{Diag}\{\sigma_1, \dots, \sigma_m\} - [M_{k\ell}]$ is diagonal-dominated and therefore is $\succeq 0$, whence $\lambda I_m - [M_{k\ell} / \sqrt{\sigma_k \sigma_\ell}] \succeq 0$ as well.

two computations of $F(\cdot)$, that is, by four matrix-vector multiplications (two by A and two by A^T).

Example 2: Semidefinite programming. The matrix game considered in Example 1 can be interpreted as follows: we are given p diagonal matrices A_i (the diagonal entries of A_i form the i th row in the game matrix A); the cost function of the first player is $\bar{f}(x) = \max_{y \in \Delta_q} x^T A y$ or, which is the same, is the largest eigenvalue of the matrix $\sum_i x_i A_i$, so that the first player is looking for a convex combination of the matrices A_i that has the smallest possible maximum eigenvalue. Now consider the latter problem in the case of symmetric matrices $A_i \in \mathbf{S}^{(d_1, \dots, d_m)}$. This problem again can be posed as a saddle point problem for the bilinear function $f(x, y) = \text{Tr}(y \sum_{i=1}^p x_i A_i)$, with x running through the simplex Δ_p and y running through the spectahedron $\Sigma^d = \{y \in \mathbf{S}^{(d_1, \dots, d_m)} : y \succeq 0, \text{Tr}(y) = 1\}$. Indeed, for $B \in \mathbf{S}^{(d_1, \dots, d_m)}$, the quantity $\max_{y \in \Sigma^d} \text{Tr}(yB)$ is exactly the maximum eigenvalue $\lambda_{\max}(B)$ of B . Thus, finding a saddle point of $f(x, y)$ on the direct product of $\Delta_p \times \Sigma^d$ is exactly the semidefinite program of minimizing the maximum eigenvalue of a convex combination of matrices $A_1, \dots, A_p$. Equipping both factors with the spectahedron setup, we get $\Theta_1 = O(1) \ln(p+1)$, $\Theta_2 = O(1) \ln(q+1)$, $q = d_1 + \dots + d_m$, $\alpha_1 = \alpha_2 = O(1)$. The associated monotone operator is $F(x, y) = \left[\begin{smallmatrix} (\text{Tr}(A_1 y), \dots, \text{Tr}(A_p y))^T \\ -\sum_i x_i A_i \end{smallmatrix} \right]$ so that $L_{11} = L_{22} = 0$ and $L_{12} = L_{21}$ is the Lipschitz constant of the mapping $x \mapsto \sum_{i=1}^p x_i A_i$ considered as a mapping from $(\mathbf{R}^p, \|\cdot\|_1)$ to $(\mathbf{S}^{(d_1, \dots, d_m)}, |\cdot|_\infty \equiv |\cdot|_1^*)$, so that $L_{12} = L_{21} = \max_{1 \leq i \leq p} |A_i|_\infty$. Applying (5.4) and (5.13), we see that the algorithm yielded by the mixed setup obeys the efficiency estimate

$$(5.15) \quad \epsilon_f(z^N) \leq O(1) \frac{\max_{1 \leq i \leq p} |A_i|_\infty \sqrt{\ln(p+1) \ln(q+1)}}{N}.$$

In the basic implementation of the resulting method, effort per step is dominated by two computations of $F(\cdot)$ and by two eigenvalue decompositions of matrices from $\mathbf{S}^{(d_1, \dots, d_m)}$.

Example 3: Computing Lovasz capacity. The Lovasz capacity number ϑ of a graph with vertices $1, \dots, n$ and an m -element set of arcs V is, by definition, the optimal value in the optimization problem

$$(5.16) \quad \min_{x \in X} \lambda_{\max}(d + x) = \min_{x \in X} \max_{y \in \Sigma_n} f(x, y), \quad f(x, y) = \text{Tr}([d + x]y),$$

where $\Sigma_n = \{x \in \mathbf{S}^n : x \succeq 0, \text{Tr}(x) = 1\}$,

$$d = \left[d_{ij} = \begin{cases} 0, & (i, j) \in V \\ 1, & (i, j) \notin V \end{cases} \right],$$

and X is the set of all symmetric $n \times n$ matrices $x = [x_{ij}]$ such that $x_{ij} = 0$ for $(i, j) \notin V$. Note that for an optimal x the matrix $\vartheta I - [d + x]$ is positive semidefinite, so that nonzero entries in x satisfy the bound $|x_{ij}| \leq \vartheta - 1$. It follows that if μ is a valid a priori upper bound on ϑ , then

$$(5.17) \quad \vartheta = \min_{x \in X_\mu} \max_{y \in \Sigma_n} \text{Tr}([d + x]y), \quad X_\mu = \{x \in X : |x_{ij}| \leq \mu - 1\}.$$

Consequently, we can approximate ϑ by our prox-method as applied to the (variational inequality associated with the) saddle point problem (5.17). Equipping X_μ with a

Euclidean setup, and Σ_n with the spectahedron setup, we have $\alpha_1 = 1$, $\alpha_2 \geq \frac{1}{2}$, $\Theta_1 = \frac{1}{2} \max_{x', x'' \in X_\mu} \|x' - x''\|_2^2 = 4m\mu^2$, where m is the number of arcs in the graph and $\Theta_2 = O(1) \ln(n+1)$. The monotone operator in question is $F(x, y) = \begin{bmatrix} y \\ -x - d \end{bmatrix}$, so that $L_{11} = L_{22} = 0$, while $L_{12} = L_{21}$ is the Lipschitz constant of the mapping $x \mapsto x$ considered as the mapping from $(\mathbf{S}^n, |\cdot|_2)$ to $(\mathbf{S}^n, |\cdot|_\infty \equiv |\cdot|_1^*)$, that is, $L_{12} = L_{21} = 1$. Applying (5.4) and (5.13), we see that the algorithm yielded by the mixed setup obeys the efficiency estimate

$$(5.18) \quad \epsilon_f(z^N) \leq O(1) \frac{\mu \sqrt{m \ln(n)}}{N}.$$

For the basic implementation, the effort per step is dominated by the necessity of finding eigenvalue decompositions of two matrices from $\mathbf{S}^n$, which requires $O(n^3)$ operations. Note that $\epsilon_f(x^N, y^N)$ is the difference of the quantities $\max_{y \in \Sigma_n} \text{Tr}([d + x^N]y) = \lambda_{\max}(d + x^N)$ (which is an upper bound on ϑ) and the easily computable quantity $\min_{x \in X_\mu} \text{Tr}([d + x]y^N)$ (which is a lower bound on ϑ). Computing these bounds, we localize the Lovasz capacity ϑ in a segment of the length $\epsilon_f(x^N, y^N)$.

We always can take $\mu = n$; with this choice, the bound (5.18) implies that, in order to approximate ϑ within absolute accuracy ϵ , it suffices to run $O(n\sqrt{m \ln(n)}\epsilon^{-1})$ steps of the algorithm. Since the arithmetic cost of a step is $O(n^3)$, we arrive at the overall complexity of $O(n^4\sqrt{m \ln(n)}\epsilon^{-1}) \leq O(n^5\sqrt{\ln(n)}\epsilon^{-1})$ operations. Note that a slight modification of the outlined construction (see section 6) allows us to approximate ϑ within relative accuracy ϵ (i.e., within absolute accuracy $\epsilon\vartheta$) in $O(n^3\sqrt{m \ln(n)}\epsilon^{-1}) \leq O(n^4\sqrt{\ln(n)}\epsilon^{-1})$ operations. It is instructive to compare these complexity estimates with those for interior point (IP) methods: for an n -node graph with no specific structure and the number arcs in both the graph and its complement of order of n^2 , the arithmetic cost of a *single* iteration in an IP method is $O(n^6)$.⁶ Thus, when approximating Lovasz capacity with fixed absolute (respectively, relative) accuracy, the prox-method outperforms the IP methods by factors of order of $O(n/\sqrt{\ln n})$ (respectively, $O(n^2/\sqrt{\ln n})$).

In the examples above, we have dealt with optimization problems which can be reformulated as saddle point problems for *biaffine* functions. An application which goes beyond this specific situation is minimization of the maximum of smooth (with Lipschitz continuous gradients) convex functions. Indeed, the problem

$$\min_{x \in X} \max_{1 \leq \ell \leq m} f_\ell(x)$$

is nothing but the saddle point problem

$$\min_{x \in X} \max_{y \in \Delta_m} f(x, y), \quad f(x, y) = \sum_{\ell=1}^m y_\ell f_\ell(x).$$

When f_ℓ are smooth on X , the convex-concave function $f(x, y)$ possesses a Lipschitz continuous gradient, and the associated saddle point problem can be solved by our prox-method.

⁶Indeed, under our assumptions, both the semidefinite program $\min_{\theta, x \in X} \{\theta : \theta I \succeq d + x\}$ defining ϑ and the dual of this program have $O(n^2)$ design variables, so that the cost of a Newton step is $O(n^6)$.

6. Numerical illustration. Below we present numerical results obtained with our prox-method on matrix games (Example 1) and the Lovasz capacity problem (Example 3), with the setups as described in the previous section. Our implementation was, essentially, the basic method with one modification: instead of always using fixed “theoretically safe” stepsizes $\gamma_t = \bar{\gamma} \equiv \frac{\alpha}{\sqrt{2}L}$, a simple policy for *on-line adjusting* the stepsizes was used. Specifically, whenever at step t the termination condition (3.4) was met after at most two inner iterations, the starting value of γ at the next step was taken to be 1.2 times larger than the final value of γ at step t . On the other hand, when in the course of a step the termination condition was not met during the first three inner iterations, the value of γ at every subsequent inner iteration was reduced according to $\gamma \mapsto \max[\gamma/2, \bar{\gamma}]$. The regularization parameter δ in (5.3) was set to 1.e–16.

Matrix games. We dealt with “square” matrix games $\min_{x \in \Delta_p} \max_{y \in \Delta_p} f_A(x, y)$, $f_A(x, y) = x^T A y$, with sparse matrices A generated as follows. We first chose at random cells ij with nonzero entries A_{ij} , with a given probability κ of choosing a particular cell. In the chosen cells, the values of A_{ij} were picked at random from the uniform distribution on $[-1, 1]$. Periodically we measured the “actual accuracy”

$$\epsilon_f(z^t) = \max_{y \in \Delta_q} (x^t)^T A y - \min_{x \in \Delta_p} x^T A y^t = \max_{1 \leq j \leq q} (A^T x^t)_j - \min_{1 \leq i \leq p} (A y^t)_i$$

of current approximate solution $z^t = (x^t, y^t)$.

Note that with $|A_{ij}| \leq 1$, the efficiency estimate (5.14) implies that

$$(6.1) \quad N(\epsilon) \equiv \min\{N : \epsilon_f(x^N, y^N) \leq \epsilon\} \leq \frac{16 \ln(2p)}{\epsilon}.$$

The numerical results are presented in Table 1. We see that, qualitatively speaking, the actual performance of the algorithm obeys the complexity bound (6.1) (see the values of the products $\#F \cdot \epsilon$). At the same time, the “empirical” constants C in the bound $\epsilon_f(x^t, y^t) \approx \frac{C}{t}$ are essentially better than the constant $16 \ln(2p)$ in (6.1); moreover, these constants seem to decrease as p grows. This phenomenon is in full accordance with the surprisingly nice behavior of our simple on-line adjustment of the stepsizes: the “theoretically safe” value for them is about 0.125, while the averaged, over t , empirical value varies from ≈ 3 for the smallest example to ≈ 400 for the largest one. Note that even those “large” values of γ_t still allow us to keep the average number of computations of F per step well below three. Of course, we do not pretend to consider our experiments as conclusive—the nice picture we observe perhaps reflects specific features of randomly generated game problems.

Computing Lovasz capacity. In order to work with the “meaningful” upper bound μ on ϑ , the execution was split into subsequent *stages*. At every stage, we applied the method to the saddle point problem (5.17) with the upper bound on the Lovasz capacity ϑ , coming from the previous stage, in the role of μ (at the very first stage, μ was set to the number n of nodes). A stage was terminated when the current upper bound on ϑ became $< \mu/2$. It is easily seen that, with this implementation of the prox-method, the total number of steps required to approximate ϑ within *relative* accuracy ϵ is, up to logarithmic factors, $O(1)n^3\sqrt{m} \leq O(1)n^4$, where m is the number

TABLE 1

Experiments with matrix games. $p = \dim x = \dim y$; κ is the density of nonzeros in the game matrix; $\epsilon = \epsilon_f(x^t, y^t)$; $\epsilon_1 = \epsilon_f(x_0, y_0)$; $\#F$ is the total number of computations of $F(\cdot)$ in the course of steps $1, \dots, t$; $\bar{\gamma} = t^{-1} \sum_{\tau=1}^t \gamma_\tau$. MATLAB code was run on a Pentium IV 1.3 GHz PC.

p κ		$t=1$	$t=32$	$t=64$	$t=128$	$t=256$	$t=512$	$t=1024$	$t=2048$	CPU sec
100 1.0	ϵ	2.5e-1	3.7e-2	1.6e-2	7.6e-3	3.7e-3	1.8e-3	8.6e-4	4.3e-4	12
	$\frac{\epsilon}{\epsilon_1}$	1.0	1.4e-1	6.2e-2	3.0e-2	1.5e-2	7.0e-3	3.4e-3	1.7e-3	
	$\#F$	2	72	144	296	592	1184	2376	4752	
	$\#F \cdot \epsilon$	5.1e-1	2.6	2.3	2.3	2.2	2.1	2.0	2.0	
	$\bar{\gamma}$	0.5	2.7	3.0	3.2	3.2	3.3	3.3	3.3	
500 0.2	ϵ	7.4e-2	1.4e-2	5.1e-3	2.2e-3	1.0e-3	4.8e-4	2.4e-4	1.2e-4	62
	$\frac{\epsilon}{\epsilon_1}$	1.0	1.9e-1	7.0e-2	3.0e-2	1.4e-2	6.5e-3	3.2e-3	1.6e-3	
	$\#F$	2	68	145	293	589	1185	2373	4753	
	$\#F \cdot \epsilon$	1.5e-1	9.7e-1	7.5e-1	6.5e-1	6.0e-1	5.7e-1	5.7e-1	5.7e-1	
	$\bar{\gamma}$	0.5	8.3	11.5	13.0	13.9	14.3	14.5	14.6	
1000 0.1	ϵ	4.0e-2	9.2e-3	2.9e-3	1.2e-3	5.5e-4	2.7e-4	1.3e-4	6.5e-5	144
	$\frac{\epsilon}{\epsilon_1}$	1.0	2.3e-1	7.3e-2	2.9e-2	1.4e-2	6.8e-3	3.3e-3	1.6e-3	
	$\#F$	2	68	140	292	588	1180	2372	4748	
	$\#F \cdot \epsilon$	8.0e-2	6.3e-1	4.1e-1	3.4e-1	3.2e-1	3.2e-1	3.2e-1	3.1e-1	
	$\bar{\gamma}$	0.5	14.3	23.4	27.2	28.4	29.2	29.6	29.9	
10,000 5.0e-3	ϵ	3.3e-3	2.9e-3	7.9e-4	2.7e-4	9.5e-5	3.8e-5	1.5e-5	6.6e-6	1607
	$\frac{\epsilon}{\epsilon_1}$	1.0	8.8e-1	2.4e-1	8.2e-2	2.9e-2	1.2e-2	4.6e-3	2.0e-3	
	$\#F$	2	64	138	285	578	1169	2355	4732	
	$\#F \cdot \epsilon$	6.6e-3	1.9e-1	1.1e-1	7.7e-2	5.5e-2	4.5e-2	3.5e-2	3.1e-2	
	$\bar{\gamma}$	0.5	26.6	106.4	135.6	181.3	241.3	289.4	335.9	
20,000 2.5e-3	ϵ	1.8e-3	1.8e-3	4.9e-4	1.5e-4	7.8e-5	2.9e-5	1.2e-5	5.3e-6	3566
	$\frac{\epsilon}{\epsilon_1}$	1.0	9.7e-1	2.7e-1	8.5e-2	4.3e-2	1.6e-2	6.9e-3	2.9e-3	
	$\#F$	2	64	139	288	578	1166	2342	4704	
	$\#F \cdot \epsilon$	3.6e-3	1.1e-1	6.7e-2	4.4e-2	4.5e-2	3.4e-2	2.9e-2	2.5e-2	
	$\bar{\gamma}$	0.5	26.6	175.0	256.3	251.2	333.4	383.4	445.7	

of arcs. Note also that every pair $(x, y) \in X \times Y$ (see (5.16)) produces a lower and an upper bound on ϑ . The upper bound, of course, is merely $\lambda_{\max}(d + x)$, while the lower bound is obtained as follows. As we remember, the x -component of a saddle point of $f(\cdot, \cdot)$ on $X \times \Sigma_n$ is a matrix from $X_{\vartheta-1}$, whence

$$\begin{aligned}
\vartheta &= \min_{x' \in X_{\vartheta-1}} \max_{y' \in \Sigma_n} \text{Tr}([d + x']y') \geq \min_{x' \in X_{\vartheta-1}} \text{Tr}([d + x']y) \\
&= \text{Tr}(dy) - (\vartheta - 1) \sum_{(i,j) \in V} |y_{ij}| \\
\Rightarrow \vartheta &\geq \frac{\text{Tr}(dy) + \sum_{(i,j) \in V} |y_{ij}|}{1 + \sum_{(i,j) \in V} |y_{ij}|}.
\end{aligned}$$

Running the method, we computed the outlined upper and lower bounds on ϑ given by iterates (x_t, y_t) and approximate solutions (x^t, y^t) , thus getting current “best observed” bounds. Computations were terminated when the difference between the “best observed so far” upper and lower bounds on ϑ became < 1 . We have run the method on six randomly generated graphs and six Hamming graphs; the results are presented in Tables 2 and 3.

TABLE 2
Experiments with Lovasz capacity and random graphs. n is the number of nodes, m is the number of arcs; $[\vartheta, \overline{\vartheta}]$ is the resulting localizer of ϑ ; $\#F$ is the total number of eigenvalue decompositions of $n \times n$ matrices in the course of running the method.

	Sizes of graph (n, m)				
	(50,616)	(100,2459)	(200,4918)	(300,11148)	(500,62230)
$[\vartheta, \overline{\vartheta}]$	[7.47, 8.46]	[10.08, 11.07]	[27.46, 28.42]	[34.17, 35.15]	[22.19, 23.17]
$\#F$	527	738	1003	3647	1867
CPU sec	6	52	515	6608	9580
					5215

TABLE 3
Experiments with Lovasz capacity and Hamming graphs. Nodes in $H_d^q(i_1, i_2, \dots)$ are all d -letter words in a q -element alphabet; two nodes are linked by an arc if and only if the Hamming distance between the corresponding words (the number of positions where the words differ from each other) belongs to the list $i_1, i_2, \dots$

	Graph and sizes (n, m)				
	$H_5^3(1)$ (243,1215)	$H_5^3(1, 2)$ (243,6075)	$H_6^3(1)$ (729,4374)	$H_6^3(1, 2)$ (729,26244)	$H_{10}^2(1, 2, 3, 4)$ (1024,197120)
$[\vartheta, \overline{\vartheta}]$	[80.59, 81.37]	[17.49, 18.13]	[242.58, 243.33]	[48.25, 49.16]	[17.37, 18.29]
$\#F$	424	458	616	683	1444
CPU sec	96	104	8752	8145	30154
					20084

REFERENCES

- [1] A. BECK AND M. TEBoulLE, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, OR Letters, 31 (2003), pp. 167–175.
- [2] A. BEN-TAL AND A. NEMIROVSKI, *Non-Euclidean restricted memory level method for large-scale convex optimization*, Math. Program., to appear.
- [3] R. BRUCK, *On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in Hilbert space*, J. Math. Anal. Appl., 61 (1977), pp. 159–164.
- [4] G. CHEN AND M. TEBoulLE, *Convergence analysis of a proximal-like minimization algorithm using Bregman functions*, SIAM J. Optim., 3 (1993), pp. 538–543.
- [5] G. COHEN, *Auxiliary problem principle and decomposition of optimization problems*, J. Optim. Theory Appl., 32 (1980), pp. 277–305.
- [6] K. C. KIWIEL, *Proximal minimization methods with generalized Bregman functions*, SIAM J. Control Optim., 35 (1997), pp. 1142–1168.
- [7] G. KORPELEVICH, *The extragradient method for finding saddle points and other problems*, Èkonom. i Mat. Metody, 12 (1976), pp. 747–756 (in Russian; English translation in Matekon).
- [8] G. KORPELEVICH, *Extrapolation gradient methods and their relation to modified Lagrange functions*, Èkonom. i Mat. Metody, 19 (1983), pp. 694–703 (in Russian; English translation in Matekon).
- [9] A. NEMIROVSKII, *Information-based complexity of linear operator equations*, J. Complexity, 8 (1992), pp. 153–175.
- [10] A. NEMIROVSKI AND D. YUDIN, *Cesari convergence of the gradient method for approximation saddle points of convex-concave functions*, Dokl. Akad. Nauk. SSSR, 239 (1978), pp. 1056–1059 (in Russian; English translation in Soviet Math. Dokl.).
- [11] A. NEMIROVSKI AND D. YUDIN, *Problem Complexity and Method Efficiency in Optimization*, John Wiley & Sons, New York, 1983.
- [12] A. NEMIROVSKI, *Efficient methods for solving variational inequalities*, Èkonom. i Mat. Metody, 17 (1981), pp. 344–359 (in Russian; English translation in Matekon).
- [13] YU. NESTEROV, *Smooth Minimization of Nonsmooth Functions*, CORE Discussion Paper 2003/12, 2003; available online at <http://www.core.ucl.ac.be/services/COREdp03.html>.
- [14] YU. NESTEROV, *Dual Extrapolation and Its Applications for Solving Variational Inequalities and Related Problems*, CORE Discussion Paper 2003/68, 2003; available online at <http://www.core.ucl.ac.be/services/COREdp03.html>.
- [15] M. A. NOOR, *New extragradient-type methods for general variational inequalities*, J. Math. Anal. Appl., 277 (2003), pp. 379–394.
- [16] M. TEBoulLE, *Convergence of proximal-like algorithms*, SIAM J. Optim., 7 (1997), pp. 1069–1083.

NONDIFFERENTIABLE MULTIPLIER RULES FOR OPTIMIZATION AND BILEVEL OPTIMIZATION PROBLEMS*

JANE J. YE[†]

Abstract. In this paper we study optimization problems with equality and inequality constraints on a Banach space where the objective function and the binding constraints are either differentiable at the optimal solution or Lipschitz near the optimal solution. Necessary and sufficient optimality conditions and constraint qualifications in terms of the Michel–Penot subdifferential are given, and the results are applied to bilevel optimization problems.

Key words. necessary optimality conditions, sufficient optimality conditions, constraint qualifications, bilevel optimization problems, Michel–Penot subdifferentials

AMS subject classifications. 49K10, 90C30, 91A65

DOI. 10.1137/S1052623403424193

1. Introduction. In this paper we study Lagrange multiplier rules and constraint qualifications (CQs) for the following optimization problem with equality and inequality constraints:

$$(P) \quad \begin{array}{ll} \min & f(x) \\ \text{s.t.} & g_i(x) \leq 0, \quad i = 1, 2, \dots, I, \\ & h_j(x) = 0, \quad j = 1, 2, \dots, J, \end{array}$$

where $f, g_i (i = 1, 2, \dots, I), h_j (j = 1, 2, \dots, J)$ are functions from a Banach space X to R and I, J are given integers. Generally one has $I \geq 1, J \geq 1$, but we allow I or $J = 0$ to signify the case in which there are no explicit constraints of the type. For any feasible solution $\bar{x}$ of problem (P), we denote by $I(\bar{x}) := \{i : g_i(\bar{x}) = 0\}$ the index set of the binding constraints.

The classical Lagrange multiplier rule (see, e.g., [4, 16]) usually requires that the objective function and the inequality constraints be Fréchet differentiable and the equality constraints be continuously differentiable. Most extensions of the classical Lagrange multipliers are given under two different assumptions: differentiability and Lipschitz continuity. On one hand, the classical multiplier rule was extended in the direction of eliminating the smoothness assumption while keeping the differentiability assumption such as in Halkin [9]. On the other hand, the classical multiplier rule was generalized in the direction of replacing the usual gradient by certain generalized gradients under Lipschitz assumptions such as in Rockafellar [22], Clarke [7], Michel and Penot [17, 18], Ioffe [11, 12], Mordukhovich [19], and Treiman [23, 24].

It is known that differentiability and Lipschitz continuity are two different kinds of assumptions and may not imply each other in general. Hence for nonlinear programming problems with mixed assumptions of differentiability and Lipschitz continuity, the only applicable optimality conditions in the literature were fuzzy multiplier

*Received by the editors May 12, 2003; accepted for publication (in revised form) March 25, 2004; published electronically December 9, 2004. The research of this paper is partially supported by NSERC and a University of Victoria internal research grant.

<http://www.siam.org/journals/siopt/15-1/42419.html>

[†]Department of Mathematics and Statistics, University of Victoria, P.O. Box 3045 STN CSC, Victoria, BC, V8W 3P4 Canada (janeye@math.uvic.ca).

rules for optimization problem with lower semicontinuous data (see, e.g., Borwein, Treiman, and Zhu [6] and Ngai and Théra [20]). Although in a finite-dimensional space the fuzzy multiplier rule reduces to an exact multiplier rule, it involves the singular subdifferential of the non-Lipschitz functions. Another issue involved is the size of the subdifferential. It is known that the Clarke generalized gradient of a differentiable function which is not strictly differentiable may contain other elements which are not the usual derivative. Our purpose is to provide an exact (not fuzzy) multiplier rule where the usual derivative (not the generalized gradient even if it is also Lipschitz continuous) is used when a function is differentiable and the generalized gradient is used when a function is not differentiable but Lipschitz continuous. Among various convex-valued generalized gradients which coincide with the usual derivative when a function is Gâteaux differentiable, including the B-generalized gradient of Treiman [23], the Michel–Penot (M-P) subdifferential is the smallest one, and hence we aim to provide a multiplier rule in terms of the M-P subdifferential. The multiplier rules in terms of other bigger generalized gradients follow immediately.

In Ye [27], under the mixed assumptions of Fréchet differentiability and Lipschitz continuity, and Fritz John and KKT Lagrange multiplier rules under generalized Mangasarian–Fromovitz, metric regularity and calmness CQs were given where the usual derivative is used when a function is differentiable.

In this paper we continue the study by considering the problem with mixed assumptions of Gâteaux differentiability, Fréchet differentiability, Hadamard differentiability (see, e.g., Definition 2.1), and Lipschitz continuity under other CQs that were not considered in [27]. Our main result includes the following generalized Lagrange multiplier rule, which summarizes the results obtained in Theorem 3.1 and Propositions 3.1–3.7.

THEOREM 1.1 (nondifferentiable KKT necessary optimality condition). *Let $\bar{x}$ be a local optimal solution of (P). Consider the following CQs at $\bar{x}$:*

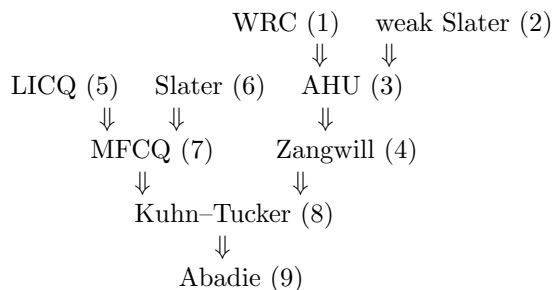
- (1) *the nondifferentiable weak reverse convex CQ as in Definition 3.12;*
- (2) *the nondifferentiable weak Slater CQ as in Definition 3.11;*
- (3) *the nondifferentiable Arrow–Hurwicz–Uzawa CQ as in Definition 3.10;*
- (4) *the generalized Zangwill CQ as in Definition 3.4;*
- (5) *the nondifferentiable linear independence CQ as in Definition 3.8;*
- (6) *the nondifferentiable Slater CQ as in Definition 3.9;*
- (7) *the nondifferentiable Mangasarian–Fromovitz CQ as in Definition 3.7;*
- (8) *the nondifferentiable Kuhn–Tucker CQ as in Definition 3.6;*
- (9) *the nondifferentiable Abadie CQ as in Definition 3.3.*

If f is either Gâteaux differentiable at $\bar{x}$ or Lipschitz near $\bar{x}$, then the KKT condition in terms of the M-P subdifferential holds at $\bar{x}$ under one of the CQs (1)–(4). If f is either Fréchet differentiable at $\bar{x}$ or Lipschitz near $\bar{x}$, then the KKT condition in terms of the M-P subdifferential holds at $\bar{x}$ under one of the CQs (1)–(9). That is, there exist scalars $\alpha_i \geq 0$ ($i \in I(\bar{x})$), $\beta_j \geq 0$, $\gamma_j \geq 0$ ($j = 1, 2, \dots, J$) such that

$$0 \in \partial^\diamond f(\bar{x}) + \sum_{i \in I(\bar{x})} \alpha_i \partial^\diamond g_i(\bar{x}) + \sum_{j=1}^J \beta_j \partial^\diamond h_j(\bar{x}) - \sum_{j=1}^J \gamma_j \partial^\diamond h_j(\bar{x}),$$

where $\partial^\diamond$ denotes the M-P subdifferential.

The relationships between the various constraint qualifications are given in the following diagram:



Note that Theorem 1.1 under CQ (7) was given in [27, Theorem 4.1]. The above KKT condition, however, provides a nondifferentiable KKT condition under all CQs (1)–(9). Moreover, the relationships between various CQs are given. Over the years, many papers have been devoted to extensions of classical CQs of type (5)–(7) to non-smooth optimization problems (see, e.g., [10, 14, 31]). To the best of the author's knowledge, CQs of type (1)–(4), (8)–(9) have never been extended to allow nondifferentiability in the literature. One of the purposes of this paper is to fill this gap since these nondifferentiable CQs are needed for studies of bilevel programming problems.

In Theorem 3.2 we also prove that the above KKT condition in terms of the M-P subdifferential becomes sufficient when the objective function is M-P pseudoconvex, the inequality constraints are M-P regular and quasiconvex, and the equality constraints are Gâteaux differentiable and quasilinear at the optimal solution $\bar{x}$.

In the last section of this paper we apply the results obtained to the bilevel optimization problem. One may reformulate the bilevel optimization as a single level optimization problem by using either the value function or the KKT condition for the lower level problem. The difficulty is that the usual CQ, such as the linear independence CQ, the Slater CQ, and the Mangasarian–Fromovitz CQ, does not hold for such a single level optimization problem. In this paper we show that the rest of the CQs (1), (3), (4), and (8)–(9) may hold for bilevel optimization problems. In particular, no CQ is required for the generalized linear bilevel optimization problem, which generalizes the known result that no CQ is needed for the linear bilevel programming problem. When the lower level problem is convex the relationship between the multiplier rule for the single level formulation by the value function approach and the one by the KKT approach is compared. It is found that the multiplier rule for the single level formulation by the value function approach is sharper than the one by the KKT approach.

We organize the paper as follows. In the next section, we provide preliminaries and preliminary results to be used in the rest of the paper. Section 3 is devoted to the discussion of CQs and the KKT necessary and sufficient optimality conditions. In section 4, applications to the bilevel optimization problem are given. In this paper unless otherwise specified, we denote by X a Banach space and by X^* its dual space equipped with the weak-star topology w^* . For $A \subseteq X$, we denote by $\text{co}A$, $\text{cl}A$ its convex hull and its closure, respectively. We denote by $B(v, \delta)$ the open ball centered at $v \in X$ with radius $\delta > 0$.

2. Preliminaries and preliminary results. We first recall some definitions of the usual derivatives.

DEFINITION 2.1 (usual derivatives). *Let X, Y be Banach spaces, let $\bar{x} \in X$, and let $f : X \rightarrow Y$. The usual directional derivative of f at $\bar{x}$ in the direction $v \in X$ is*

given by

$$f'(\bar{x}; v) := \lim_{t \downarrow 0} \frac{f(\bar{x} + tv) - f(\bar{x})}{t}$$

when this limit exists. f is said to be Gâteaux differentiable if there exists $Df(\bar{x})$, an element of the space $L(X, Y)$ of continuous linear functionals from X to Y such that for every $v \in X$, $f'(\bar{x}; v) = \langle Df(\bar{x}), v \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the canonic pairing. f is said to be Hadamard differentiable at $\bar{x}$ if $Df(\bar{x}) \in L(X, Y)$ and, for every $v \in X$,

$$\lim_{t \downarrow 0, v' \rightarrow v} \frac{f(\bar{x} + tv') - f(\bar{x})}{t} = \langle Df(\bar{x}), v \rangle.$$

f is said to be Fréchet differentiable at $\bar{x}$ if $Df(\bar{x}) \in L(X, Y)$ and the convergence in

$$f'(\bar{x}; v) := \lim_{t \downarrow 0} \frac{f(\bar{x} + tv) - f(\bar{x})}{t} = \langle Df(\bar{x}), v \rangle$$

is uniform with respect to v in bounded sets.

Remark 2.1. It is clear from the above definition that Fréchet differentiability is stronger than Hadamard differentiability, which in turn is stronger than Gâteaux differentiability.

DEFINITION 2.2 (M-P subdifferential). Let $\bar{x} \in X$ and let $f : X \rightarrow R$ be any function. The M-P directional derivative of f at $\bar{x}$ in the direction $v \in X$ introduced in [17] is given by

$$f^\diamond(\bar{x}; v) := \sup_{w \in X} \limsup_{t \downarrow 0} \frac{f(\bar{x} + t(v + w)) - f(\bar{x} + tw)}{t},$$

and the M-P subdifferential of f at $\bar{x}$ is given by the set

$$\partial^\diamond f(\bar{x}) := \{x^* \in X^* : \langle x^*, v \rangle \leq f^\diamond(\bar{x}; v) \quad \forall v \in X\}.$$

The M-P subdifferential is a natural generalization of the Gâteaux derivative since it is known (see [17, Proposition 1.3]) that when a function f is Gâteaux differentiable at $\bar{x}$, $f^\diamond(\bar{x}; v) = f'(\bar{x}; v)$ and $\partial^\diamond f(\bar{x}) = \{Df(\bar{x})\}$. Moreover when a function f is convex, the M-P subdifferential coincides with the subdifferential in the sense of convex analysis.

Whenever the Clarke generalized directional derivative $f^\circ(\bar{x}; v)$ and the Clarke generalized gradient $\partial^\circ f(\bar{x})$ exist, one always has

$$f^\diamond(\bar{x}; v) \leq f^\circ(\bar{x}; v), \quad \partial^\diamond f(\bar{x}) \subseteq \partial^\circ f(\bar{x}).$$

Note that the above inequality and the inclusion may be strict even in the case when f is Lipschitz continuous. For example, the function

$$(1) \quad f(x) = \begin{cases} x^2 \sin \frac{1}{x}, & x \neq 0, \\ 0, & x = 0 \end{cases}$$

on R is Lipschitz near 0 and Fréchet differentiable at 0 with $Df(0) = 0$, and hence $\partial^\diamond f(0) = \{0\}$ and $f^\diamond(0; v) = f'(0; v) = 0$. However, $\partial^\circ f(0) = [-1, 1]$ and $f^\circ(0; v) = |v|$.

Similar to the Clarke regularity [7], the following regularity concept was introduced in [5] as semiregularity (see also [25]) for Lipschitz continuous functions. We now extend the definition to any functions so that a Gâteaux differentiable function is also M-P regular.

DEFINITION 2.3 (M-P regularity). *Let $f : X \rightarrow R$ be a function on X and let $\bar{x} \in X$. We say that f is M-P regular at $\bar{x}$ if the usual directional derivative $f'(\bar{x}; v)$ exists and $f'(\bar{x}; v) = f^\diamond(\bar{x}; v)$ for all $v \in X$.*

The following properties of the M-P directional derivative and the M-P subdifferential will be useful.

PROPOSITION 2.1 (see [17, 18, 5]). *Let X be a Banach space, let $\bar{x} \in X$, and let $f, g : X \rightarrow R$ be either Gâteaux differentiable at $\bar{x}$ or Lipschitz near $\bar{x}$. Then the following hold:*

- (i) *The function $v \rightarrow f^\diamond(\bar{x}; v)$ is finite, positively homogeneous, and subadditive on X .*
- (ii) *For any scalar λ , $\partial^\diamond(\lambda f)(\bar{x}) = \lambda \partial^\diamond f(\bar{x})$, and for every $v \in X$, $f^\diamond(\bar{x}; -v) = (-f)^\diamond(\bar{x}; v)$.*
- (iii) *$\partial^\diamond(f + g)(\bar{x}) \subseteq \partial^\diamond f(\bar{x}) + \partial^\diamond g(\bar{x})$ and $(f + g)^\diamond(\bar{x}; v) \leq f^\diamond(\bar{x}; v) + g^\diamond(\bar{x}; v)$ for all $v \in X$. The equalities hold if both f and g are M-P regular at $\bar{x}$.*
- (iv) *$\partial^\diamond f(\bar{x})$ is a nonempty, convex, weak*-compact subset of X^* , and for every v in X , one has $f^\diamond(\bar{x}; v) = \max\{\langle \xi^*, v \rangle : \xi^* \in \partial^\diamond f(\bar{x})\}$.*
- (v) *If $\bar{x}$ is a local minimum of f , then $0 \in \partial^\diamond f(\bar{x})$ and $f^\diamond(\bar{x}; v) \geq 0$ for all $v \in X$.*

In [26, Proposition 3.1], it was shown that a Lipschitz function f is strictly differentiable if and only if both f and $-f$ are Clarke regular. Similarly we have the following conclusion.

PROPOSITION 2.2. *Let $f : X \rightarrow R$ be a function which is Lipschitz near $\bar{x} \in X$. Then f and $-f$ are both M-P regular if and only if f is Gâteaux differentiable at $\bar{x}$.*

Proof. It is obvious that if f is Gâteaux differentiable at $\bar{x}$, then both f and $-f$ are M-P regular. Now suppose that both f and $-f$ are M-P regular; then by (iii) of Proposition 2.1, one has

$$\partial^\diamond f(\bar{x}) + \partial^\diamond(-f)(\bar{x}) = \partial^\diamond(f - f)(\bar{x}) = \{0\},$$

which implies that $\partial^\diamond f(\bar{x})$ is a singleton since both $\partial^\diamond f(\bar{x})$ and $\partial^\diamond(-f)(\bar{x})$ are nonempty. Let $\partial^\diamond f(\bar{x}) = \{\xi\}$. Since $\xi \in X^*$, to prove that f is Gâteaux differentiable at $\bar{x}$ it suffices to prove that $f^\diamond(\bar{x}; v) = \langle \xi, v \rangle$ for each $v \in X$. By (iv) of Proposition 2.1, for each $v \in X$, $f^\diamond(\bar{x}; v) \geq \langle \xi, v \rangle$. By the Hahn–Banach theorem there exists $\xi' \in X^*$ majorized by $f^\diamond(\bar{x}; \cdot)$ and agreeing with $f^\diamond(\bar{x}; \cdot)$ at v . It follows that $\xi' \in \partial^\diamond f(\bar{x})$, and we have $f^\diamond(\bar{x}; v) = \langle \xi', v \rangle \geq \langle \xi, v \rangle$. If $\langle \xi, v \rangle$ were less than $f^\diamond(\bar{x}; v)$, then $\xi \neq \xi'$, contrary to the fact that $\partial^\diamond f(\bar{x}) = \{\xi\}$, and hence f is Gâteaux differentiable at $\bar{x}$. $\square$

Based on the M-P subdifferential, we extend the notions of pseudoconvexity and pseudoconcavity to allow nondifferentiability. For a definition of this kind of generalization to a class of generalized gradients, we refer the reader to [21].

DEFINITION 2.4 (M-P pseudoconvexity and pseudoconcavity). *Let f be a function defined on a Banach space X . f is said to be M-P pseudoconvex at $\bar{x} \in X$ if for all $x \in X$,*

$$f^\diamond(\bar{x}; x - \bar{x}) \geq 0 \Rightarrow f(x) \geq f(\bar{x}).$$

f is said to be M-P pseudoconcave at $\bar{x} \in X$ if for all $x \in X$,

$$f^\diamond(\bar{x}; x - \bar{x}) \leq 0 \Rightarrow f(x) \leq f(\bar{x}).$$

f is said to be M-P pseudoconvex (pseudoconcave) if it is M-P pseudoconvex (pseudoconcave) at all $x \in X$.

f is said to be M-P pseudoaffine if it is both M-P pseudoconvex and M-P pseudoconcave.

Remark 2.2. It is obvious that if *f* is Gâteaux differentiable at $\bar{x}$, then *f* is M-P pseudoconvex at $\bar{x}$ if and only if $-f$ is M-P pseudoconcave at $\bar{x}$. Using Proposition 2.1 it is easy to show that if *f* is Lipschitz near $\bar{x}$ and M-P pseudoconvex at $\bar{x}$, then $-f$ is M-P pseudoconcave at $\bar{x}$. However, the definitions for M-P pseudoconvexity and pseudoconcavity for a nondifferentiable function are not symmetric since M-P pseudoconcavity of *f* at x may not imply M-P pseudoconvexity of $-f$ at $\bar{x}$. For example, $\|x\|$ is both M-P pseudoconvex and pseudoconcave and hence M-P pseudoaffine, but $-\|x\|$ is M-P pseudoconcave but not M-P pseudoconvex.

As in the differentiable case we have the following necessary and sufficient optimality condition under the M-P pseudoconvexity.

THEOREM 2.1. *Let $\bar{x} \in X$ and *f* be M-P pseudoconvex at $\bar{x}$. Then $\bar{x}$ is a global minimum of the function *f*(*x*) if and only if $f^\diamond(\bar{x}; x - \bar{x}) \geq 0$ for all $x \in X$, i.e., $0 \in \partial^\diamond f(\bar{x})$.*

Proof. Assume that $\bar{x}$ is a global minimum of the function *f*(*x*); then for any $t \in (0, 1)$ one has

$$\frac{f(\bar{x} + t(x - \bar{x})) - f(\bar{x})}{t} \geq 0,$$

and hence

$$f^\diamond(\bar{x}; x - \bar{x}) \geq 0 \quad \forall x \in X.$$

Conversely if the above inequality holds, then by definition of M-P pseudoconvexity, one has $f(x) \geq f(\bar{x})$ and the proof is complete. $\square$

We now recall the definition for strictly quasiconvex (also referred to as semistrictly quasiconvex) functions and quasiconvex functions.

DEFINITION 2.5 (quasiconvexity and strict quasiconvexity). *Let *f* be a function defined on a Banach space *X*. *f* is said to be quasiconvex at $\bar{x} \in X$ if for all $x \in X$,*

$$f(x) \leq f(\bar{x}), 0 < \lambda < 1 \Rightarrow f((1 - \lambda)\bar{x} + \lambda x) \leq f(\bar{x}).$$

f is said to be strictly quasiconvex at $\bar{x} \in X$ if for all $x \in X$,

$$f(x) < f(\bar{x}), 0 < \lambda < 1 \Rightarrow f((1 - \lambda)\bar{x} + \lambda x) < f(\bar{x}).$$

f is said to be quasiconcave (strictly quasiconcave) at $\bar{x}$ if $-f$ is quasiconvex (strictly quasiconvex) at $\bar{x}$.

f is said to be (strictly) quasiconvex (quasiconcave) if it is (strictly) quasiconvex (quasiconcave) at all $x \in X$.

f is said to be quasilinear if it is both quasiconvex and quasiconcave.

We relate M-P pseudoconvex functions to strictly quasiconvex functions and quasiconvex functions in the following proposition, which can be proved similarly to the proof of [16, Theorem 9.5].

PROPOSITION 2.3. *Let *f* be a continuous and Gâteaux differentiable function on *X*. If *f* is M-P pseudoconvex (M-P pseudoconcave), then *f* is strictly quasiconvex (quasiconcave) and hence also quasiconvex (quasiconcave).*

3. Nondifferentiable multiplier rules and constraint qualifications. We first recall the notions of the contingent cone (also called the cone of tangents) and the cone of feasible directions.

DEFINITION 3.1 (contingent cone). *Let $\Omega \subseteq X$ and $\bar{x} \in \text{cl}\Omega$. The contingent cone of Ω at $\bar{x}$ is the closed cone defined by*

$$T_\Omega(\bar{x}) := \{v \in X : \exists t_n \downarrow 0, v_n \rightarrow v \text{ s.t. } \bar{x} + t_n v_n \in \Omega \forall n\}.$$

DEFINITION 3.2 (cone of feasible directions). *Let $\Omega \subseteq X$ and $\bar{x} \in \text{cl}\Omega$. The cone of feasible directions of Ω at $\bar{x}$ is the cone defined by*

$$D_\Omega(\bar{x}) := \{v \in X : \exists \delta > 0 \text{ s.t. } \bar{x} + tv \in \Omega \forall t \in (0, \delta)\}.$$

Based on the notions of contingent cone and M-P subdifferential, we extend the Abadie CQ introduced in [2] to our nondifferentiable setting.

DEFINITION 3.3 (nondifferentiable Abadie CQ). *Let $\bar{x} \in \Omega := \{x \in X : g_i(x) \leq 0, i = 1, 2, \dots, I, h_j(x) = 0, j = 1, 2, \dots, J\}$. We say that the nondifferentiable Abadie CQ holds at $\bar{x}$ if g_i ($i \in I(\bar{x})$) and h_j ($j = 1, 2, \dots, J$) are either Gâteaux differentiable at $\bar{x}$ or Lipschitz near $\bar{x}$, the convex cone generated by*

$$(2) \quad A := \bigcup_{i \in I(\bar{x})} \partial^\diamond g_i(\bar{x}) \cup \bigcup_{j=1}^J \partial^\diamond h_j(\bar{x}) \cup \bigcup_{j=1}^J [-\partial^\diamond h_j(\bar{x})]$$

is closed, and

$$\left. \begin{aligned} g_i^\diamond(\bar{x}; v) &\leq 0 \quad \forall i \in I(\bar{x}), \\ h_j^\diamond(\bar{x}; v) &= 0 \quad \forall j = 1, 2, \dots, J \end{aligned} \right\} \Rightarrow v \in T_\Omega(\bar{x}).$$

Based on the notions of the cone of feasible directions and the M-P subdifferential, we extend the Zangwill CQ introduced in [32] from inequality constraints to inequality and equality constraints in the nondifferentiable setting.

DEFINITION 3.4 (generalized Zangwill CQ). *Let $\bar{x} \in \Omega := \{x \in X : g_i(x) \leq 0, i = 1, 2, \dots, I, h_j(x) = 0, j = 1, 2, \dots, J\}$. We say that the generalized Zangwill CQ holds at $\bar{x}$ if g_i ($i \in I(\bar{x})$) and h_j ($j = 1, 2, \dots, J$) are either Gâteaux differentiable at $\bar{x}$ or Lipschitz near $\bar{x}$, the convex cone generated by the set A defined by (2) is closed, and*

$$\left. \begin{aligned} g_i^\diamond(\bar{x}; v) &\leq 0 \quad \forall i \in I(\bar{x}), \\ h_j^\diamond(\bar{x}; v) &= 0 \quad \forall j = 1, 2, \dots, J \end{aligned} \right\} \Rightarrow v \in \text{cl}D_\Omega(\bar{x}).$$

LEMMA 3.1. *Let Ω be a closed subset of X and let $f : X \rightarrow \mathbb{R}$ be either Gâteaux differentiable at $\bar{x}$ or Lipschitz near $\bar{x}$. If $\bar{x}$ is a local minimum of f over Ω , then*

$$(3) \quad f^\diamond(\bar{x}; v) \geq 0 \quad \forall v \in \text{cl}D_\Omega(\bar{x}).$$

Moreover if f is either Fréchet differentiable at $\bar{x}$ or Lipschitz near $\bar{x}$, then

$$(4) \quad f^\diamond(\bar{x}; v) \geq 0 \quad \forall v \in T_\Omega(\bar{x}).$$

Proof. We first show that (3) holds. Suppose there exists $v \in D_\Omega(\bar{x})$ such that $f^\diamond(\bar{x}; v) < 0$. Then

$$\limsup_{t \downarrow 0} \frac{f(\bar{x} + tv) - f(\bar{x})}{t} \leq f^\diamond(\bar{x}; v) < 0,$$

which implies that

$$f(\bar{x} + tv) - f(\bar{x}) < 0 \quad \forall t > 0 \text{ small enough.}$$

But this contradicts the fact that $\bar{x}$ is a local minimum of f over Ω , and hence

$$f^\diamond(\bar{x}; v) \geq 0 \quad \forall v \in D_\Omega(\bar{x}).$$

Consequently (3) follows from the continuity of $f^\diamond(\bar{x}; \cdot)$ (see (i) of Proposition 2.1).

Now suppose that there exists $v \in T_\Omega(\bar{x})$ such that $f^\diamond(\bar{x}; v) < 0$. Then there exist $r > 0, \epsilon > 0$ such that

$$f(\bar{x} + tv) - f(\bar{x}) \leq -rt \quad \forall t \in (0, \epsilon).$$

If f is Lipschitz near $\bar{x}$, then there exists $\delta > 0$ such that

$$f(\bar{x} + tv') - f(\bar{x} + tv) \leq L_f t \|v' - v\| \quad \forall v' \in B(v, \delta),$$

where L_f is the Lipschitz constant. By definition of the contingent cone, there exists $t_n \downarrow 0, v_n \rightarrow v$ such that $\bar{x} + t_n v_n \in \Omega$ for all n . Therefore for n large enough, one has

$$\|v_n - v\| < \frac{r}{2L_f}, \quad t_n \in (0, \epsilon)$$

and

$$\begin{aligned} & f(\bar{x} + t_n v_n) - f(\bar{x}) \\ &= f(\bar{x} + t_n v_n) - f(\bar{x} + t_n v) + f(\bar{x} + t_n v) - f(\bar{x}) \\ &\leq L_f t_n \|v_n - v\| - r t_n \\ &< -\frac{r}{2} t_n < 0. \end{aligned}$$

But this contradicts the fact that $\bar{x}$ is a local minimum of f over Ω , and hence (4) holds.

We omit the proof for (4) under the Fréchet differentiability assumption since it is a classical result (see, e.g., [13, Theorem 4.14]). $\square$

We now show that under the nondifferentiable Abadie CQ and the generalized Zangwill CQ, the KKT condition holds. It is interesting to note that although the nondifferentiable Abadie CQ is weaker than the generalized Zangwill CQ, the KKT condition under the nondifferentiable Abadie CQ requires stronger assumptions on the objective function.

THEOREM 3.1 (KKT conditions under Abadie CQ and Zangwill CQ). *Let $\bar{x}$ be a local optimal solution of (P). Under one of the following conditions,*

- (i) *the generalized Zangwill CQ holds at $\bar{x}$, and f is either Gâteaux differentiable at $\bar{x}$ or Lipschitz near $\bar{x}$;*
- (ii) *the nondifferentiable Abadie CQ holds at $\bar{x}$, and f is either Fréchet differentiable at $\bar{x}$ or Lipschitz near $\bar{x}$,*

the KKT condition holds at $\bar{x}$, i.e., there exist scalars $\alpha_i \geq 0$ ($i \in I(\bar{x})$), $\beta_j \geq 0$, $\gamma_j \geq 0$ ($j = 1, 2, \dots, J$) such that

$$0 \in \partial^\diamond f(\bar{x}) + \sum_{i \in I(\bar{x})} \alpha_i \partial^\diamond g_i(\bar{x}) + \sum_{j=1}^J \beta_j \partial^\diamond h_j(\bar{x}) - \sum_{j=1}^J \gamma_j \partial^\diamond h_j(\bar{x}).$$

Proof. Under the assumptions of the theorem, by Lemma 3.1 and Proposition 2.1 it is easy to show that $f^\diamond(\bar{x}; v) \geq 0$ for all v satisfying the following system:

$$\begin{aligned} g_i^\diamond(\bar{x}; v) &\leq 0 \quad \forall i \in I(\bar{x}), \\ h_j^\diamond(\bar{x}; v) &\leq 0 \quad \forall j = 1, 2, \dots, J, \\ (-h_j)^\diamond(\bar{x}; v) &\leq 0 \quad \forall j = 1, 2, \dots, J. \end{aligned}$$

Since g_i ($i \in I(\bar{x})$) are either Gâteaux differentiable at $\bar{x}$ or Lipschitz near $\bar{x}$ by (iv) of Proposition 2.1, v satisfies the above system if and only if $\max_{a \in A} \langle a, v \rangle \leq 0$, where A is the set defined by (2). Consequently,

$$f^\diamond(\bar{x}; v) \geq 0 \text{ whenever } \max_{a \in C} \langle a, v \rangle \leq 0,$$

where C denotes the convex cone generated by A . Thus the function $f^\diamond(\bar{x}; \cdot) + \delta_{C^0}(\cdot)$ attains its minimum at 0, where $C^0 := \{v \in X : \langle v, c \rangle \leq 0 \text{ for all } v \in C\}$ is the polar cone of C and δ_{C^0} is the indicator function of set C^0 . Since $f^\diamond(\bar{x}; \cdot)$ is convex and continuous by virtue of Proposition 2.1(i), by the sum rule (see, e.g., [7, Corollary 1, p. 105]), one has

$$0 \in \partial\phi(0) + \partial\delta_{C^0}(0),$$

where $\phi(\cdot) := f^\diamond(\bar{x}; \cdot)$ and $\partial\phi$ is the subdifferential in the sense of convex analysis. Since $\partial\phi(0) = \partial^\diamond f(\bar{x})$ and $\partial\delta_{C^0}(0) = C^{00} = C$ the above inclusion is the same as

$$0 \in \partial^\diamond f(\bar{x}) + C.$$

Therefore there exist some $\xi^* \in \partial^\diamond f(\bar{x})$, $\eta_i^* \in \partial^\diamond g_i(\bar{x})$ ($i \in I(\bar{x})$), $\zeta_j^*, \nu_j^* \in \partial^\diamond h_j(\bar{x})$ ($j = 1, 2, \dots, J$), $\alpha_i \geq 0$ ($i \in I(\bar{x})$), $\beta_j \geq 0$, $\gamma_j \geq 0$ ($j = 1, 2, \dots, J$) such that

$$0 = \xi^* + \sum_{i \in I(\bar{x})} \alpha_i \eta_i^* + \sum_{j=1}^J \beta_j \zeta_j^* - \sum_{j=1}^J \gamma_j \nu_j^*,$$

which implies that the KKT condition holds. $\square$

In the following theorem, we extend the classical KKT sufficient condition as given in [4, Theorem 4.3.8] to our nondifferentiable setting.

THEOREM 3.2 (nondifferentiable KKT sufficient condition). *Let $\bar{x}$ be a feasible solution of (P). Suppose that f, g_i ($i \in I(\bar{x})$), h_j ($j = 1, 2, \dots, J$) are either Gâteaux differentiable at $\bar{x}$ or Lipschitz near $\bar{x}$ and there exist scalars $\alpha_i \geq 0$ ($i \in I(\bar{x})$), β_j ($j = 1, 2, \dots, J$) such that*

$$(5) \quad 0 \in \partial^\diamond f(\bar{x}) + \sum_{i \in I(\bar{x})} \alpha_i \partial^\diamond g_i(\bar{x}) + \sum_{j=1}^J \beta_j \partial^\diamond h_j(\bar{x}).$$

Let $J^+ = \{j : \beta_j > 0\}$ and $J^- = \{j : \beta_j < 0\}$. Further suppose that f is M-P pseudoconvex at $\bar{x}$, g_i ($i \in I(\bar{x})$), h_j ($j \in J^+$), and $-h_j$ ($j \in J^-$) are M-P regular and quasiconvex at $\bar{x}$. Then $\bar{x}$ is a global optimal solution of (P).

Proof. Note that (5) is equivalent to the existence of

$$\begin{aligned} \xi^* &\in \partial^\diamond f(\bar{x}), \quad \eta_i^* \in \partial^\diamond g_i(\bar{x}) \quad (i \in I(\bar{x})), \\ \gamma_j^* &\in \partial^\diamond h_j(\bar{x}) \quad (j \in J^+), \quad \zeta_j^* \in \partial^\diamond (-h_j)(\bar{x}) \quad (j \in J^-) \end{aligned}$$

such that

$$(6) \quad 0 = \xi^* + \sum_{i \in I(\bar{x})} \alpha_i \eta_i^* + \sum_{j \in J^+} \beta_j \gamma_j^* - \sum_{j \in J^-} \beta_j \zeta_j^*.$$

Let x be any feasible solution of (P); then for any $i \in I(\bar{x})$,

$$g_i(x) \leq 0 = g_i(\bar{x}).$$

By the quasiconvexity of g_i at $\bar{x}$ it follows that

$$(7) \quad g_i(\bar{x} + \lambda(x - \bar{x})) = g_i(\lambda x + (1 - \lambda)\bar{x}) \leq g_i(\bar{x})$$

for all $\lambda \in (0, 1)$. This implies that

$$(8) \quad g_i^\diamond(\bar{x}; x - \bar{x}) = g'_i(\bar{x}; x - \bar{x}) \leq 0 \quad \forall i \in I(\bar{x})$$

by the M-P regularity. Similarly since $h_j(j \in J^+)$ and $-h_j(j \in J^-)$ are M-P regular and quasiconvex at $\bar{x}$, we have

$$(9) \quad h_j^\diamond(\bar{x}; x - \bar{x}) \leq 0 \quad \forall j \in J^+,$$

$$(10) \quad (-h_j)^\diamond(\bar{x}; x - \bar{x}) \leq 0 \quad \forall j \in J^-.$$

Note that (8)–(10) imply

$$(11) \quad \langle \eta_i^*, x - \bar{x} \rangle \leq 0 \quad \forall i \in I(\bar{x}),$$

$$(12) \quad \langle \gamma_j^*, x - \bar{x} \rangle \leq 0 \quad \forall j \in J^+,$$

$$(13) \quad \langle \zeta_j^*, x - \bar{x} \rangle \leq 0 \quad \forall j \in J^-.$$

Multiplying (11), (12), and (13) by $\alpha_i \geq 0$ ($i \in I(\bar{x})$), $\beta_j > 0$ ($j \in J^+$), and $-\beta_j > 0$ ($j \in J^-$), respectively, and adding we get

$$\left\langle \sum_{i \in I(\bar{x})} \alpha_i \eta_i^* + \sum_{j \in J^+} \beta_j \gamma_j^* - \sum_{j \in J^-} \beta_j \zeta_j^*, x - \bar{x} \right\rangle \leq 0.$$

By virtue of (6), the above inequality implies that

$$\langle \xi^*, x - \bar{x} \rangle \geq 0,$$

which implies by (iv) of Proposition 2.1 that

$$f^\diamond(\bar{x}; x - \bar{x}) \geq \langle \xi^*, x - \bar{x} \rangle \geq 0$$

since $\xi^* \in \partial^\diamond f(\bar{x})$. By the M-P pseudoconvexity of f at $\bar{x}$, we must have $f(x) \geq f(\bar{x})$, and the proof is complete. $\square$

We now extend the Kuhn–Tucker CQ introduced by Kuhn and Tucker in [15] to the nondifferentiable setting.

DEFINITION 3.5 (cone of attainable directions). *Let $\Omega \subseteq X$ and $\bar{x} \in \text{cl}\Omega$. We say that $v \in A_\Omega(\bar{x})$, the cone of attainable directions of Ω at $\bar{x}$ if there exist $\delta > 0$, and a mapping $\alpha : R \rightarrow X$ such that $\alpha(\tau) \in \Omega$ for all $\tau \in (0, \delta)$, $\alpha(0) = \bar{x}$, and $\lim_{\tau \downarrow 0} \frac{\alpha(\tau) - \alpha(0)}{\tau} = v$.*

The cone of attainable directions is also known as the adjacent cone (see, e.g., [1]) or the incident cone. In fact

$$A_{\Omega}(\bar{x}) = \liminf_{\tau \downarrow 0} \frac{\Omega - \bar{x}}{\tau}$$

and hence is a closed set.

DEFINITION 3.6 (nondifferentiable Kuhn–Tucker CQ). *Let $\bar{x} \in \Omega := \{x \in X : g_i(x) \leq 0, i = 1, 2, \dots, I, h_j(x) = 0, j = 1, 2, \dots, J\}$. We say the nondifferentiable Kuhn–Tucker CQ is satisfied at $\bar{x}$ if g_i ($i \in I(\bar{x})$) and h_j ($j = 1, 2, \dots, J$) are either Gâteaux differentiable at $\bar{x}$ or Lipschitz near $\bar{x}$, the convex cone generated by the set (2) is closed, and*

$$\left. \begin{aligned} g_i^{\diamond}(\bar{x}; v) &\leq 0 \quad \forall i \in I(\bar{x}), \\ h_j^{\diamond}(\bar{x}; v) &= 0 \quad \forall j = 1, 2, \dots, J \end{aligned} \right\} \Rightarrow v \in A_{\Omega}(\bar{x}).$$

It is easy to see that $\text{cl}D_{\Omega}(\bar{x}) \subseteq A_{\Omega}(\bar{x}) \subseteq T_{\Omega}(\bar{x})$, and hence the following relationship among the generalized Zangwill CQ, the nondifferentiable Kuhn–Tucker CQ, and the nondifferentiable Abadie CQ is obvious.

PROPOSITION 3.1. *The nondifferentiable Zangwill CQ implies the nondifferentiable Kuhn–Tucker CQ, and the nondifferentiable Kuhn–Tucker CQ implies the nondifferentiable Abadie CQ. That is,*

$$\text{Zangwill CQ} \implies \text{Kuhn–Tucker CQ} \implies \text{Abadie CQ}.$$

DEFINITION 3.7 (nondifferentiable Mangasarian–Fromovitz CQ). *Let $\bar{x}$ be a feasible solution of (P). We say that the nondifferentiable Mangasarian–Fromovitz CQ is satisfied if g_i ($i \in I(\bar{x})$) are either Hadamard differentiable at $\bar{x}$ or Lipschitz near $\bar{x}$, g_i ($i \notin I(\bar{x})$) are continuous at $\bar{x}$, h_j ($j = 1, 2, \dots, J$) are Fréchet differentiable at $\bar{x}$ and continuous in a neighborhood of $\bar{x}$, $\{Dh_1(\bar{x}), \dots, Dh_J(\bar{x})\}$ are linearly independent, and there exists $v \in X$ such that*

$$(14) \quad g_i^{\diamond}(\bar{x}; v) < 0 \quad \forall i \in I(\bar{x}),$$

$$(15) \quad \langle Dh_j(\bar{x}), v \rangle = 0 \quad \forall j = 1, 2, \dots, J.$$

LEMMA 3.2. *If g_i ($i \in I(\bar{x})$) are either Hadamard differentiable at $\bar{x}$ or Lipschitz near $\bar{x}$, g_i ($i \notin I(\bar{x})$) are continuous at $\bar{x}$, h_j ($j = 1, 2, \dots, J$) are Fréchet differentiable at $\bar{x}$ and continuous in a neighborhood of $\bar{x}$, then the nondifferentiable Mangasarian–Fromovitz CQ is equivalent to the nonexistence of $(\alpha, \beta) \in R_+^I \times R^J$ such that $(\alpha, \beta) \neq 0$ and*

$$(16) \quad 0 \in \sum_{i \in I(\bar{x})} \alpha_i \partial^{\diamond} g_i(\bar{x}) + \sum_{j=1}^J \beta_j Dh_j(\bar{x}).$$

Proof. We prove the lemma by contradiction.

Suppose the nondifferentiable Mangasarian–Fromovitz CQ holds but there exists a nonzero vector $(\alpha, \beta) \in R_+^I \times R^J$ such that

$$(17) \quad 0 = \sum_{i \in I(\bar{x})} \alpha_i \xi_i^* + \sum_{j=1}^J \beta_j Dh_j(\bar{x})$$

for some $\xi_i^* \in \partial^\diamond g_i(\bar{x})$, $i \in I(\bar{x})$. Since the vectors $Dh_j(\bar{x})$ are linearly independent, at least one α_i is nonzero. By (17), for v which is a solution of (14), (15),

$$\sum_{i \in I(\bar{x})} \alpha_i \langle \xi_i^*, v \rangle = - \sum_{j=1}^J \beta_j \langle Dh_j(\bar{x}), v \rangle.$$

But this is impossible since the right-hand side of the equation is zero while the left-hand side of the equation is nonzero. Therefore the nondifferentiable Mangasarian–Fromovitz CQ implies that there is no nonzero vector $(\alpha, \beta) \in R_+^I \times R^J$ such that (16) holds.

Conversely suppose that there is no nonzero vector $(\alpha, \beta) \in R_+^I \times R^J$ such that (16) holds. It is obvious that under this assumption, $Dh_1(\bar{x}), \dots, Dh_J(\bar{x})$ are linearly independent. We first prove that for any given $i \in I(\bar{x})$, there exists $v \in X$ such that

$$(18) \quad g_i^\diamond(\bar{x}; v) < 0,$$

$$(19) \quad \langle Dh_j(\bar{x}), v \rangle = 0 \quad \forall j = 1, 2, \dots, J.$$

If, on the contrary, the above system has no solution, then $v = 0$ is a solution to the following optimization problem:

$$\begin{aligned} \min \quad & g_i^\diamond(\bar{x}; v) \\ \text{s.t.} \quad & \langle Dh_j(\bar{x}), v \rangle = 0 \quad \forall j = 1, \dots, J. \end{aligned}$$

Since the objective function is convex and the constraints are linear, by the Lagrange multiplier rule and the fact that $\partial\phi(0) = \partial^\diamond g_i(\bar{x})$ for $\phi(\cdot) := g_i^\diamond(\bar{x}; \cdot)$ there must exist $\beta \in R^J$ such that

$$0 \in \partial^\diamond g_i(\bar{x}) + \sum_{j=1}^J \beta_j Dh_j(\bar{x}),$$

which is a contradiction. Now we can show that for any two given $i, i' \in I(\bar{x})$, there exists $v \in X$ such that

$$\begin{aligned} g_{i'}^\diamond(\bar{x}; v) &< 0, \\ g_i^\diamond(\bar{x}; v) &< 0, \\ \langle Dh_j(\bar{x}), v \rangle &= 0 \quad \forall j = 1, 2, \dots, J. \end{aligned}$$

On the contrary, suppose that the above system does not have a solution. Then $g_{i'}^\diamond(\bar{x}; v) \geq 0$ for all v satisfying the system (18)–(19), which implies that $v = 0$ is a solution to the following optimization problem with convex constraints:

$$\begin{aligned} \min \quad & g_{i'}^\diamond(\bar{x}; v) \\ \text{s.t.} \quad & g_i^\diamond(\bar{x}; v) \leq 0, \\ & \langle Dh_j(\bar{x}), v \rangle = 0 \quad \forall j = 1, \dots, J. \end{aligned}$$

Indeed, let v be any feasible solution of the above problem and let u be a solution of (18)–(19); then for any $t > 0$, $u + tv$ is a solution of (18)–(19), and hence $g_{i'}^\diamond(\bar{x}; v + tu) \geq 0$ by the assumption, which implies that $g_{i'}^\diamond(\bar{x}; v) \geq 0$ after taking limits

as $t \rightarrow 0$. By the Lagrange multiplier rule, since the Slater condition holds for the above optimization problem, there must exist $(\alpha_i, \beta) \in R_+ \times R^J$ such that

$$0 \in \partial^\diamond g_{i'}(\bar{x}) + \alpha_i \partial^\diamond g_i(\bar{x}) + \sum_{j=1}^J \beta_j Dh_j(\bar{x}),$$

which is a contradiction. The rest of the proof follows from the mathematical induction. $\square$

DEFINITION 3.8 (nondifferentiable linear independence CQ). *Let $\bar{x}$ be a feasible solution of (P). We say that the nondifferentiable linear independence CQ is satisfied if g_i ($i \in I(\bar{x})$) are either Hadamard differentiable at $\bar{x}$ or Lipschitz near $\bar{x}$, g_i ($i \notin I(\bar{x})$) are continuous at $\bar{x}$, h_j ($j = 1, 2, \dots, J$) are Fréchet differentiable at $\bar{x}$ and continuous in a neighborhood of $\bar{x}$, and for any $\xi_i^* \in \partial^\diamond g_i(\bar{x})$ $\{\xi_i^* (i \in I(\bar{x})), Dh_1(\bar{x}), \dots, Dh_J(\bar{x})\}$ are linearly independent.*

The following is a straightforward consequence of Lemma 3.2.

PROPOSITION 3.2 (LICQ implies MFCQ). *The nondifferentiable linear independence CQ implies the nondifferentiable Mangasarian–Fromovitz CQ.*

DEFINITION 3.9 (nondifferentiable Slater CQ). *Let $\bar{x}$ be a feasible solution of (P). We say that the nondifferentiable Slater CQ is satisfied at $\bar{x}$ if g_i ($i \in I(\bar{x})$) are M-P pseudoconvex at $\bar{x}$ and either Hadamard differentiable at $\bar{x}$ or Lipschitz near $\bar{x}$; g_i ($i \notin I(\bar{x})$) are continuous at $\bar{x}$; h_j ($j = 1, 2, \dots, J$) are Fréchet differentiable at $\bar{x}$, continuous in a neighborhood of $\bar{x}$, and quasilinear at $\bar{x}$; $\{Dh_1(\bar{x}), \dots, Dh_J(\bar{x})\}$ are linearly independent; and there exists $\hat{x} \in X$ such that*

$$\begin{aligned} g_i(\hat{x}) &< 0 \quad \forall i \in I(\bar{x}), \\ h_j(\hat{x}) &= 0 \quad \forall j = 1, 2, \dots, J. \end{aligned}$$

PROPOSITION 3.3 (Slater CQ implies MFCQ). *The nondifferentiable Slater CQ implies the nondifferentiable Mangasarian–Fromovitz CQ.*

Proof. Since $g_i(\hat{x}) < g_i(\bar{x})$ for all $i \in I(\bar{x})$, by the M-P pseudoconvexity of g_i ($i \in I(\bar{x})$) we have

$$g_i^\diamond(\bar{x}; \hat{x} - \bar{x}) < 0, \quad i \in I(\bar{x}).$$

Also since $h_j(\hat{x}) = h_j(\bar{x})$ ($j = 1, \dots, J$), quasiconvexity and quasiconcavity of h_j at $\bar{x}$ implies that

$$\langle Dh_j(\bar{x}), \hat{x} - \bar{x} \rangle = 0, \quad j = 1, \dots, J.$$

Thus the system (14)–(15) has a solution $v = \hat{x} - \bar{x}$ and the nondifferentiable Mangasarian–Fromovitz CQ is satisfied. $\square$

PROPOSITION 3.4 (MFCQ implies Kuhn–Tucker CQ). *The nondifferentiable Mangasarian–Fromovitz CQ implies the nondifferentiable Kuhn–Tucker CQ.*

Proof. We first show that the convex cone generated by the set

$$A = \bigcup_{i \in I(\bar{x})} \partial^\diamond g_i(\bar{x}) \cup \bigcup_{j=1}^J \{\pm Dh_j(\bar{x})\}$$

is closed. It is easy to see that

$$\text{cone} A = \text{cone} \bigcup_{i \in I(\bar{x})} \partial^\diamond g_i(\bar{x}) + \left\{ \sum_{j=1}^J \beta_j Dh_j(\bar{x}) : \beta_j \in R \right\},$$

where $\text{cone} A$ denotes the convex cone generated by set A . Since $\text{co} \bigcup_{i \in I(\bar{x})} \partial^\diamond g_i(\bar{x})$ is a nonempty convex weak*-compact subset of X^* not containing zero, as it is shown in [22, Corollary 9.6.1], $\text{cone} \text{co} \bigcup_{i \in I(\bar{x})} \partial^\diamond g_i(\bar{x})$ is closed. But

$$\text{cone} \bigcup_{i \in I(\bar{x})} \partial^\diamond g_i(\bar{x}) = \text{cone} \text{co} \bigcup_{i \in I(\bar{x})} \partial^\diamond g_i(\bar{x});$$

hence $\text{cone} \bigcup_{i \in I(\bar{x})} \partial^\diamond g_i(\bar{x})$ is a closed convex cone. By Lemma 3.2, for any nonzero $\beta \in R^J$,

$$\sum_{j=1}^J \beta_j Dh_j(\bar{x}) \notin \text{cone} \bigcup_{i \in I(\bar{x})} \partial^\diamond g_i(\bar{x}).$$

By the linear independence of $\{Dh_1(\bar{x}), \dots, Dh_J(\bar{x})\}$, $0 \neq \sum_{j=1}^J \beta_j Dh_j(\bar{x})$ for any nonzero $\beta \in R^J$. Therefore any nonzero element in the cone $\{\sum_{j=1}^J \beta_j Dh_j(\bar{x}) : \beta_j \in R\}$ is not in the closed convex cone $\text{cone} \bigcup_{i \in I(\bar{x})} \partial^\diamond g_i(\bar{x})$, which implies that $\text{cone} A$ is closed by [22, Corollary 9.1.2] (which obviously holds in any general Banach space as well).

Let v be a solution of (14) and (15). Since $\{Dh_1(\bar{x}), \dots, Dh_J(\bar{x})\}$ are linearly independent, by the correction theorem of Halkin [9, Theorem F], there exist a neighborhood U of $\bar{x}$ and a continuous mapping ζ from U into X such that $\zeta(\bar{x}) = 0$, ζ is Fréchet differentiable at $\bar{x}$ with $D\zeta(\bar{x}) = 0$, and

$$(20) \quad h_j(x + \zeta(x)) = \langle Dh_j(\bar{x}), x - \bar{x} \rangle \quad \forall x \in U, j = 1, 2, \dots, J.$$

For all $t \in R$ such that $\bar{x} + tv \in U$, denote

$$\alpha(t) := \bar{x} + tv + \zeta(\bar{x} + tv).$$

Then $h_j(\alpha(t)) = 0$, $j = 1, 2, \dots, J$, for all $t > 0$ small enough. Let $i \in I(\bar{x})$. If g_i is Hadamard differentiable at $\bar{x}$, then since $\lim_{t \downarrow 0} \zeta(\bar{x} + tv)/t = D\zeta(\bar{x}) = 0$ by (14)

$$g_i(\alpha(t)) < 0 \quad \forall t > 0 \text{ small enough.}$$

Now suppose that g_i is Lipschitz near $\bar{x}$. By (14) one has that there exists $r > 0$ such that for $t > 0$ small enough

$$g_i(\bar{x} + tv) - g_i(\bar{x}) < -rt.$$

Since g_i is Lipschitz near $\bar{x}$, for $t > 0$ small enough and any $v' \in X$

$$g_i(\bar{x} + tv') - g_i(\bar{x} + tv) \leq L_{g_i} t \|v' - v\|.$$

Since $D\zeta(\bar{x}) = 0$, $\zeta(\bar{x}) = 0$, one has

$$\|\zeta(\bar{x} + tv)\| < \frac{r}{2L_{g_i}} t \quad \forall t > 0 \text{ small enough,}$$

and hence

$$\begin{aligned} g_i(\alpha(t)) &= g_i(\bar{x} + tv + \zeta(\bar{x} + tv)) \\ &= g_i\left(\bar{x} + t\left[v + \frac{\zeta(\bar{x} + tv)}{t}\right]\right) - g_i(\bar{x} + tv) + g_i(\bar{x} + tv) - g_i(\bar{x}) \\ &\leq L_{g_i} \|\zeta(\bar{x} + tv)\| - rt \\ &< -\frac{r}{2} t < 0 \quad \text{for } t > 0 \text{ small enough.} \end{aligned}$$

By the continuity assumptions on $g_i (i \notin I(\bar{x}))$, one also has

$$g_i(\alpha(t)) < 0 \quad \forall t > 0 \text{ small enough, } i \notin I(\bar{x}).$$

Hence $v \in A_\Omega(\bar{x})$.

Now let v satisfy

$$\begin{aligned} g_i^\diamond(\bar{x}; v) &\leq 0 \quad \forall i \in I(\bar{x}), \\ \langle Dh_j(\bar{x}), v \rangle &= 0 \quad \forall j = 1, 2, \dots, J. \end{aligned}$$

By the assumption of the Mangasarian–Fromovitz CQ, there exists a sequence $\{v_k\}$ such that

$$\begin{aligned} g_i^\diamond(\bar{x}; v_k) &< 0 \quad \forall i \in I(\bar{x}), \\ \langle Dh_j(\bar{x}), v_k \rangle &= 0 \quad \forall j = 1, 2, \dots, J, \end{aligned}$$

and $v = \lim_{k \rightarrow \infty} v_k$. By the proof above, $v_k \in A_\Omega(\bar{x})$, and so $v = \lim_{k \rightarrow \infty} v_k \in A_\Omega(\bar{x})$. $\square$

We now extend the Arrow–Hurwicz–Uzawa CQ introduced in [3] to the nondifferentiable setting.

DEFINITION 3.10 (nondifferentiable Arrow–Hurwicz–Uzawa CQ). *Let $\bar{x}$ be a feasible solution of (P). We say that the nondifferentiable Arrow–Hurwicz–Uzawa CQ is satisfied at $\bar{x}$ if g_i ($i \in I(\bar{x})$), h_j ($j = 1, 2, \dots, J$) are either Gâteaux differentiable at $\bar{x}$ or Lipschitz near $\bar{x}$, g_i ($i \notin I(\bar{x})$) are continuous at $\bar{x}$, h_j ($j = 1, 2, \dots, J$) are M-P pseudoaffine at $\bar{x}$, the convex cone generated by the set (2) is closed, and there exists $v \in X$ such that*

$$\begin{aligned} (21) \quad & g_i^\diamond(\bar{x}; v) < 0 \quad \forall i \in W, \\ (22) \quad & g_i^\diamond(\bar{x}; v) \leq 0 \quad \forall i \in V, \\ (23) \quad & h_j^\diamond(\bar{x}; v) = 0 \quad \forall j = 1, 2, \dots, J, \end{aligned}$$

where

$$\begin{aligned} V &:= \{i \in I(\bar{x}) : g_i \text{ is M-P pseudoconcave at } \bar{x}\}, \\ W &:= I(\bar{x}) \setminus V. \end{aligned}$$

PROPOSITION 3.5 (AHUCQ implies Zangwill CQ). *The nondifferentiable Arrow–Hurwicz–Uzawa CQ implies the generalized Zangwill CQ.*

Proof. Suppose that v satisfies (21)–(23). For any $i \in W$ by virtue of (21), for all $\tau \in (0, 1]$ small enough,

$$g_i(\bar{x} + \tau v) < g_i(\bar{x}) = 0.$$

For $i \in V$ by virtue of (22), $g_i^\diamond(\bar{x}; v) \leq 0$, which implies by the definition of the M-P pseudoconcavity that $g_i(\bar{x} + \tau v) \leq g_i(\bar{x})$ for all $\tau \geq 0$ small enough. By the continuity assumptions at $\bar{x}$ for g_i ($i \notin I(\bar{x})$), for all $\tau \in (0, 1]$ small enough,

$$g_i(\bar{x} + \tau v) < 0 \quad \forall i \notin I(\bar{x}).$$

Hence for all $\tau > 0$ small enough,

$$\begin{aligned} g_i(\bar{x} + \tau v) &\leq 0, \quad i = 1, 2, \dots, I, \\ h_j(\bar{x} + \tau v) &= 0 \quad \forall j = 1, 2, \dots, J, \end{aligned}$$

which implies that $v \in D_\Omega(\bar{x})$ and the proof of the proposition is complete due to the continuity of $g_i^\diamond(\bar{x}, \cdot)$ ($i \in I(\bar{x})$) and $h_j^\diamond(\bar{x}, \cdot)$ ($j = 1, 2, \dots, J$). $\square$

DEFINITION 3.11 (nondifferentiable weak Slater CQ). *Let $\bar{x}$ be a feasible solution of (P). We say the nondifferentiable weak Slater CQ holds at $\bar{x}$ if g_i ($i \in I(\bar{x})$) are M-P pseudoconvex at $\bar{x}$ and either Gâteaux differentiable at $\bar{x}$ or Lipschitz near $\bar{x}$; g_i ($i \notin I(\bar{x})$) are continuous at $\bar{x}$; h_j ($j = 1, 2, \dots, J$) are Gâteaux differentiable, continuous, and M-P pseudoaffine; the convex cone generated by the set (2) is closed; and there exists $\hat{x} \in X$ such that*

$$\begin{aligned} g_i(\hat{x}) &< 0, \quad i \in I(\bar{x}), \\ h_j(\hat{x}) &= 0, \quad j = 1, 2, \dots, J. \end{aligned}$$

PROPOSITION 3.6 (weak Slater CQ implies AHUCQ). *The nondifferentiable weak Slater CQ implies the nonsmooth Arrow–Hurwicz–Uzawa CQ.*

Proof. Take $V = \emptyset$ and $W = I(\bar{x})$. Since $g_i(\hat{x}) < g_i(\bar{x})$ for all $i \in I(\bar{x})$, by the M-P pseudoconvexity of g_i ($i \in I(\bar{x})$) we have

$$(24) \quad g_i^\diamond(\bar{x}; \hat{x} - \bar{x}) < 0, \quad i \in I(\bar{x}).$$

By Proposition 2.3, h_j ($j = 1, 2, \dots, J$) are quasilinear. Since $h_j(\hat{x}) = h_j(\bar{x})$ ($j = 1, \dots, J$), quasiconvexity and quasiconcavity of h_j at $\bar{x}$ implies that

$$h_j'(\bar{x}; \hat{x} - \bar{x}) = 0.$$

Hence the system (21)–(23) has a solution $v = \hat{x} - \bar{x}$. The proof of the proposition is complete. $\square$

DEFINITION 3.12 (nondifferentiable weak reverse convex CQ). *We say that the nondifferentiable weak reverse convex CQ holds at $\bar{x}$ if g_i ($i \in I(\bar{x})$), h_j ($j = 1, 2, \dots, J$) are either Gâteaux differentiable at $\bar{x}$ or Lipschitz near $\bar{x}$, g_i ($i \in I(\bar{x})$) are M-P pseudoconcave at $\bar{x}$, g_i ($i \notin I(\bar{x})$) are continuous at $\bar{x}$, h_j ($j = 1, 2, \dots, J$) are M-P pseudoaffine, and the convex cone generated by the set (2) is closed.*

Since (22)–(23) always has a solution $v = 0$, the following relationship between the nondifferentiable weak reverse convex constraint CQ and the nondifferentiable Arrow–Hurwicz–Uzawa CQ is immediate.

PROPOSITION 3.7 (weak reverse convex CQ implies AHUCQ). *The nondifferentiable weak reverse convex constraint CQ implies the nondifferentiable Arrow–Hurwicz–Uzawa CQ.*

Finally we end this section with an equivalent condition to the nondifferentiable Arrow–Hurwicz–Uzawa CQ. We omit the proof since the proof is similar to that of Lemma 3.2.

PROPOSITION 3.8. *Suppose that g_i ($i \in I(\bar{x})$), h_j ($j = 1, 2, \dots, J$) are either Gâteaux differentiable at $\bar{x}$ or Lipschitz near $\bar{x}$, g_i ($i \notin I(\bar{x})$) are continuous at $\bar{x}$, h_j ($j = 1, 2, \dots, J$) are M-P pseudoaffine at $\bar{x}$, and the convex cone generated by the set (2) is closed. The nondifferentiable Arrow–Hurwicz–Uzawa CQ is equivalent to the nonexistence of $(\alpha, \beta) \in R_+^I \times R^J$ such that $\alpha_W \neq 0$ and*

$$(25) \quad 0 \in \sum_{i \in W} \alpha_i \partial^\diamond g_i(\bar{x}) + \sum_{i \in V} \alpha_i \partial^\diamond g_i(\bar{x}) + \sum_{j=1}^J \beta_j \partial^\diamond h_j(\bar{x}),$$

where

$$\begin{aligned} V &:= \{i \in I(\bar{x}) : g_i \text{ is M-P pseudoconcave at } \bar{x}\}, \\ W &:= I(\bar{x}) \setminus V. \end{aligned}$$

4. Bilevel optimization. In this section we apply the results obtained in the previous section to the bilevel optimization problem,

$$\begin{aligned}
 \text{(BP)} \quad & \min \quad F(x, y) \\
 & \text{s.t.} \quad y \in S(x), \\
 & \quad G_i(x, y) \leq 0, \quad i = 1, 2, \dots, I, \\
 & \quad H_j(x, y) = 0, \quad j = 1, 2, \dots, J,
 \end{aligned}$$

where $S(x)$ denotes the set of solutions of the lower level problem:

$$\begin{aligned}
 \text{(P}_x\text{)} \quad & \min_y \quad f(x, y) \\
 & \text{s.t.} \quad g_i(x, y) \leq 0, \quad i = 1, 2, \dots, m, \\
 & \quad h_j(x, y) = 0, \quad j = 1, 2, \dots, n,
 \end{aligned}$$

and F, G_i, H_j, f, g_i, h_j are functions on the Banach space $X \times Y$. For simplicity we assume that $S(x)$ is nonempty for all $x \in X$.

Define the value function of the lower level problem by

$$V(x) := \min_y \{f(x, y) : g_i(x, y) \leq 0, h_j(x, y) = 0, i = 1, 2, \dots, m, j = 1, 2, \dots, n\}.$$

Then (BP) can be reformulated as the following single level optimization problem:

$$\begin{aligned}
 \text{(SP)} \quad & \min \quad F(x, y) \\
 & \text{s.t.} \quad f(x, y) - V(x) \leq 0, \\
 & \quad g_i(x, y) \leq 0, \quad i = 1, 2, \dots, m, \\
 & \quad h_j(x, y) = 0, \quad j = 1, 2, \dots, n, \\
 & \quad G_i(x, y) \leq 0, \quad i = 1, 2, \dots, I, \\
 & \quad H_j(x, y) = 0, \quad j = 1, 2, \dots, J.
 \end{aligned}$$

It is known that $V(x)$ may not be differentiable in general, even in the case where all problem data f, g_i, h_j are continuously differentiable, and hence a nonsmooth multiplier rule must be used as in [28, 29]. Moreover it was shown in [28, Proposition 3.2] that the CQs such as the linear independence CQ, the Slater CQ, and the Mangasarian–Fromovitz CQ do not hold for (SP). It is obvious that the nondifferentiable weak Slater CQ will never be satisfied since the inequality constraint $f(x, y) - V(x) \leq 0$ is actually an equality constraint. In this section, we show that it is possible for the nondifferentiable weak reverse convex CQ to hold; hence the nondifferentiable Arrow–Hurwicz–Uzawa CQ, the generalized Zangwill CQ, the nondifferentiable Kuhn–Tucker CQ, and the nondifferentiable Abadie CQ are also applicable CQs for (SP).

Excluding the CQs that will never hold for (SP) such as (2), (5)–(7) in Theorem 1.1, we derive the KKT condition for (SP) by using the calculus rules for the M-P subdifferential in (ii)–(iii) of Proposition 2.1 as follows.

THEOREM 4.1. *Let $(\bar{x}, \bar{y})$ be a local optimal solution of (SP). Suppose that the objective function $F(x, y)$ is either Gâteaux differentiable at $(\bar{x}, \bar{y})$ or Lipschitz near $(\bar{x}, \bar{y})$ and the value function $V(x)$ is Lipschitz near $\bar{x}$. If one of the CQs such as the nondifferentiable weak reverse convex CQ, the nondifferentiable Arrow–Hurwicz–Uzawa CQ, the generalized Zangwill CQ, the nondifferentiable Kuhn–Tucker CQ, and*

the nondifferentiable Abadie CQ holds at $(\bar{x}, \bar{y})$, then the KKT condition holds; i.e., there exist scalars $\lambda \geq 0$, $\alpha_i \geq 0$ ($i = 1, 2, \dots, m$), β_j ($j = 1, 2, \dots, p$), γ_i ($i = 1, 2, \dots, I$), η_j ($j = 1, 2, \dots, J$) such that

$$\begin{aligned} 0 \in & \partial^\diamond F(\bar{x}, \bar{y}) + \lambda(\partial^\diamond f(\bar{x}, \bar{y}) - \partial^\diamond V(\bar{x}) \times \{0\}) + \sum_{i=1}^m \alpha_i \partial^\diamond g_i(\bar{x}, \bar{y}) \\ & + \sum_{j=1}^n \beta_j \partial^\diamond h_j(\bar{x}, \bar{y}) + \sum_{i=1}^I \gamma_i \partial^\diamond G_i(\bar{x}, \bar{y}) + \sum_{j=1}^J \eta_j \partial^\diamond H_j(\bar{x}, \bar{y}), \\ & \alpha_i g_i(\bar{x}, \bar{y}) = 0, \quad i = 1, 2, \dots, m, \\ & \gamma_i G_i(\bar{x}, \bar{y}) = 0, \quad i = 1, 2, \dots, I. \end{aligned}$$

In the above KKT condition, we need to give an upper estimate for the term $\partial^\diamond V(\bar{x})$. Such an estimate usually involves a convex combination of solutions and multipliers for the lower level problem as in [28, Proposition 2.1], and a growth hypothesis assumption is usually needed [7, Theorem 6.2]. However, in the case when the value function is convex, no such convex combination and growth hypothesis are needed, as the following result indicates.

PROPOSITION 4.1. *Let $x \in X$ and $y \in S(x)$. Suppose that f, g_i ($i = 1, 2, \dots, I$), h_j ($j = 1, 2, \dots, J$) are Gâteaux differentiable at (x, y) and the KKT condition holds for problem (26). If the value function $V(x)$ is convex, then for any $y \in S(x)$,*

$$\partial V(x) \subseteq \left\{ D_x f(x, y) + \sum_{i=1}^m \nu_i D_x g_i(x, y) + \sum_{j=1}^n \pi_j D_x h_j(x, y) : (\nu, \pi) \in M^1(y) \right\},$$

where $\partial V(x)$ denotes the subdifferential in the sense of convex analysis and $M^1(y)$ is the set of multipliers for (P_x) :

$$M^1(y) := \left\{ (\nu, \pi) \in R_+^m \times R^n : \begin{array}{l} D_y f(x, y) + \sum_{i=1}^m \nu_i D_y g_i(x, y) + \sum_{j=1}^n \pi_j D_y h_j(x, y) = 0 \\ \nu_i g_i(x, y) = 0, \quad i = 1, 2, \dots, m \end{array} \right\}.$$

Proof. Now let $\xi \in \partial V(x)$. Then by definition of the subdifferential in the sense of convex analysis,

$$V(x') - V(x) \geq \langle \xi, x' - x \rangle \quad \forall x' \in X,$$

which implies by definition of the value function that for all (x', y') satisfying the constraints

$$g_i(x', y') \leq 0, \quad i = 1, 2, \dots, m, \quad h_j(x', y') = 0, \quad j = 1, 2, \dots, n,$$

one has

$$f(x', y') - f(x, y) \geq \langle \xi, x' - x \rangle.$$

That is, $(x', y') = (x, y)$ is a solution to the following optimization problem:

$$(26) \quad \begin{cases} \min_{x', y'} & f(x', y') - \langle \xi, x' \rangle, \\ & g_i(x', y') \leq 0, \quad i = 1, 2, \dots, m, \\ & h_j(x', y') = 0, \quad i = 1, 2, \dots, n. \end{cases}$$

By the KKT condition there exists $(\nu, \pi) \in R_+^m \times R^n$ such that

$$\begin{aligned} 0 &= D_x f(x, y) - \xi + \sum_{i=1}^m \nu_i D_x g_i(x, y) + \sum_{j=1}^n \pi_j D_x h_j(x, y), \\ 0 &= D_y f(x, y) + \sum_{i=1}^m \nu_i D_y g_i(x, y) + \sum_{j=1}^n \pi_j D_y h_j(x, y), \\ \nu_i g_i(x, y) &= 0, \quad i = 1, 2, \dots, m, \end{aligned}$$

and hence the proof is complete. $\square$

Combining Theorem 4.1 and Proposition 4.1, we now have a KKT condition for (SP) which does not involve the value function.

THEOREM 4.2. *Let $(\bar{x}, \bar{y})$ be a local optimal solution of (SP). Suppose that $F(x, y)$ is either Gâteaux differentiable at $(\bar{x}, \bar{y})$ or Lipschitz continuous near $(\bar{x}, \bar{y})$, f, g_i ($i = 1, 2, \dots, I$), h_j ($j = 1, 2, \dots, J$) are Gâteaux differentiable; the KKT condition for problem (26) holds; and the value function $V(x)$ is convex. If one of the CQs, such as the nondifferentiable weak reverse convex CQ, the nondifferentiable Arrow–Hurwicz–Uzawa CQ, the generalized Zangwill CQ, the nondifferentiable Kuhn–Tucker CQ, or the nondifferentiable Abadie CQ, holds at $(\bar{x}, \bar{y})$, then there exist scalars $\lambda \geq 0, \alpha_i \geq 0, \nu_i \geq 0$ ($i = 1, 2, \dots, m$), β_j, π_j ($j = 1, 2, \dots, n$), $\gamma_i \geq 0$ ($i = 1, 2, \dots, I$), η_j ($j = 1, 2, \dots, J$) such that*

$$\begin{aligned} 0 &\in \partial^\diamond F(\bar{x}, \bar{y}) + \sum_{i=1}^m (\alpha_i - \lambda \nu_i) Dg_i(\bar{x}, \bar{y}) + \sum_{j=1}^n \beta_j Dh_j(\bar{x}, \bar{y}) \\ &\quad + \sum_{i=1}^I \gamma_i \partial^\diamond G_i(\bar{x}, \bar{y}) + \sum_{j=1}^J \eta_j \partial^\diamond H_j(\bar{x}, \bar{y}), \\ 0 &= D_y f(\bar{x}, \bar{y}) + \sum_{i=1}^m \nu_i D_y g_i(\bar{x}, \bar{y}) + \sum_{j=1}^n \pi_j D_y h_j(\bar{x}, \bar{y}), \\ \nu_i g_i(\bar{x}, \bar{y}) &= 0, \quad i = 1, 2, \dots, m, \\ \alpha_i g_i(\bar{x}, \bar{y}) &= 0, \quad i = 1, 2, \dots, m, \\ \gamma_i G_i(\bar{x}, \bar{y}) &= 0, \quad i = 1, 2, \dots, I. \end{aligned}$$

Proof. Applying Theorem 4.1 and Proposition 4.1, we find scalars $\lambda \geq 0, \alpha_i \geq 0, \nu_i \geq 0$ ($i = 1, 2, \dots, m$), ζ_j, π_j ($j = 1, 2, \dots, n$), $\gamma_i \geq 0$ ($i = 1, 2, \dots, I$), η_j ($j = 1, 2, \dots, J$) such that

$$\begin{aligned} 0 &\in \partial^\diamond F(\bar{x}, \bar{y}) - \lambda \left[\sum_{i=1}^m \nu_i D_x g_i(\bar{x}, \bar{y}) + \sum_{j=1}^n \pi_j D_x h_j(\bar{x}, \bar{y}) \right] \times \{-D_y f(\bar{x}, \bar{y})\} \\ (27) \quad &+ \sum_{i=1}^m \alpha_i Dg_i(\bar{x}, \bar{y}) + \sum_{j=1}^n \zeta_j Dh_j(\bar{x}, \bar{y}) + \sum_{i=1}^I \gamma_i \partial^\diamond G_i(\bar{x}, \bar{y}) + \sum_{j=1}^J \eta_j \partial^\diamond H_j(\bar{x}, \bar{y}), \end{aligned}$$

$$\begin{aligned} (28) \quad 0 &= D_y f(\bar{x}, \bar{y}) + \sum_{i=1}^m \nu_i D_y g_i(\bar{x}, \bar{y}) + \sum_{j=1}^n \pi_j D_y h_j(\bar{x}, \bar{y}), \\ \nu_i g_i(\bar{x}, \bar{y}) &= 0, \quad i = 1, 2, \dots, m, \\ \alpha_i g_i(\bar{x}, \bar{y}) &= 0, \quad i = 1, 2, \dots, m, \\ \gamma_i G_i(\bar{x}, \bar{y}) &= 0, \quad i = 1, 2, \dots, I. \end{aligned}$$

From (28),

$$-D_y f(\bar{x}, \bar{y}) = \sum_{i=1}^m \nu_i D_y g_i(\bar{x}, \bar{y}) + \sum_{j=1}^n \pi_j D_y h_j(\bar{x}, \bar{y}).$$

Substituting the above into (27) and denoting $\beta_j = \zeta_j - \lambda \pi_j$ completes the proof. $\square$

We now consider the special case when the lower level problem is linear; i.e., the functions $f(x, y)$, $g_i(x, y)$, $h_j(x, y)$ are all jointly linear. It is known that for linear bilevel programming problems, which are the bilevel optimization problems where the lower level problem is jointly linear and there is no upper level constraints, no CQs are needed. By Theorem 4.2 and the weak reverse convex CQ, we have the following KKT condition for the following “generalized linear” bilevel optimization problem where no constraint qualification is needed.

COROLLARY 4.1. *Let $(\bar{x}, \bar{y})$ be a local optimal solution of (SP). Suppose that $F(x, y)$ is either Gâteaux differentiable at $(\bar{x}, \bar{y})$ or Lipschitz continuous near $(\bar{x}, \bar{y})$, $f(x, y)$, $g_i(x, y)$ ($i = 1, 2, \dots, m$), $h_j(x, y)$ ($j = 1, 2, \dots, n$) are jointly linear; $G_i(x, y)$ ($i = 1, 2, \dots, I$) are Gâteaux differentiable and M-P pseudoconcave at $(\bar{x}, \bar{y})$; and $H_j(x, y)$ ($j = 1, 2, \dots, J$) are Gâteaux differentiable and M-P pseudoaffine at $(\bar{x}, \bar{y})$. Then there exist scalars $\lambda \geq 0$, $\alpha_i \geq 0$, $\nu_i \geq 0$ ($i = 1, 2, \dots, m$), β_j, π_j ($j = 1, 2, \dots, n$), $\gamma_i \geq 0$ ($i = 1, 2, \dots, I$), η_j ($j = 1, 2, \dots, J$) such that*

$$\begin{aligned} 0 &\in \partial^\diamond F(\bar{x}, \bar{y}) + \sum_{i=1}^m (\alpha_i - \lambda \nu_i) Dg_i(\bar{x}, \bar{y}) + \sum_{j=1}^n \beta_j Dh_j(\bar{x}, \bar{y}) \\ &\quad + \sum_{i=1}^I \gamma_i DG_i(\bar{x}, \bar{y}) + \sum_{j=1}^J \eta_j DH_j(\bar{x}, \bar{y}), \\ 0 &= D_y f(\bar{x}, \bar{y}) + \sum_{i=1}^m \nu_i D_y g_i(\bar{x}, \bar{y}) + \sum_{j=1}^n \pi_j D_y h_j(\bar{x}, \bar{y}), \\ \nu_i g_i(\bar{x}, \bar{y}) &= 0, \quad i = 1, 2, \dots, m, \\ \alpha_i g_i(\bar{x}, \bar{y}) &= 0, \quad i = 1, 2, \dots, m, \\ \gamma_i G_i(\bar{x}, \bar{y}) &= 0, \quad i = 1, 2, \dots, I. \end{aligned}$$

Proof. By Theorem 4.2 and the weak reverse convex CQ, under the assumptions of the corollary it suffices to prove that the convex cone generated by the set

$$A := [Df(\bar{x}, \bar{y}) - \partial V(\bar{x}) \times \{0\}] \cup \bigcup_{i=1}^m \{Dg_i\} \cup \bigcup_{j=1}^n \{\pm Dh_j\} \cup \bigcup_{i=1}^I \{DG_i\} \cup \bigcup_{j=1}^J \{\pm DH_j\}$$

is closed, where the dependence of the derivatives on $(\bar{x}, \bar{y})$ is omitted whenever there is no confusion.

Since the lower level problem is linear, by [8, Proposition 2.13] (which obviously holds in any general Banach space as well), the value function $V(x)$ is a polyhedral convex function, which implies by [22, Theorem 23.10] (which obviously holds in any general Banach space as well) that $\partial V(\bar{x})$ is a polyhedral convex set. Since by the assumptions for the problem (SP) $V(x)$ is finite and convex on X , $\partial V(\bar{x})$ is bounded and hence $Df(\bar{x}, \bar{y}) - \partial V(\bar{x}) \times \{0\}$ is a bounded polyhedral convex set. Therefore, by definition, $Df(\bar{x}, \bar{y}) - \partial V(\bar{x}) \times \{0\}$ is a convex hull of a finite set of points. Consequently, the convex hull of the set $[A \cup \{0\}]$ is a polyhedral convex

set containing the origin. By [22, Corollary 19.7.1] (which also holds in any general Banach space) the convex cone generated by $\text{co}[A \cup \{0\}]$ is polyhedral. But the convex cone generated by A is the same as the convex cone generated by $\text{co}[A \cup \{0\}]$, so it is also polyhedral and hence closed. The proof of the corollary is therefore complete. $\square$

It is interesting to compare the value function approach with the classical approach, in which the lower level problem is replaced by the KKT condition of the lower level problem. Suppose the KKT condition is necessary and sufficient for optimality of the lower level problem; then BP is equivalent to the following single level optimization problem:

$$\begin{aligned}
 & \min_{x, y, \nu, \pi} F(x, y) \\
 & \text{s.t.} \quad 0 = D_y f(x, y) + \sum_{i=1}^m \nu_i D_y g_i(x, y) + \sum_{j=1}^n \pi_j D_y h_j(x, y), \\
 & \quad \sum_{i=1}^m \nu_i g_i(x, y) \geq 0 \\
 & \quad \nu_i \geq 0, \quad g_i(x, y) \leq 0, \quad i = 1, 2, \dots, m, \\
 & \quad h_j(x, y) = 0, \quad j = 1, 2, \dots, n, \\
 & \quad G_i(x, y) \leq 0, \quad i = 1, 2, \dots, I, \\
 & \quad H_j(x, y) = 0, \quad j = 1, 2, \dots, J.
 \end{aligned}
 \tag{KP}$$

It is known [30, Proposition 1.1] that usual CQs such as the Mangasarian–Fromovitz CQ do not hold for problem (KP). However, if a suitable CQ is satisfied and the functions f, g_i ($i = 1, \dots, I$), h_j ($j = 1, \dots, J$) are second order continuously differentiable, then the KKT condition for problem (KP) is the existence of scalars $\lambda \geq 0, \mu, \alpha_i \geq 0, \nu_i \geq 0$ ($i = 1, 2, \dots, m$), β_j, π_j ($j = 1, 2, \dots, n$), $\gamma_i \geq 0$ ($i = 1, 2, \dots, I$), η_j ($j = 1, 2, \dots, J$) such that

$$\begin{aligned}
 & 0 \in \partial^\diamond F(\bar{x}, \bar{y}) + \sum_{i=1}^m (\alpha_i - \lambda \nu_i) D g_i(\bar{x}, \bar{y}) + \sum_{j=1}^n \beta_j D h_j(\bar{x}, \bar{y}) \\
 & \quad + \sum_{i=1}^I \gamma_i \partial^\diamond G_i(\bar{x}, \bar{y}) + \sum_{j=1}^J \eta_j \partial^\diamond H_j(\bar{x}, \bar{y}) \\
 & \quad + \mu D \left[D_y f + \sum_{i=1}^m \nu_i D_y g_i + \sum_{j=1}^n \pi_j D_y h_j \right] (\bar{x}, \bar{y}), \\
 & 0 = D_y f(\bar{x}, \bar{y}) + \sum_{i=1}^m \nu_i D_y g_i(\bar{x}, \bar{y}) + \sum_{j=1}^n \pi_j D_y h_j(\bar{x}, \bar{y}), \\
 & \nu_i g_i(\bar{x}, \bar{y}) = 0, \quad i = 1, 2, \dots, m, \\
 & \alpha_i g_i(\bar{x}, \bar{y}) = 0, \quad i = 1, 2, \dots, m, \\
 & \gamma_i G_i(\bar{x}, \bar{y}) = 0, \quad i = 1, 2, \dots, I.
 \end{aligned}$$

Comparing the KKT condition for (KP) with the KKT condition for (SP) in Theorem 4.2, it is easy to see that if the value function is convex, then the fact that the KKT condition holds for problem (SP) implies that the KKT condition for problem

(KP) holds with $\mu = 0$, and in the case when the lower level problem is linear, the KKT condition for (SP) coincides with the KKT condition for problem (KP). This establishes the relationship between the two approaches. Hence the nondifferentiable Arrow–Hurwicz–Uzawa CQ is also an applicable CQ for problem (KP).

Acknowledgments. I would like to thank Jean-Paul Penot for pointing out an error in an earlier version and for many helpful conversations and suggestions about the paper. I would also like to thank the anonymous referees for their extremely valuable comments and suggestions. In particular, I thank one of the referees for suggesting the current version of the statement and the proof of Theorem 3.1.

REFERENCES

- [1] J.-P. AUBIN AND H. FRANKOWSKA, *Set-Valued Analysis*, Birkhäuser Boston, Boston, 1990.
- [2] J. M. ABADIE, *On the Kuhn-Tucker theorem*, in *Nonlinear Programming*, J. Abadie, ed., John Wiley, New York, 1967, pp. 21–36.
- [3] K. J. ARROW, L. HURWICZ, AND H. UZAWA, EDS., *Studies in Linear and Nonlinear Programming*, Stanford University Press, Stanford, CA, 1958.
- [4] M. S. BAZARAA, H. D. SHERALI, AND C. M. SHETTY, *Nonlinear Programming Theory and Algorithms*, 2nd ed., John Wiley, New York, 1993.
- [5] J. R. BIRGE AND L. QI, *Semiregularity and generalized subdifferentials with applications to optimization*, *Math. Oper. Res.*, 18 (1993), pp. 982–1005.
- [6] J. M. BORWEIN, J. S. TREIMAN, AND Q. J. ZHU, *Necessary conditions for constrained optimization problems with semicontinuous and continuous data*, *Trans. Amer. Math. Soc.*, 350 (1998), pp. 2409–2429.
- [7] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.
- [8] A. V. Fiacco AND J. Kyparisis, *Convexity and concavity properties of the optimal value function in parametric nonlinear programming*, *J. Optim. Theory Appl.*, 48 (1986), pp. 95–126.
- [9] H. HALKIN, *Implicit functions and optimization problems without continuous differentiability of the data*, *SIAM J. Control*, 12 (1974), pp. 229–236.
- [10] M. S. GOWDA AND M. TEBoulLE, *A comparison of constraint qualifications in infinite-dimension convex programming*, *SIAM J. Control Optim.*, 28 (1990), pp. 925–935.
- [11] A. D. IOFFE, *Necessary conditions in nonsmooth optimization*, *Math. Oper. Res.*, 9 (1984), pp. 159–189.
- [12] A. D. IOFFE, *A Lagrange multiplier rule with small convex-valued subdifferentials for nonsmooth problems of mathematical programming involving equality and nonfunctional constraints*, *Math. Programming*, 58 (1993), pp. 137–145.
- [13] J. JAHN, *Introduction to the Theory of Nonlinear Optimization*, Springer, Berlin, 1996.
- [14] A. JOURANI, *Constraint qualifications and Lagrange multipliers in nondifferentiable programming problems*, *J. Optim. Theory Appl.*, 81 (1994), pp. 533–548.
- [15] H. W. KUHN AND A. W. TUCKER, *Nonlinear programming*, in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, J. Neyman, ed., University of California Press, Berkeley, CA, 1951, pp. 481–492.
- [16] O. L. MANGASARIAN, *Nonlinear Programming*, McGraw-Hill, New York, 1969; reprinted as *Classics. Appl. Math.* 10, SIAM, Philadelphia, 1994.
- [17] P. MICHEL AND J.-P. PENOT, *Calcul sous-différentiel pour des fonctions lipschitziennes et non lipschitziennes*, *C. R. Acad. Sci. Paris Sér. I Math.*, 12 (1984), pp. 269–272.
- [18] P. MICHEL AND J.-P. PENOT, *A generalized derivative for calm and stable functions*, *Differential Integral Equations*, 5 (1992), pp. 433–454.
- [19] B. S. MORDUKHOVICH, *On necessary conditions for an extremum in nonsmooth optimization*, *Soviet Math. Dokl.*, 283 (1985), pp. 215–220.
- [20] H. V. NGAI AND M. THÉRA, *A fuzzy necessary optimality condition for non-Lipschitz optimization in Asplund spaces*, *SIAM J. Optim.*, 12 (2002), pp. 656–668.
- [21] J.-P. PENOT, *Are generalized derivatives useful for generalized convex functions?*, in *Generalized Convexity, Generalized Monotonicity: Recent Results*, J. P. Crouzeix et al., eds., Kluwer Academic, Dordrecht, The Netherlands, 1998, pp. 3–59.
- [22] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [23] J. S. TREIMAN, *Shrinking generalized gradients*, *Nonlinear Analysis*, 12 (1988), pp. 1429–1450.

- [24] J. S. TREIMAN, *Lagrange multipliers for nonconvex generalized gradients with equality, inequality, and set constraints*, SIAM J. Control Optim., 37 (1999), pp. 1313–1329.
- [25] Z. WU AND J. J. YE, *Some results on integration of subdifferentials*, Nonlinear Anal., 39 (2000), pp. 955–976.
- [26] Z. WU AND J. J. YE, *Equivalence among various derivatives and subdifferentials of the distance function*, J. Math. Anal. Appl., 282 (2003), pp. 629–647.
- [27] J. J. YE, *Multiplier rules under mixed assumptions of differentiability and Lipschitz continuity*, SIAM J. Control Optim., 39 (2001), pp. 1441–1460.
- [28] J. J. YE AND D. L. ZHU, *Optimality conditions for bilevel programming problems*, Optimization, 33 (1995), pp. 9–27.
- [29] J. J. YE AND D. L. ZHU, *A note on optimality conditions for bilevel programming problems*, Optimization, 39 (1997), pp. 361–366.
- [30] J. J. YE, D. L. ZHU, AND Q. J. ZHU, *Exact penalization and necessary optimality conditions for generalized bilevel programming problems*, SIAM J. Optim., 7 (1997), pp. 481–507.
- [31] C. ZĂLINESCU, *A comparison of constraint qualifications in infinite-dimensional convex programming revisited*, J. Austral. Math. Soc. Ser. B, 40 (1999), pp. 353–378.
- [32] W. I. ZANGWILL, *Nonlinear Programming: A Unified Approach*, Prentice-Hall, Englewood Cliffs, NJ, 1969.

GLOBAL MINIMIZATION OF NORMAL QUARTIC POLYNOMIALS BASED ON GLOBAL DESCENT DIRECTIONS*

LIQUN QI[†], ZHONG WAN[‡], AND YU-FEI YANG[§]

Abstract. A normal quartic polynomial is a quartic polynomial whose fourth degree term coefficient tensor is positive definite. Its minimization problem is one of the simplest cases of nonconvex global optimization, and has engineering applications. We call a direction a global descent direction of a function at a point if there is another point with a lower function value along this direction. For a normal quartic polynomial, we present a criterion to find a global descent direction at a noncritical point, a saddle point, or a local maximizer. We give sufficient conditions to judge whether a local minimizer is global and give a method for finding a global descent direction at a local, but not global, minimizer. We also give a formula at a critical point and a method at a noncritical point to find a one-dimensional global minimizer along a global descent direction. Based upon these, we propose a global descent algorithm for finding a global minimizer of a normal quartic polynomial when $n = 2$. For the case $n \geq 3$, we propose an algorithm for finding an ϵ -global minimizer. At each iteration of a second algorithm, a system of constrained nonlinear equations is solved. Numerical tests show that these two algorithms are promising.

Key words. global optimization, normal quartic polynomial, tensor

AMS subject classifications. 90C30, 65K10

DOI. 10.1137/S1052623403420857

1. Introduction. The multivariate polynomial optimization problem has attracted some attention recently [6, 12, 14, 16, 17, 18, 25]. It has applications in signal processing [2, 5, 18, 22, 23]; merit functions of polynomial equations [6]; 0–1 integer, linear, and quadratic programs [12]; nonconvex quadratic programs [12]; and bilinear matrix inequalities [12]. It is related to Hilbert’s 17th problem on the representation of nonnegative polynomials [14, 20].

In [18], Qi and Teo raised the concept of *normal polynomial*. They used tensors to denote coefficients of a multivariate polynomial. They called an even degree polynomial a normal polynomial if its leading degree term coefficient tensor is positive definite. They showed that the multivariate polynomials resulting from signal processing [2, 5, 22, 23] are normal quartic polynomials. They gave a bound for the norms of all global minimizers of a normal polynomial. They pointed out that normal quartic optimization is one of the simplest nontrivial cases of nonconvex global optimization.

In [17], Qi further studied the extrema structure of a real polynomial, in particular, a normal quartic polynomial. Let f be a real polynomial. Qi [17] called a

*Received by the editors January 8, 2003; accepted for publication (in revised form) March 9, 2004; published electronically December 9, 2004.

<http://www.siam.org/journals/siopt/15-1/42085.html>

[†]Department of Applied Mathematics, The Hong Kong Polytechnic University, Kowloon, Hong Kong (maqlq@polyu.edu.hk). The work of this author was supported by the Research Grant Council of Hong Kong and the Australia Research Council.

[‡]College of Mathematics and Econometrics, Hunan University, Changsha, China (wan-math@163.com). This work was done when this author visited the Department of Mathematics and Statistics, Curtin University of Technology, Perth, Australia and was supported by the Australia Research Council.

[§]College of Mathematics and Econometrics, Hunan University, Changsha, China. Current address: Department of Applied Mathematics, the Hong Kong Polytechnic University, Kowloon, Hong Kong (mayfyang@polyu.edu.hk). The work of this author was supported by the Postdoctoral Fellowship of Hong Kong Polytechnic University.

polynomial factor of $f - c_0$, where c_0 is a constant, an essential factor of f . He showed that essential factors of f play an important role in defining critical and extremum surfaces of f . He proved that a normal polynomial has no odd degree essential factors, and all of its even degree essential factors are normal polynomials, up to a sign change. He also showed that a normal quartic polynomial can have at most one local maximizer.

The results of [18] and [17] indicate that there should be some better methods for finding a global minimizer of a normal quartic polynomial than in the general case.

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a nonconvex function. Let $x, y \in \mathbb{R}^n$, $y \neq 0$. We call y a *global descent direction* of f at x if there is a $t \in \mathbb{R}$ such that

$$f(x + ty) < f(x).$$

Clearly, x is a global minimizer of f if and only if it has no global descent directions. It is also obvious that a local descent direction of f at x is a global descent direction of f at x . Hence, if x is not a local minimizer of f , it is easy to find a global descent direction of f at x .

Thus, the next questions are as follows: Given x and y , can we judge if y is or is not a global descent direction of f at x ? If it is, can we easily find a one-dimensional global minimizer of f along this global descent direction? In general, how can we find such a global descent direction? We seek answers to these questions for the case when f is a normal quartic polynomial.

The paper is organized as follows. In section 2, we review the current knowledge on tensors and normal quartic polynomials.

In section 3, we present a criterion for finding a global descent direction of a normal quartic polynomial f at a saddle point or a local maximizer and for judging whether a given direction is a global descent direction of f at a local minimizer. At a critical point of f , we give a formula for finding a one-dimensional global minimizer along a global descent direction.

In section 4, we give a method at a noncritical point of f to find a one-dimensional global minimizer along a global descent direction.

In section 5, we present a method for finding a global descent direction of f at a local minimizer when $n = 2$. For the case $n \geq 3$, we propose a constrained nonlinear equation approach to find a global descent direction of f at a local minimizer. The latter is valid for general global optimization. Some sufficient conditions for judging whether a local minimizer is global are given in section 6.

Based upon the above analysis, in section 7 we propose a global descent algorithm for finding a global minimizer of f when $n = 2$. When $n \geq 3$, we form another algorithm based upon the constrained nonlinear equation approach described in section 5. This algorithm will find an ϵ -global minimizer of f .

In section 8, we describe an application of the proposed method in signal processing and report numerical testing results for these two algorithms. For the first algorithm, which can find a global minimizer of f when $n = 2$, we solve the example in [23] in eight iterations. We then solve ten problems randomly generated. The results show that the maximum iteration number is 10 and the minimum is 4. The computer time for each example is no more than one second. For the second algorithm which finds an ϵ -global minimizer of f for $n \geq 3$, we solve ten problems randomly generated for $n = 3$ and five problems randomly generated for $n = 4$. The maximum iteration number is 21 and the minimum is 6. The computer time for each example is approximately one second. We then solve four problems for a class of special normal quartic polynomials with $n = 6$. After 21–25 iterations, we obtain the global minimizers of these four problems. Some final remarks are given in section 9.

2. Tensor analysis and normal quartic polynomial. We use A and B to denote fourth order totally symmetric tensors, and use $A_{i_1 i_2 i_3 i_4}$, $i_l \in \{1, \dots, n\}$ for $l = 1, 2, 3, 4$ to denote the elements of a fourth order totally symmetric tensor A , i.e.,

$$A_{i_1 i_2 i_3 i_4} = A_{j_1 j_2 j_3 j_4}$$

if $\{i_1, i_2, i_3, i_4\}$ is any reordering of $\{j_1, j_2, j_3, j_4\}$. Let $x \in \mathbb{R}^n$ define

$$Ax^4 := \sum_{i_1, i_2, i_3, i_4=1}^n A_{i_1 i_2 i_3 i_4} x_{i_1} x_{i_2} x_{i_3} x_{i_4}.$$

Similarly, we use M and N to denote third order totally symmetric tensors, P and Q to denote second order totally symmetric tensors, p and q to denote first order totally symmetric tensors, and p_0 and q_0 to denote constants. Thus, the elements of M , P , and p are M_{ijk} , P_{ij} , and p_i , respectively, for $i, j, k = 1, \dots, n$. We also have

$$Mx^3 = \sum_{i,j,k=1}^n M_{ijk} x_i x_j x_k,$$

$$Px^2 = \sum_{i,j=1}^n P_{ij} x_i x_j,$$

and

$$px = \sum_i^n p_i x_i.$$

Then we may denote a quartic polynomial $f : \mathbb{R}^n \rightarrow \mathbb{R}$ as

$$(1) \quad f(x) = Ax^4 + Mx^3 + Px^2 + px + p_0,$$

where A is a fourth order tensor, M is a third order tensor, P is a second order tensor, and p is a first order tensor, while p_0 is a constant.

Actually, a first order tensor p is equivalent to a vector, and a second order tensor P is equivalent to a square matrix. Thus, we have $px = p^T x$ and $Px^2 = x^T P x$. But we will prefer to use the tensor notation px and Px^2 in this paper.

Let $\|\cdot\|$ be a norm in $\mathbb{R}^n$. Denote

$$S := \{x \in \mathbb{R}^n : \|x\| = 1\}.$$

We say that a fourth order totally symmetric tensor A is *positive definite* if

$$Ax^4 > 0$$

for all $x \in S$. This definition extends the definition of positive definite matrices. For a fourth order tensor A , it was defined in [18] that

$$[A] := \min\{Ax^4 : x \in S\}.$$

Clearly, A is positive definite if and only if $[A] > 0$. Similarly, for a second order totally symmetric tensor P ,

$$[P] := \min\{Px^2 : x \in S\},$$

which is the smallest eigenvalue of the symmetric matrix P when we use the 2-norm. We may also define positive semidefinite totally symmetric tensors similarly.

We also define the norm of A , M , P , a , etc. as

$$\|A\| := \max\{|Ax^4| : x \in S\}, \text{ etc.}$$

A quartic or quadratic polynomial is called a normal quartic or quadratic polynomial, respectively, if its leading coefficient tensor is positive definite.

Here are some nice properties for a normal quartic polynomial f , expressed by

(1) [17, 18].

(a) When $\|x\|$ tends to infinity, the value of f will also tend to infinity.

(b) f always has a global minimizer. If x^* is a global minimizer of f , then

$$\|x^*\| \leq L := \max \left\{ 1, \frac{\|M\| + \|P\| + \|p\|}{[A]} \right\}.$$

For a normal quartic polynomial arising in signal processing, a computational bound for its global minimizers was also given in [18].

(c) If f can be written as

$$f(x) = g(x)h(x) + c_0,$$

where g and h are two nonconstant polynomials and c_0 is a constant, then g and h are normal quadratic polynomials up to a sign change of both g and h . In this case, if at least one of the zero sets of g and h is nonempty, then a global minimizer of f can be found in the interiors of one or two ellipsoids defined by the zero sets of g and h , or at the points defined by the zero sets of g and h . See [17] for details.

(d) f has at most one local maximizer.

According to the Bézout theorem [3, 10], if a quartic polynomial f of n variables has only isolated critical points, then the number of these isolated critical points is less than 3^n . In [4], it was shown that a quartic polynomial of two variables has at most five isolated local extremum points if it has only isolated critical points. See also [21]. But a quartic polynomial or even a normal quartic polynomial may have extremum manifold. So these results have not given a real bound on the number of local extremum points.

Surely, if a function f has a connected extremum manifold C , it has infinitely many local minimizers. Thus, we may not count the number of local extremum points in this case. But we may count the number of extremum levels. It is easy to see that if a function f has a connected critical point manifold C , then f has the same value at this manifold [17]. This motivates us to define the following concept: We call a real number c_0 a *minimum level (critical level)* of f if there is a local minimizer (critical point) x of f such that

$$f(x) = c_0.$$

PROPOSITION 1. *The number of minimum (critical) levels of a polynomial is finite.*

Proof. The critical point set of a polynomial is a real algebraic variety [26]. It has at most a finite number of topological components [26]. As stated above, f has the same value on each topological component of this algebraic variety. Hence, the number of critical levels of f is finite. As a minimum level is a critical level, the number of minimum levels of f is also finite. $\square$

This proposition has important implications for developing global descent algorithms for finding a global minimizer of a normal quartic polynomial. We say an algorithm is a global descent algorithm if it can always proceed from a local minimizer, when it is not a global minimizer, to another local minimizer with lower function value. By this proposition, such an algorithm will converge in finitely many iterations.

3. Global descent directions. In this section we study global descent directions of f at critical points, where $f : \mathfrak{R}^n \rightarrow \mathfrak{R}$ is assumed to be a normal quartic polynomial, i.e., its fourth degree term coefficient tensor is positive definite. We first give a criterion to judge the type of critical points. Then we give a sufficient and necessary condition for a direction y to be a global descent direction of f at a critical point x , and a formula to compute a global minimizer of f along y if y is a global descent direction. We further investigate the way to find a global descent direction if x is a local maximizer or a saddle point of f . If x is a local minimizer, we give a criterion to judge whether it is a global minimizer or not. The issue of finding a global descent direction when x is a local minimizer, but not global, will be treated in section 5.

Let $F : \mathfrak{R}^n \rightarrow \mathfrak{R}^n$ be the gradient function of f , i.e.,

$$F = \nabla f.$$

Throughout this section, we assume that x is a critical point of f , i.e.,

$$F(x) = 0.$$

Then we may rewrite f as

$$(2) \quad f(x + ty) = f(x) + t^2 Qy^2 + t^3 Ny^3 + t^4 Ay^4,$$

where $y \in \mathfrak{R}^n, y \neq 0, t \in \mathfrak{R}$, Q is a second order tensor, and N is a third order tensor. It is obvious that

$$Q = \frac{1}{2} f''(x)$$

and

$$N = \frac{1}{6} f'''(x).$$

Hence, it is easy to calculate Q and N .

The next proposition gives a criterion to judge whether a given critical point x of f is a local minimizer, a local maximizer, or a saddle point.

PROPOSITION 2. *Let $f : \mathfrak{R}^n \rightarrow \mathfrak{R}$ be a normal quartic polynomial. Assume that x is a critical point of f . Then the following statements hold.*

- (i) *If Q is negative definite, then x is the unique local maximizer of f .*
- (ii) *If Q is positive definite, or if Q is positive semidefinite and*

$$Ny^3 = 0$$

for all $y \in \{y \in \mathfrak{R}^n : Qy^2 = 0\}$, then x is a local minimizer of f .

- (iii) *Otherwise, x is a saddle point of f .*

Proof. If Q is positive definite, negative definite, or has both positive and negative eigenvalues, the conclusions are clear, as the sign of $f(x + ty) - f(x)$ is dominated by $t^2 Qy^2$ when y is fixed and t is small.

If Q is positive semidefinite and for all

$$y \in \{y \in \mathbb{R}^n : Qy^2 = 0\},$$

$$Ny^3 = 0,$$

then when t is sufficiently small, for

$$y \notin \{y \in \mathbb{R}^n : Qy^2 = 0\},$$

the sign of $f(x + ty) - f(x)$ is dominated by $t^2 Qy^2$, which is always positive, while for

$$y \in \{y \in \mathbb{R}^n : Qy^2 = 0\},$$

we have

$$f(x + ty) - f(x) = t^4 Ay^4 \geq 0,$$

as A is positive definite. Hence, x is a local minimizer in this case.

If Q is positive or negative semidefinite and for some fixed

$$y \in \{y \in \mathbb{R}^n : Qy^2 = 0\},$$

$$Ny^3 \neq 0,$$

then when t is sufficiently small, the sign of $f(x + ty) - f(x)$ is dominated by $t^3 Ny^3$, which changes sign when t changes sign. Hence, x is a saddle point in this case.

If Q is negative semidefinite and for all

$$y \in \{y \in \mathbb{R}^n : Qy^2 = 0\},$$

$$Ny^3 = 0,$$

then for

$$y \notin \{y \in \mathbb{R}^n : Qy^2 = 0\},$$

we have $f(x + ty) - f(x) < 0$ when t is sufficiently small, while for

$$y \in \{y \in \mathbb{R}^n : Qy^2 = 0\},$$

we have

$$f(x + ty) - f(x) = t^4 Ay^4 \geq 0,$$

as A is positive definite. Hence, x is a saddle point in this case.

This exhausted all the cases. $\square$

We still assume that x is a critical point of f . The next issue is to determine whether a given direction $y \in \mathbb{R}^n$ is a global descent direction of f at x .

Denote

$$a \equiv a(y) = Ay^4,$$

$$b \equiv b(y) = -Ny^3,$$

and

$$c \equiv c(y) = Qy^2.$$

Since A is positive definite, we have $a > 0$. Define $\phi : \mathbb{R} \rightarrow \mathbb{R}$ by

$$\phi(t) := f(x + ty) - f(x) \equiv at^4 - bt^3 + ct^2.$$

Clearly, y is a global descent direction of f at x if and only if there is a $t \in \mathbb{R}$ such that $\phi(t) < 0$. If t^* is a global minimizer of ϕ , then $x + t^*y$ is the best candidate for the next iterate of a global descent algorithm for finding a global minimizer of f if we regard x as the current iterate.

Hence, ϕ plays a fundamental role in our discussion of the global descent direction. We call ϕ the *fundamental polynomial* of f at x along the direction y .

We also denote

$$\Delta \equiv \Delta(y) = b^2 - 4ac.$$

The next theorem provides a sufficient and necessary condition for y to be a global descent direction of f at x and provides a formula to compute exactly a global minimizer of the fundamental polynomial.

THEOREM 3 (fundamental polynomial test). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a normal quartic polynomial and x be a critical point of f . Then $y \in \mathbb{R}^n$ is a global descent direction of f at x if and only if*

$$\Delta > 0.$$

Furthermore, if $\Delta > 0$, a global minimizer of ϕ is

$$(3) \quad t^* = \begin{cases} \frac{3b + \sqrt{9b^2 - 32ac}}{8a} & \text{if } b \geq 0, \\ \frac{3b - \sqrt{9b^2 - 32ac}}{8a} & \text{otherwise} \end{cases}$$

and we have

$$(4) \quad \phi(t^*) = -\frac{c^2}{4a} + \frac{9b^2c}{32a^2} - \frac{27b^4}{512a^3} - \frac{\sqrt{b^2(9b^2 - 32ac)^3}}{512a^3} < 0 = \phi(0).$$

Proof. By calculus, if and only if $\Delta \leq 0$, we have $\phi(t) \geq 0 = \phi(0)$, i.e.,

$$f(x + ty) \geq f(x)$$

for all $t \in \mathbb{R}$. This proves the first conclusion.

If $\Delta > 0$, then ϕ has three critical points:

$$t_0 = 0, \quad t_1 = \frac{3b + \sqrt{9b^2 - 32ac}}{8a}, \quad t_2 = \frac{3b - \sqrt{9b^2 - 32ac}}{8a}.$$

In the case $b \geq 0$, we have $\phi(0) = 0$, $\phi(t_1) = \phi(t^*)$ is given by (4), and

$$\phi(t_2) = -\frac{c^2}{4a} + \frac{9b^2c}{32a^2} - \frac{27b^4}{512a^3} + \frac{\sqrt{b^2(9b^2 - 32ac)^3}}{512a^3}.$$

It is not difficult to see that

$$\phi(0) = 0 > \phi(t_1)$$

and

$$\phi(t_2) \geq \phi(t_1).$$

This shows that $t^* = t_1$ is a global minimizer of ϕ .

In the case $b < 0$, the proof is similar. This completes our proof. $\square$

With this theorem, we may analyze global descent directions at a critical point case by case.

The next theorem shows that eigenvectors corresponding to any negative eigenvalue of Q are global descent directions.

THEOREM 4 (negative eigenvalue of Q). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a normal quartic polynomial and x be a critical point of f .*

If $y \in \mathbb{R}^n$ is an eigenvector of Q corresponding to a negative eigenvalue of Q , then y is a global descent direction of f at x , and $x + t^*y$ is a one-dimensional global minimizer of f from x along y , with the function value

$$f(x + t^*y) = f(x) + \phi(t^*) < f(x),$$

where t^* and $\phi(t^*)$ are given by (3) and (4), respectively.

Proof. Since $a > 0$ and $c = Qy^2 < 0$, we have $\Delta > 0$. The conclusions of the theorem follow from Theorem 3. $\square$

When Q has a zero eigenvalue but no negative eigenvalue, by the next theorem, we may also determine whether an eigenvector corresponding to a zero eigenvalue of Q is a global descent direction or not.

THEOREM 5 (zero eigenvalue of Q). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a normal quartic polynomial and x be a critical point of f .*

If $y \in \mathbb{R}^n$ is an eigenvector of Q corresponding to a zero eigenvalue of Q , then y is a global descent direction of f at x if and only if $b \neq 0$.

Furthermore, if $b \neq 0$, then $x + t^*y$ is a one-dimensional global minimizer of f from x along y , with the function value

$$f(x + t^*y) = f(x) + \phi(t^*) < f(x),$$

where

$$t^* = \frac{3b}{4a}$$

and

$$\phi(t^*) = -\frac{27b^4}{256a^3}.$$

Proof. We have $c = Qy^2 = 0$. We now have $\Delta > 0$ if and only if $b \neq 0$. Again, the conclusions of the theorem follow from Theorem 3. $\square$

Note that the formulas about t^* and $\phi(t^*)$ in the above theorem are special cases of (3) and (4) when $c = 0$.

Theorems 4 and 5 actually address the issue of how to compute a global descent direction when x is a local maximizer or a saddle point of f . When x is a local minimizer but not global, it is not an easy task to compute a global descent direction of f . In section 5 we will deal with this case. In the following we give a criterion for judging whether a local minimizer is global.

THEOREM 6 (global minimizers). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a normal quartic polynomial and x be a local minimizer of f . Then x is a global minimizer of f if and only if for all $y \in \mathbb{R}^n, y \neq 0$ with $c \neq 0$, we have $\Delta \leq 0$.*

If for all $y \in \mathbb{R}^n, y \neq 0$ with $c \neq 0$, we have $\Delta < 0$, then x is the unique global minimizer of f .

If for all $y \in \mathbb{R}^n, y \neq 0$ with $c \neq 0$, we have $\Delta \leq 0$, and for some $y \in \mathbb{R}^n, y \neq 0$ with $c \neq 0$, we have $\Delta = 0$, then x is a global minimizer of f but there are other global minimizers of f .

Proof. Since x is a local minimizer, we have $c \geq 0$, and $c = 0$ implies $b = 0$ by Proposition 2(ii).

If x is a global minimizer, then there does not exist a global descent direction. Thus by Theorem 3, $\Delta(y) \leq 0$ for all $y \neq 0$. This proves the “only if” part of the first assertion.

For the “if” part of the first assertion, we proceed by contradiction. Assume the contrary, i.e., x is not a global minimizer. Then there exists a global descent direction $y \neq 0$. By Theorem 3 we have $\Delta(y) > 0$. By the “if” condition, we have $c = 0$. This implies $b = 0$ by the argument at the beginning of this proof. Hence $\phi(t) = at^4 > 0$, which contradicts the assumption that y is a global descent direction.

We now prove the two other assertions.

Suppose that $\Delta < 0$ for all $y \in \mathbb{R}^n, y \neq 0$ with $c \neq 0$. Let $y \neq 0$. Again, by the argument at the beginning of this proof, if $c = 0$, then $b = 0$ and thus

$$\phi(t) = at^4 > 0$$

for all $t \neq 0$. If $c \neq 0$, then since $a > 0$ and $\Delta < 0$, we also have

$$\phi(t) > 0$$

for all $t \neq 0$. Hence x is the unique global minimizer of f .

Similarly, if $\Delta \leq 0$ for all $y \in \mathbb{R}^n, y \neq 0$ with $c \neq 0$, then x is a global minimizer of f . But if $\Delta = 0$ for some $y \in \mathbb{R}^n, y \neq 0$ with $c \neq 0$, then $x + t^*y$ will be another global minimizer of f , where $t^* = \frac{b}{2a}$. In this case, x is a global minimizer but not a unique global minimizer of f . $\square$

4. One-dimensional normal quartic minimization at a noncritical point.

In this section we study the global minimizers of the normal quartic polynomial f at a noncritical point along a global descent direction, which can be reformulated as a one-dimensional global minimization. To address this issue, we propose an approach based on the one-dimensional Newton method.

Assume that x is not a critical point of f , i.e.,

$$q := F(x) \neq 0.$$

Let $y \in \mathbb{R}^n$ such that

$$(5) \quad qy < 0.$$

Then y is a descent direction, and hence a global descent direction of f at x .

Our aim is to find a one-dimensional global minimizer of f on the line $\{x + ty : t \in \mathbb{R}\}$, which is equivalent to finding a global minimizer of

$$\phi(t) := f(x + ty) - f(x).$$

We may rewrite

$$(6) \quad \phi(t) = at^4 - bt^3 + ct^2 - dt,$$

where the meanings of a, b , and c are defined in the last section, while $d = -qy$.

We now have $a > 0$ and $d > 0$. Denote

$$\psi(t) := \phi'(t) = 4at^3 - 3bt^2 + 2ct - d$$

and

$$\eta(t) := \psi'(t) = 12at^2 - 6bt + 2c.$$

Let

$$(7) \quad \Delta_1 = 9b^2 - 24ac.$$

The next proposition states a useful property of ψ .

PROPOSITION 7. *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a normal quartic polynomial and x be a noncritical point of f . Suppose that y is a descent direction satisfying (5). If $\Delta_1 > 0$, then ψ has a local maximizer*

$$(8) \quad \bar{t}_1 = \frac{3b - \sqrt{\Delta_1}}{12a}$$

and a local minimizer

$$(9) \quad \bar{t}_2 = \frac{3b + \sqrt{\Delta_1}}{12a}$$

with $\bar{t}_1 < \bar{t}_2$ and $\psi(\bar{t}_1) > \psi(\bar{t}_2)$.

Proof. If $\Delta_1 > 0$, then $\bar{t}_1$ and $\bar{t}_2$ are two zeros of the quadratic polynomial η , and hence two critical points of the cubic polynomial ψ . We also have

$$\eta'(\bar{t}_1) = -2\sqrt{\Delta_1} < 0$$

and

$$\eta'(\bar{t}_2) = 2\sqrt{\Delta_1} > 0,$$

i.e., $\bar{t}_1$ is a local maximizer of ψ and $\bar{t}_2$ is a local minimizer of ψ . Since $a > 0$, we have

$$\bar{t}_2 - \bar{t}_1 = \frac{\sqrt{\Delta_1}}{6a} > 0.$$

Since ψ is a cubic polynomial, $\bar{t}_1$ is a local maximizer and $\bar{t}_2$ is a local minimizer, we have $\psi(\bar{t}_1) > \psi(\bar{t}_2)$. $\square$

Denote

$$(10) \quad \bar{t} = \frac{b}{4a},$$

$$(11) \quad \bar{t}_0 = \max \left\{ 1, \frac{3b_+ + 2(-c)_+ + d}{4a} \right\},$$

and

$$(12) \quad \hat{t}_0 = \min \left\{ -1, -\frac{3(-b)_+ + 2(-c)_+}{4a} \right\},$$

where for any $\alpha \in \mathbb{R}$, $\alpha_+ := \max\{\alpha, 0\}$. It is not difficult to deduce $\psi(t) > 0$ for all $t > \bar{t}_0$ and $\psi(t) < 0$ for all $t < \hat{t}_0$.

Since ϕ is a one-dimensional normal quartic polynomial, it has one or two local minimizers. We distinguish these two different cases in the next proposition and give illustrations of the various situations in Figures 1–4.

PROPOSITION 8. *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a normal quartic polynomial and x be a noncritical point of f . Suppose that y is a descent direction satisfying (5).*

(i) *If one of the following three cases occurs, then ϕ has a unique local, hence global minimizer t^* :*

- (a) $\Delta_1 \leq 0$;
- (b) $\Delta_1 > 0$ and $\psi(\bar{t}_2) \geq 0$;
- (c) $\Delta_1 > 0$ and $\psi(\bar{t}_1) \leq 0$.

Furthermore, in case (a), ϕ is convex, and

$$0 < t^* \leq \bar{t}_0.$$

In case (b), we have

$$0 < t^* < \bar{t}_1.$$

In case (c), we have

$$(\bar{t}_2)_+ < t^* < \bar{t}_0.$$

(ii) *If $\Delta_1 > 0$ and $\psi(\bar{t}_1) > 0 > \psi(\bar{t}_2)$, then ϕ has two local minimizers t_1^* and t_2^* such that*

$$t_1^* < \bar{t}_1 < \bar{t}_2 < t_2^*.$$

Proof. (i) If case (a) occurs, then $\eta(t) \geq 0$ for all t . This implies that ϕ is convex. Since ϕ is a nonconstant convex polynomial, it can have one local minimizer t^* , which is also a global minimizer. Since $\psi(0) = -d < 0$, $\psi(t^*) = 0$, and $\psi(\bar{t}_0) \geq 0$, we have $0 < t^* \leq \bar{t}_0$.

If case (b) occurs, we have

$$\psi(\bar{t}_1) > \psi(\bar{t}_2) \geq 0.$$

Since $\psi(0) < 0$, we have $t^* \in (0, \bar{t}_1)$ such that $\psi(t^*) = 0$. We now have

$$0 < t^* < \bar{t}_1 < \bar{t} < \bar{t}_2.$$

Since $\eta(t) > 0$ for all $t < \bar{t}_1$, t^* is a local minimizer of ϕ .

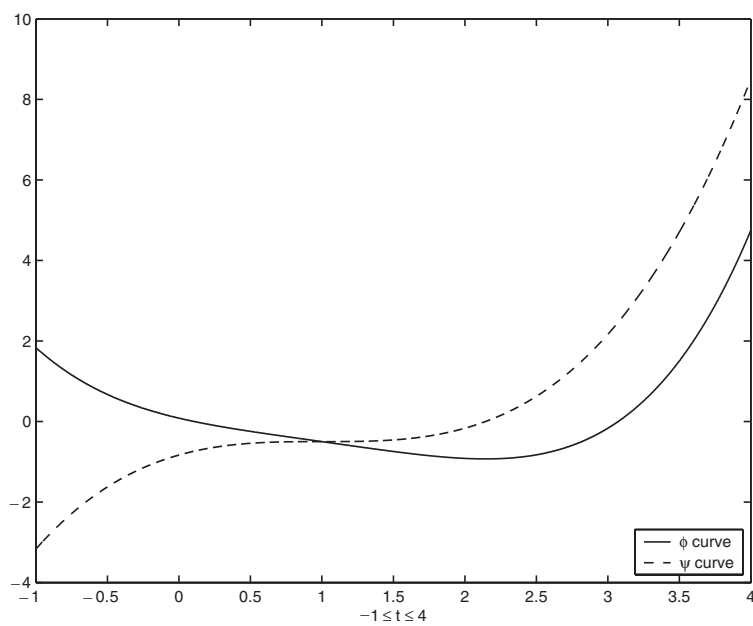


FIG. 1. Graphs of ϕ and ψ when $\phi(t) = \frac{1}{12}(t-1)^4 - 0.5t$.

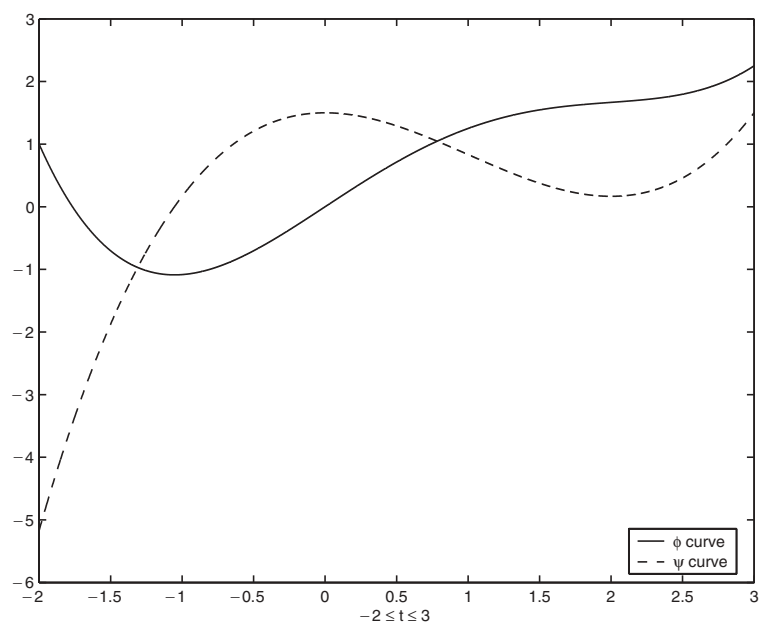


FIG. 2. Graphs of ϕ and ψ when $\phi(t) = \frac{1}{12}t^4 - \frac{1}{3}t^3 + 1.5t$.

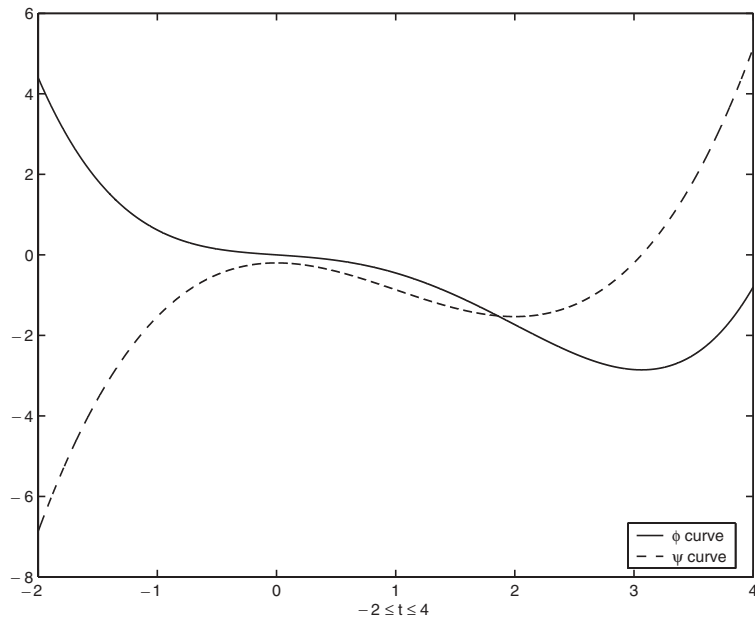


FIG. 3. Graphs of ϕ and ψ when $\phi(t) = \frac{1}{12}t^4 - \frac{1}{3}t^3 - 0.2t$.

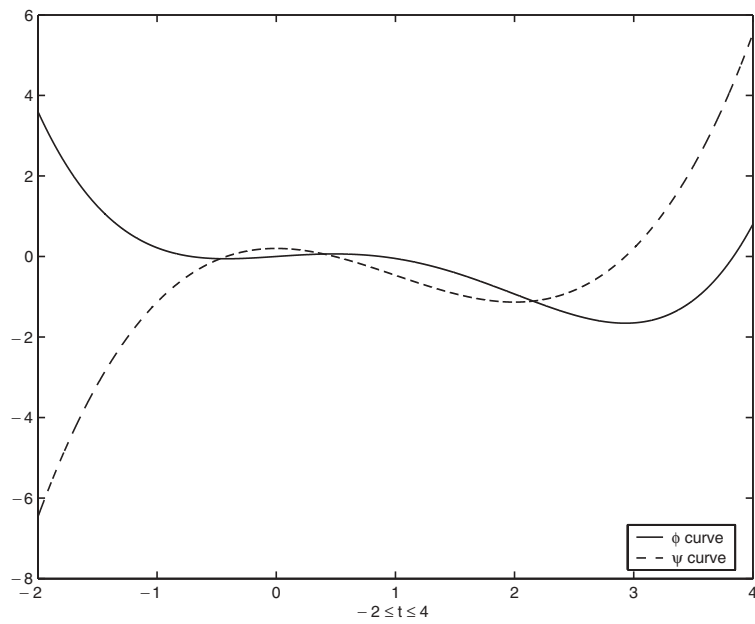


FIG. 4. Graphs of ϕ and ψ when $\phi(t) = \frac{1}{12}t^4 - \frac{1}{3}t^3 + 0.2t$.

If $\psi(\bar{t}_2) > 0$, then t^* is the only zero of ψ . If $\psi(\bar{t}_2) = 0$, then $\bar{t}_2$ is also a zero of ψ . But since $\psi(t) > 0$ when t is close to but not equal to $\bar{t}_2$, in this case $\bar{t}_2$ is a saddle point of ϕ . In both cases, t^* is the unique minimizer of ϕ . Since ϕ is a normal quartic polynomial, it is also the unique global minimizer of ϕ .

If case (c) occurs, we have

$$\bar{t} < (\bar{t}_2)_+ < t^* < \bar{t}_0.$$

The remaining proof is similar to that of the last case.

(ii) Since $a > 0$, $\psi(\bar{t}_1) > 0 > \psi(\bar{t}_2)$, and $\bar{t}_1 < \bar{t}_2$, we have

$$t_1^* < \bar{t}_1 < \bar{t}_2 < t_2^*.$$

Actually, we have

$$\hat{t}_0 \leq t_1^* < \bar{t}_1 < \bar{t} < \bar{t}_2 < t_2^* \leq \bar{t}_0. \quad \square$$

It follows from the above proposition that if ϕ has only one local minimizer, then it is also the unique global minimizer of ϕ . We now state an algorithm for finding the global minimizer of ϕ in this case.

ALGORITHM 1.

Step 0. Compute Δ_1 , $\bar{t}$, and $\bar{t}_0$ by (7), (10), and (11), respectively. Set $j := 0$.

Step 1. If $\psi(\bar{t}_0) = 0$, then set $t^* := \bar{t}_0$ and stop.

Step 2. If $\Delta_1 \leq 0$, then go to Step 3; otherwise go to Step 5.

Step 3. If $\psi(\bar{t}) = 0$, then set $t^* := \bar{t}$ and stop.

Step 4. If $\psi(\bar{t}) > 0$, then set $t_0 := 0$; otherwise, set $t_0 := \bar{t}_0$. Go to Step 6.

Step 5. If $\psi(\bar{t}_2) \geq 0$, then set $t_0 := 0$; otherwise set $t_0 := \bar{t}_0$.

Step 6. Compute the next iterate point by using the one-dimensional Newton method

$$(13) \quad t_{j+1} := t_j - \frac{\psi(t_j)}{\eta(t_j)}.$$

Step 7. Set $j := j + 1$. Go to Step 6.

The following is the convergence statement of the above algorithm.

THEOREM 9 (single minimizer of ϕ). Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a normal quartic polynomial and x be a noncritical point of f . Suppose that y is a descent direction satisfying (5) and that the assumption in Proposition 8(i) holds.

Then Algorithm 1 is well defined and the generated sequence $\{t_j\}$ converges to the unique global minimizer t^* of ϕ quadratically. Furthermore, $\{t_j\}$ strictly increases to t^* in the case where we choose $t_0 = 0$, and strictly decreases to t^* in the case where we choose $t_0 = \bar{t}_0$.

Proof. We prove only the convergence result when Step 4 in Algorithm 1 is executed. For the other two cases, the proof is similar.

Assume that $\psi(1) \neq 0$ and $\psi(\bar{t}) \neq 0$. Then $\psi(\bar{t}_0) > 0$; otherwise, $\psi(\bar{t}_0) = 0$. It is not difficult to deduce $\bar{t}_0 = 1$, a contradiction. This, together with (i)(a) of the last proposition, implies $0 < t^* < \bar{t}_0$.

If $\psi(\bar{t}) > 0$, then $0 < t^* < \bar{t}$. If $t_j \in [0, t^*)$, then $\psi(t_j) < 0$. We also have $\eta(t_j) > 0$, as $\bar{t}$ is the only possible zero of η (only when $\Delta_1 = 0$). Then

$$t_{j+1} := t_j - \frac{\psi(t_j)}{\eta(t_j)} > t_j.$$

By the mean value theorem, there is an $\xi \in (t_j, t^*)$ such that

$$\eta(\xi)(t^* - t_j) = \psi(t^*) - \psi(t_j) = -\psi(t_j).$$

Since $\eta'(t) < 0$ for all $t < \bar{t}$, as $t_j < \xi$, we have

$$\eta(t_j) > \eta(\xi).$$

Hence,

$$\eta(t_j)(t^* - t_j) > \eta(\xi)(t^* - t_j) = -\psi(t_j).$$

As $\eta(t_j) > 0$, we have

$$t^* - t_j > -\frac{\psi(t_j)}{\eta(t_j)}.$$

This implies that

$$t_j < t_{j+1} < t^*.$$

Hence, if we let $t_0 = 0$, the one-dimensional Newton method (13) is well defined and $\{t_j\}$ strictly increases to a limit. By the property of the Newton method, this limit must be a zero of ψ , which is t^* . As $\eta'(t^*) \neq 0$, the sequence converges quadratically to t^* .

If $\psi(\bar{t}) < 0$, then $\bar{t} < t^* < \bar{t}_0$. By taking $t_0 = \bar{t}_0$, similar to the above proof, we can deduce the corresponding convergence result. $\square$

When ϕ has two local minimizers t_1^* and t_2^* , we can use a similar method to find them. The differences lie in that

(i) we also calculate $\hat{t}_0$ by (12);

(ii) if $\psi(\hat{t}_0) = 0$, we have $t_1^* = \hat{t}_0$. If $\psi(\bar{t}_0) = 0$, we have $t_2^* = \bar{t}_0$. Otherwise, we use the one-dimensional Newton method (13) to find t_1^* and t_2^* with different choices of the starting point t_0 . Specifically speaking, we may use $t_0 = \bar{t}_0$ to find t_2^* . If $0 < \bar{t}_1$, we may use $t_0 = 0$ to find t_1^* ; otherwise, we use $t_0 = \hat{t}_0$ to find t_1^* .

We call such a modified algorithm Algorithm 2. The next theorem gives its convergence statement.

THEOREM 10 (two local minima of ϕ). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a normal quartic polynomial and x be a noncritical point of f . Suppose that y is a descent direction satisfying (5) and that the assumption in Proposition 8(ii) holds.*

Then Algorithm 2 is well defined, and the two sequences generated by (13) converge to two local minimizers t_1^* and t_2^* of ϕ quadratically, with one sequence strictly increasing to t_1^* and another sequence strictly decreasing to t_2^* .

The proof of the above theorem is omitted since it is similar to that of Theorem 9.

We may also find one of t_1^* and t_2^* , say t_1^* , first. Then we may find another of them, say t_2^* , by solving the quadratic polynomial

$$\frac{d}{dt} \left(\frac{\phi(t)}{t - t_1^*} \right).$$

After finding t_1^* and t_2^* , comparing the values of $\phi(t_1^*)$ and $\phi(t_2^*)$, we may find a global minimizer t^* of ϕ .

Since minimizing ϕ is only a subproblem of minimizing f , we only need to find an approximate global minimizer t^* of f such that $\phi(t^*) < 0$ and $|\psi(t^*)| \leq \epsilon$ for some given $\epsilon > 0$.

5. Global descent directions at a local minimizer. At a noncritical point of f , we may use the method in section 4 to find a one-dimensional global minimizer of f along this descent direction and its opposite direction. At a saddle point or the unique local maximizer of f , we may identify a global descent direction by Theorems 4 and 5 and use the formula in section 3 to calculate a one-dimensional global minimizer of f along this global descent direction. Actually, if we proceed in this way, we will not meet the unique local maximizer of f , unless we by chance use it as the starting point, since each time, the iterate is a one-dimensional global minimizer along a line. Now, the only difficult points are local minimizers of f . Theorem 3 provides a criterion for checking if a given direction is a global descent direction of f at a local minimizer, but does not provide a constructive way for finding or identifying a global descent direction.

In this section, we discuss this problem. Assume that x is a local minimizer of f . We use the same notation as in section 3. By Theorem 3, we may solve the following problem to obtain a global descent direction y :

$$(14) \quad \max \Delta(y) \quad \text{subject to } y \in S.$$

If we use the infinity norm, then the maximization problem (14) can be converted to $2n(n-1)$ -dimensional maximization problems:

$$(15) \quad \max \Delta(y) \quad \text{subject to } y_i = 1, |y_j| \leq 1 \text{ for } j \neq i,$$

and

$$(16) \quad \max \Delta(y) \quad \text{subject to } y_i = -1, |y_j| \leq 1 \text{ for } j \neq i$$

for $i = 1, \dots, n$.

When $n = 2$, (15) and (16) are four one-dimensional sixth degree polynomial maximization problems. They are practically solvable. In fact, it follows from Theorem 3 that the following result holds.

THEOREM 11. *Assume that x is a local minimizer of f and $n = 2$. If one of the four one-dimensional sixth degree polynomial maximization problems in (15) and (16) has a feasible point y such that $\Delta(y) > 0$, then y is a global descent direction of f at x . Otherwise, i.e., when all of them have no positive optimal function values, x is a global minimizer of f .*

When $n \geq 3$, it seems that there are no advantages to solving (15) and (16) other than solving the normal quartic minimization problem

$$(17) \quad \min \{Ay^4 + Ny^3 + Qy^2 : y \in \mathbb{R}^n\},$$

which is equivalent to the original problem. If (17) has a feasible solution y such that

$$Ay^4 + Ny^3 + Qy^2 < 0,$$

then y is a global descent direction of f at x . Otherwise, x is already a global minimizer of f .

For $n \geq 3$, we propose a constrained nonlinear equation approach for finding a global descent direction at a local minimizer. This approach is also valid for the general global optimization. Hence, in the following proposition only, f is a general nonconvex function.

Let $\epsilon > 0$. We say x^* is an ϵ -global minimizer of a nonconvex function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ if for all $x \in \mathbb{R}^n$,

$$f(x) > f(x^*) - \epsilon.$$

The next proposition gives an approach to compute a global descent direction at a local minimizer.

PROPOSITION 12. x^* is an ϵ -global minimizer of a nonconvex function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ if and only if the bound constrained system of nonlinear equations

$$(18) \quad \begin{cases} F(x) &= 0, \\ f(x) - x_{n+1} &= 0, \\ x_{n+1} - f(x^*) + \epsilon &\leq 0 \end{cases}$$

has no feasible solution. If (18) has a feasible solution $\bar{x}$, then $y = \bar{x} - x^*$ is a global descent direction of f at x^* .

This proposition is easy to prove. In the literature, there are many methods for solving bound constrained systems of nonlinear equations [1, 7, 8, 9, 11, 13, 19, 24].

6. Criteria for identifying a global minimizer. Analogous to local optimality conditions in the ordinary nonlinear programming, it is very important to provide appropriate computable termination rules for the algorithms to compute a global minimizer. We call such a termination rule the *global optimality condition*.

Although Theorem 6 presents a sufficient and necessary condition for a local minimizer to be global, it is usually not checkable in the general case. For $n = 2$, Theorem 11 shows that a local minimizer is global if the problems (15) and (16) have no positive optimal function value; however, Theorem 11 is not usable for the case $n \geq 3$. In this section, we present some checkable criteria to judge whether a local minimizer is global when $n \geq 3$.

Throughout this section, we assume that f is a normal quartic polynomial and x is a local minimizer of f . In the next proposition we give a sufficient condition for x to be a global minimizer of f .

PROPOSITION 13. Assume that f is a normal quartic polynomial and x is a local minimizer of f . If $\|N\|^2 \leq 4[A][Q]$, then x is a global minimizer of f .

This follows directly from Theorem 3.

The following corollary is a direct consequence of Propositions 13 and 2.

COROLLARY 14. Assume that f is a normal quartic polynomial and x is a local minimizer of f . If $N = 0$, then x is a global minimizer of f .

As mentioned in section 2, $[Q]$ is the smallest eigenvalue of the symmetric matrix Q when we use the 2-norm. We also have

$$\|N\|_2 \leq \|N\|_\infty \leq \sum_{i,j,k=1}^n |N_{ijk}|.$$

But this upper bound for $\|N\|_2$ may be too large. We may improve it.

Let $N_{i..}$ denote the second order totally symmetric tensor with elements N_{ijk} , where i is fixed. By means of $N_{i..}$, we obtain an upper bound of $\|N\|_2$.

PROPOSITION 15. For any third order tensor N , we have

$$\|N\|_2 \leq \|z\|_2,$$

where $z \in \mathbb{R}^n$ and

$$z_i = \|N_{i..}\|_2.$$

Proof.

$$\begin{aligned}
\|N\|_2 &= \max\{\sum_{i,j,k=1}^n |N_{ijk}x_i x_j x_k| : \|x\|_2 = 1\} \\
&= \max\{\sum_{i=1}^n |x_i N_{i..} x^2| : \|x\|_2 = 1\} \\
&\leq \max\{\sum_{i=1}^n |x_i| \|N_{i..}\|_2 : \|x\|_2 = 1\} \\
&\leq \|z\|_2.
\end{aligned}
\quad \square$$

Let $P^{(0)}$ and $Q^{(0)}$ be two second order totally symmetric tensors. Then we use $P^{(0)} \times Q^{(0)}$ to denote the outproduct of $P^{(0)}$ and $Q^{(0)}$. In particular, we have

$$P^{(0)} \times Q^{(0)} x^4 = P^{(0)} x^2 Q^{(0)} x^2.$$

We may also have the addition of two fourth order totally symmetric tensors. It is easy to prove the following proposition, which provides a lower bound of $[A]$.

PROPOSITION 16. *If $P^{(0)}$ and $Q^{(0)}$ are two second order positive definite totally symmetric tensors, and $P^{(i)}$ and $Q^{(i)}$ for $i = 1, \dots, k$ are second order positive semidefinite totally symmetric tensors, then*

$$(19) \quad A = \sum_{i=0}^k P^{(i)} \times Q^{(i)}$$

is a fourth order positive definite totally symmetric tensor, and

$$[A] \geq \sum_{i=0}^k [P^{(i)}][Q^{(i)}].$$

Based on Propositions 13–16, we deduce a checkable sufficient condition for a local minimizer to be global as follows.

COROLLARY 17. *Assume that f is a normal quartic polynomial and x is a local minimizer of f . If the leading term coefficients tensor A can be expressed by (19) and*

$$(20) \quad \|z\|_2^2 \leq 4 \sum_{i=0}^k [P^{(i)}][Q^{(i)}][Q]$$

holds, where z is defined as in Proposition 15, $P^{(i)}$, $Q^{(i)}$ come from (19), and Q comes from (2), then x is a global minimizer of f .

Note that (19) also gives a way to generate normal quartic polynomials. Actually, we may generate A by (19) and generate M , P , p , and p_0 randomly.

7. Global descent algorithms. We use x_i to denote the i th component of $x \in \mathbb{R}^n$ and $x^{(k)}$ to denote different points in $\mathbb{R}^n$.

We now state an algorithm for minimizing f when $n = 2$.

ALGORITHM 3.

Step 0. Have an initial point $x^{(0)} \in \mathbb{R}^n$, $\rho > 2$, $\epsilon_0 \geq 0$, $\epsilon_1, \epsilon_2 > 0$. Usually we may choose $x^{(0)} = 0$. Let $k = 0$.

Step 1 (noncritical point). If

$$\|F(x^{(k)})\| \leq \epsilon_0,$$

go to Step 2. Otherwise, calculate the Newton direction

$$d = -(\nabla F(x^{(k)}))^{-1} F(x^{(k)}).$$

If the Newton direction d does not exist or if

$$d^T F(x^{(k)}) \geq -\epsilon_1 \|d\|^\rho,$$

then let $y = -F(x^{(k)})$. Otherwise, let $y = d$. Then we use Algorithm 1 or Algorithm 2 to calculate t^* as an approximate global minimizer of ϕ defined by (6) such that $\phi(t^*) < 0$ and $|\psi(t^*)| \leq \epsilon_2$. Go to Step 6.

Step 2 (negative eigenvalue). If Q has a negative eigenvalue, let y be an eigenvector corresponding to this eigenvalue. Go to Step 5.

Step 3 (zero eigenvalue). If Q has a zero eigenvalue and there is an eigenvector y corresponding to this eigenvalue such that $b \neq 0$, go to Step 5.

Step 4 (local minimizer). Solve four maximization problems in (15) and (16) to find a feasible point y of (15) and (16) such that $\Delta(y) > 0$. If such a y can be found, go to the next step. If no such y can be found, then $x^{(k)}$ is a global minimizer. Stop.

Step 5. Use (3) to calculate t^* .

Step 6. Let

$$x^{(k+1)} = x^{(k)} + t^* y.$$

Let

$$k := k + 1$$

and go to Step 1.

At Step 1 of the above algorithm we determine whether the current iterate x^k is a critical point of f .

In the case where x^k is a critical point of f , when x^k is a saddle point or a local maximizer, it follows from Theorems 4 and 5 that y obtained from Step 2 or Step 3 is a global descent direction. When x^k is a local minimizer, by Theorem 11, y obtained from Step 4 is a global descent direction, and if all of the problems in (15) and (16) have no positive optimal function value, then $x^{(k)}$ is a global minimizer.

In the case where x^k is not a critical point of f , we first compute the Newton direction. If it exists and is acceptable, then it is used as a global descent direction; otherwise, the negative gradient direction of f is used. Then we use Algorithm 1 or Algorithm 2 to calculate the global minimizer of ϕ . Theorems 9 and 10 provided convergence results for these two algorithms.

Combining the above analysis with the discussion in sections 2–5, we have the following convergence theorem.

THEOREM 18. *Let $\epsilon_0 = 0$ in Algorithm 3. Assume that $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ is a normal quartic polynomial. Then the sequence generalized by Algorithm 3 is globally convergent to a global minimizer x^* of f . If $\nabla F(x^*)$ is nonsingular, the convergence is quadratic.*

Proof. By Proposition 1, the sequence generalized by Algorithm 3 is globally convergent to a global minimizer x^* of f .

The second assertion follows directly from the properties of the Newton method. $\square$

When $n \geq 3$, we may use some sufficient condition given in section 6 to determine whether a local minimizer x of f is global. If x is not a global minimizer of f , we may solve (18) to find a global descent direction y .

Denote $z = (x^T, x_{n+1})^T$, $\Omega = \{z : x_{n+1} \leq f(x^*) - \epsilon\}$,

$$\Phi(z) = \begin{pmatrix} F(x) \\ f(x) - x_{n+1} \end{pmatrix}, \quad \text{and} \quad \Psi(z) = \frac{1}{2} \|\Phi(z)\|^2.$$

In the following, we describe an algorithm for solving the bound constrained system of nonlinear equations (18), which is similar to that in [13].

ALGORITHM 4.

Step 0. Given constants $\sigma \in (0, 1)$, $\theta \in [0, 1)$, $\eta \in (0, 1)$, $\varepsilon > 0$ and $M > 0$. Let $\gamma = 1 - \theta^2$, $\alpha_0 = 1$, and $z^{(0)} = (x^{*T}, f(x^*) - 1.1\epsilon)^T \in \Omega$. Let $l = 0$.

Step 1. If $\|\Phi(z^{(l)})\| \leq \varepsilon$, stop.

Step 2. Choose $d^{(l)} \in \mathbb{R}^{n+1}$ such that

$$(21) \quad \begin{aligned} & z^{(l)} + d^{(l)} \in \Omega, \quad \|d^{(l)}\|_\infty \leq M \quad \text{and} \\ & \|\Phi'(z^{(l)})d^{(l)} + \Phi(z^{(l)})\| \leq \theta \|\Phi(z^{(l)})\|. \end{aligned}$$

If such a choice is not possible, stop.

Step 3. If

$$\Psi(z^{(l)} + \alpha_l d^{(l)}) < \Psi(z^{(l)}),$$

then set $z^{l+1} = z^{(l)} + \alpha_l d^{(l)}$. Otherwise, let $\alpha_l := \eta \alpha_l$ and go to Step 3.

Step 4. If

$$\Psi(z^{(l)} + \alpha_l d^{(l)}) \leq (1 - \sigma \gamma \alpha_l) \Psi(z^{(l)}),$$

then set $\alpha_{l+1} := 1$. Otherwise, set $\alpha_{l+1} := \eta \alpha_l$.

Let $l := l + 1$ and go to Step 1.

At Step 2 of Algorithm 4, any reasonable approach for computing $d^{(l)}$ can be used. In this paper we consider the auxiliary subproblem

$$(22) \quad \min \|\Phi'(z^{(l)})d + \Phi(z^{(l)})\|^2 \quad \text{subject to} \quad z^{(l)} + d \in \Omega, \quad \|d\|_\infty \leq M.$$

Many methods for solving the above problem have been proposed in the literature; see, e.g., [7, 8]. Following the proof of Theorem 2.2 in [13], it is not difficult to deduce the following convergence result.

THEOREM 19. Assume that $\{z \in \Omega : \Psi(z) \leq \Psi(z^{(0)})\}$ is bounded. Let $\{z^{(l)}\}$ be a sequence generated by Algorithm 4. Then every limit point of $\{z^{(l)}\}$ is a solution of (18).

If Algorithm 4 stops at Step 1, then $y = x^{(l)} - x^*$ can be used as a global descent direction of f at x^* .

If Algorithm 4 stops at Step 2, i.e., there is no direction $d^{(l)}$ satisfying (21), then when we impose that θ is close to 1 and M is large, $z^{(l)}$ is close to a stationary point of Ψ , which is not a solution of $\Phi(z) = 0$. In this case, we may use the method proposed in [15] to go further to find a solution of $\Phi(z) = 0$. The method in [15] always works.

Actually, since we need only find a global descent direction of f at x^* , we do not need to solve (18) completely. Hence, we may change Step 1 of Algorithm 4 to the following:

Step 1'. If $|\Phi_{n+1}(z^{(l)})| \leq 0.5\varepsilon$, stop.

In Step 4 of Algorithm 3, we may use Algorithm 4 instead of solving (15) and (16) to find a global descent direction of f . We call such an algorithm Algorithm 5 and

use it to find an ϵ -global minimizer of f when $n \geq 3$. Sufficient conditions given in section 6 can be used as the termination condition for identifying the global minimizer.

THEOREM 20. *Let $\epsilon_0 = 0$ in Algorithm 5. Assume that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a normal quartic polynomial. Then the sequence generalized by Algorithm 5 is globally convergent to a ϵ -global minimizer x^* of f . If $\nabla F(x^*)$ is nonsingular, the convergence is quadratic.*

The proof of this theorem is the same as the proof of Theorem 18. Note that without the property indicated in Proposition 1, both Theorems 18 and 20 do not hold. This exploited the characteristics of polynomials.

8. An application in signal processing and numerical tests. In this section, we first describe an application of the proposed method in signal processing. Then we report some results of numerical experiments of the algorithms described in the last section.

In the area of broadband antenna array signal processing, the following global optimization problem often arises (see [22, 23]):

$$(23) \quad \begin{array}{ll} \min_w & f_0(w) \\ \text{subject to} & g_l(w) = 0, l = 1, \dots, (N_l + N_q), \end{array}$$

where $f_0 : \mathbb{R}^n \rightarrow \mathbb{R}$ is a strictly convex multivariate polynomial of degree 2, $g_l : \mathbb{R}^n \rightarrow \mathbb{R}, l = 1, \dots, (N_l + N_q)$ are multivariate polynomials of degree at most 2, N_l denotes the number of linear constraints, N_q denotes the number of quadratic constraints, and $w \in \mathbb{R}^n$ is the n -tuple real weight vector.

In [23], Thng, Cantoni, and Leung showed that the problem (23) is equivalent to the global minimization of a quartic multivariate polynomial. Instead of finding all common zeros of a set of multivariate cubic polynomials as in [23], we use the algorithms proposed in this paper to compute the global minimizer of the quartic polynomial.

As an example, we apply our method to solve a 70-tuple problem listed in Appendix C in [23]. As shown in [23], this problem can be transformed into globally minimizing a bivariate quartic polynomial as follows:

$$\begin{aligned} \mathcal{X}(\alpha_1, \alpha_2) = & 0.337280011659804177 - 0.122071359035091510\alpha_1^2 \\ & + 0.077257128600040819\alpha_1^4 - 0.217646697603541049\alpha_1\alpha_2 \\ & + 0.233083387816363887\alpha_1^3\alpha_2 - 0.129244611969892874\alpha_2^2 \\ & + 0.286227131697582205\alpha_1^2\alpha_2^2 + 0.1755719525003619673\alpha_1\alpha_2^3 \\ & + 0.0567691913792773433\alpha_2^4. \end{aligned}$$

Before solving the above problem, we give a general computable criteria to judge whether or not any quartic polynomial is normal for the case $n = 2$.

Given a quartic polynomial $f : \mathbb{R}^n \rightarrow \mathbb{R}$ in the form of (1), f is normal if and only if $[A] := \min\{Ax^4 : x \in S\} > 0$. If we also use the infinity norm, then we can compute $[A]$ for $n = 2$ by solving the following four minimization problems:

$$(24) \quad \min\{Ax^4 : x_1 = 1, |x_2| \leq 1\}, \quad \min\{Ax^4 : x_2 = 1, |x_1| \leq 1\},$$

and

$$(25) \quad \min\{Ax^4 : x_1 = -1, |x_2| \leq 1\}, \quad \min\{Ax^4 : x_2 = -1, |x_1| \leq 1\}.$$

Since (24) and (25) are four at most fourth degree univariate polynomial minimization problems on the interval $[-1, 1]$, it is easy to find their global minimum

TABLE 1
The numerical results when $n = 2$.

Question	GMP (x^*, y^*)	MOF	IN
Q21	(1.11753212587544, 0.17682220411240)	-10.78897112065468	10
Q22	(-0.20087198302720, 0.39084435467521)	-17.36200487514549	5
Q23	(0.09135514315091, -0.84352027358099)	-31.14019175696217	4
Q24	(-0.49521459227644, -0.74362935904160)	-92.28342983346143	7
Q25	(0.30109168886, -1.96636703726)	1.55127331116065	5
Q26	(6.50340840, -9.42272282)	1288925.94605849	7
Q27	(-1.348191548098, 0.675076327936)	-103.344964605849	7
Q28	(0.34648622547009, 0.4125003776761)	56.94847262588482	4
Q29	(0.71784344987734, 0.25545015129099)	-83.95500103898701	5
Q210	(0.3397935539523, -0.15034069197633)	-15.38564238623246	4

solutions. Thus, we can determine the value of $[A]$ by choosing the smallest minimum value of the above four minimization problems.

Using the above method, we can judge that the quartic polynomial $\mathcal{X}(\alpha_1, \alpha_2)$ is normal, and $[A] = 0.007286157241353 > 0$.

Since $n = 2$, we use Algorithm 3 to compute the global minimizer of $\mathcal{X}(\alpha_1, \alpha_2)$ in eight iterations.

We further test Algorithm 3 for $n = 2$, and test Algorithm 5 for $n = 3, 4$ and 6. All computer procedure codes are written in MATLAB language and are implemented on a Pentium III 550e PC. Some parameters are taken as

$$\rho = 3, \quad \varepsilon = \varepsilon_1 = \varepsilon_2 = 10^{-8}, \quad \sigma = 0.5, \quad \theta = 0.25, \quad \eta = 0.5, \quad M = 1.$$

In addition, in our numerical experiments, if $\|F(x^{(k)})\| \leq 10^{-8}$ at iteration k , we view $x^{(k)}$ as a critical point. For this time, we find a global descent direction at $x^{(k)}$ by Step 2, 3, or 4 (instead of Step 1) of Algorithms 3 and 5.

When $n = 2$, using Proposition 16 and the routine RANDPOLY in MAPLE, we randomly generate 10 normal quartic polynomials and find their global minimizers by Algorithm 3. The results show that the maximal iteration number is 10 and the minimal iteration number is 4. The computer time for each example is no more than one second. Table 1 lists the global minimal point (GMP), the corresponding minimal value of the objective function (MOF), and the iteration number (IN) of Algorithm 3 for each problem. We recorded the times Algorithm 3 visited Step 2–4 before it obtained a global minimizer for the first 10 examples in the appendix. We found that Algorithm 3 visited Step 2–4 three times before it obtained a global minimizer for test example Q21, and only once for the other 9 test examples.

In Tables 2 and 3, we give numerical results by implementing Algorithm 5 for the cases $n = 3$ and 4, respectively. Here, 15 normal quartic polynomials are generated randomly in a similar fashion to that for the case $n = 2$, of which ten minimization problems are for $n = 3$, and five are for $n = 4$. This time, the maximum IN is 21, the minimum is 6. The computer time for each example is approximately one second.

The termination criterion when $n \geq 3$ is Corollary 17 following Proposition 16. Hence, the obtained global minimizers satisfy Proposition 13.

When n further increases, the number of coefficients of a quartic polynomial increases rapidly. This poses a storage problem. Also, in practice, when n is large, the coefficient tensors may be sparse. Hence, at last in this section, we use Algorithm 5 to solve a class of special normal quartic polynomials with $n = 6$ and sparse coefficient tensors. These results further verify the effectiveness of the proposed algorithm.

TABLE 2
The numerical results when $n = 3$.

Question	GMP (x^*, y^*, z^*)	MOF	IN
Q31	(-1.2618991, 0.481777475, 0.511554542)	-112.374935	7
Q32	(0.45198184, 0.28181568, -0.04439019)	-22.11006871	6
Q33	(0.42542661, 1.21252312, -1.858553196)	-228.406952	8
Q34	(1.52134791, 1.95307499, -1.95758625)	-748.700592	21
Q35	(1.63015225, 0.79920839, -1.37311600)	-280.1038486	8
Q36	(7.715035687, -1.54753605, -2.52152795)	-17687.1075	12
Q37	(0.8565932583, -0.26891694, 0.162867234)	-70.1167646	8
Q38	(0.752808377, 0.287362024, -0.822919492)	-102.236381	6
Q39	(0.814835158, -0.308996855, -1.38411083)	-147.45786	8
Q310	(-0.693137416, 3.75111948, -11.5578088)	-31876.9903	10

TABLE 3
The numerical results when $n = 4$.

Question	GMP (x^*, y^*, z^*, u^*)	MOF	IN
Q41	(-2.16567286, -1.05504009, 1.91472660, 3.19936935)	-1395.23409	17
Q42	(-0.05721938, -1.34223106, 0.71604593, 1.29859919)	-222.071582	9
Q43	(-1.2292586, 0.63703738, 0.73368052, -0.74318163)	-117.921805	9
Q44	(-0.33345841, -0.20877798, 1.50120003, -0.19617979)	-122.44731	16
Q45	(-0.67299109, -0.10065256, 0.83060242, -0.01484504)	-201.373676	6

It is seen that the following polynomial is normal:

$$f(x) = \sum_{i=1}^6 a_i x_i^4 + x^T P x + p^T x,$$

where $a_i, i = 1, \dots, 6$, are six randomly generated positive numbers, P is a randomly generated 6×6 matrix, and p is a randomly generated six-dimensional vector. Furthermore, if we use the infinity norm, then we have $[A] = \min\{a_i\}$ (here A represents the fourth order coefficient tensor of the leading degree term of the polynomial).

We first compute the global minimizer of a simple example:

$$x^4 + y^4 + z^4 + u^4 + v^4 + w^4 + x^2 + y^2 - z^2 - u - v + w.$$

It is not very difficult to verify that

$$(x, y, z, u, v, w) = (0, 0, 0.70710678118655, 0.62996052494744, 0.62996052494744, -0.62996052494744)$$

is a global minimizer by directly solving the KKT equations. By implementing Algorithm 5, we get the above solution after five iterations.

Table 4 lists the numerical results when we randomly generate the coefficients.

All 29 randomly generated tested examples in this section are listed in the appendix.

9. Concluding remarks. In this paper, we introduced the concept of global descent directions. For a normal quartic polynomial, we give ways to find a global descent direction at a noncritical point, saddle point, or local maximizer in the general case and at a local minimizer when $n = 2$. For $n \geq 3$, we propose a constrained nonlinear equation approach to find a global descent direction at a local minimizer. We also give a formula at a critical point and a method at a noncritical point to find a

TABLE 4
The numerical results when $n = 6$.

Question	Q61	Q62	Q63	Q64
x^*	0.54521881	-0.65466417	-0.363974062	-1.350391459
y^*	-1.46441019	-1.86951600	-1.028303631	-1.483150332
z^*	-0.72060665	-0.368135071	0.563190492	-1.369006772
u^*	1.17814427	0.819086646	0.9486927097	-1.10594118
v^*	0.79406511	0.77562232	-0.707559149	1.54353024
w^*	-0.465794119	-0.53132279	0.470942714	2.33088412
MOF	28.94281730	-23.00564783	-16.27241853	-70.87818171
IN	25	23	21	23

one-dimensional global minimizer along a global descent direction. Based upon these, two global descent algorithms were proposed. For $n = 2$, the proposed algorithm can find a global minimizer of a normal quartic polynomial. For $n \geq 3$, the proposed algorithm can find an ϵ -global minimizer of the objective function when some global optimality condition holds.

10. Appendix. In this appendix, we list all 29 randomly generated tested examples:

$$\text{Q21: } 32x^4 + 54x^3y + 127x^2y^2 + 78xy^3 + 73y^4 + 75 - 2x^3 - 45x^2y - 74x^2 - 60xy^2 - xy - 31x - 43y^3 + 73y^2 - 82y;$$

$$\text{Q22: } 108x^4 + 142x^3y + 266x^2y^2 + 160xy^3 + 96y^4 - 4 - 50x^3 - 47x^2y + 75x^2 + 67xy^2 - 79xy + 53x + 63y^3 - 24y^2 - 29y;$$

$$\text{Q23: } 32x^4 + 32x^3y + 150x^2y^2 + 124xy^3 + 81y^4 + 21 + 68x^3 - 84x^2y + 80x^2 + 23xy^2 - 20xy - 7x + 4y^3 - 77y^2 + 40y;$$

$$\text{Q24: } 16x^4 + 12x^3y + 92x^2y^2 + 18xy^3 + 127y^4 - 49 + 22x^3 - 8x^2y + 53x^2 + 84xy^2 - 98xy - 5x + 20y^3 - 72y^2 + 10y;$$

$$\text{Q25: } 86x^4 + 140x^3y + 170x^2y^2 + 76xy^3 + 39y^4 - 15 - 21x^3 - 75x^2 + 93xy^2 + 23xy - 16x + 83y^3 - 17y^2 + 51y;$$

$$\text{Q26: } 42x^4 + 102x^3y + 128x^2y^2 + 76xy^3 + 24y^4 + 63 - 85x^3 - 55x^2y - 37x^2 - 35xy^2 + 97xy + 50x + 79y^3 + 56y^2 + 49y;$$

$$\text{Q27: } 57x^4 + 126x^3y + 222x^2y^2 + 172xy^3 + 119y^4 - 1 + x^3 - 47x^2y - 91x^2 - 47xy^2 - 61xy + 41x - 58y^3 - 90y^2 + 53y;$$

$$\text{Q28: } 121x^4 + 218x^3y + 277x^2y^2 + 168xy^3 + 76y^4 + 88 + 43x^3 - 66x^2y - 53x^2 - 61xy^2 - 23xy - 37x + 31y^3 - 34y^2 - 42y;$$

$$\text{Q29: } 74x^4 + 128x^3y + 182x^2y^2 + 126xy^3 + 52y^4 - 32 - 76x^3 - 65x^2y + 25x^2 + 28xy^2 - 61xy - 60x + 9y^3 + 29y^2 - 66y;$$

$$\text{Q210: } 80x^4 + 172x^3y + 260x^2y^2 + 176xy^3 + 80y^4 + 5 + 78x^3 + 39x^2y + 94x^2 + 68xy^2 - 17xy - 98x - 36y^3 + 40y^2 + 22y;$$

$$\begin{aligned} \text{Q31: } & 117x^4 + 86x^3y + 182x^3z + 185x^2y^2 + 241x^2z^2 + 204x^2yz + 44xy^3 + 172xy^2z + \\ & 114xyz^2 + 136xz^3 + 57y^4 + 84y^3z + 159y^2z^2 + 56yz^3 + 50z^4 + 88 + 66x^3 - 29x^2y - \\ & 91x^2z - 53x^2 - 19xy^2 - 47xyz + 68xy - 72xz^2 - 87xz + 79x + 43y^3 - 66y^2z - 53y^2 - \\ & 61yz^2 - 23yz - 37y + 31z^3 - 34z^2 - 42z; \end{aligned}$$

$$\begin{aligned} \text{Q32: } & 105x^4 + 96x^3y + 90x^3z + 223x^2y^2 + 209x^2z^2 + 148x^2yz + 102xy^3 + 168xy^2z + \\ & 176xyz^2 + 114xz^3 + 98y^4 + 136y^3z + 275y^2z^2 + 142yz^3 + 108z^4 + 5 - 76x^3 - 65x^2y + \\ & 25x^2z + 28x^2 - 61xy^2 - 60xyz + 9xy + 29xz^2 - 66xz - 32x + 78y^3 + 39y^2z + 94y^2 + \\ & 68yz^2 - 17yz - 98y - 36z^3 + 40z^2 + 22z; \end{aligned}$$

$$\begin{aligned} \text{Q33: } & 61x^4 + 42x^3y + 130x^3z + 104x^2y^2 + 225x^2z^2 + 114x^2yz + 44xy^3 + 114xy^2z + \\ & 180xyz^2 + 162xz^3 + 36y^4 + 48y^3z + 142y^2z^2 + 138yz^3 + 112z^4 + 81 - 88x^3 - 43x^2y - \\ & 73x^2z + 25x^2 + 4xy^2 - 59xyz + 62xy - 55xz^2 + 25xz + 9x + 40y^3 + 61y^2z + 40y^2 - \\ & 78yz^2 + 62yz + 11y + 88z^3 + z^2 + 30z; \end{aligned}$$

$$\begin{aligned} \text{Q34: } & 96x^4 + 38x^3y + 262x^3z + 169x^2y^2 + 414x^2z^2 + 202x^2yz + 30xy^3 + 256xy^2z + \\ & 238xyz^2 + 282xz^3 + 75y^4 + 138y^3z + 253y^2z^2 + 146yz^3 + 110z^4 - 73 - 5x^3 - 28x^2y + \\ & 4x^2z - 11x^2 + 10xy^2 + 57xyz - 82xy - 48xz^2 - 11xz + 38x - 7y^3 + 58y^2z - 94y^2 - \\ & 68yz^2 + 14yz - 35y - 14z^3 - 9z^2 - 51z; \end{aligned}$$

$$\begin{aligned} \text{Q35: } & 80x^4 + 130x^3y + 206x^3z + 248x^2y^2 + 302x^2z^2 + 258x^2yz + 130xy^3 + 356xy^2z + \\ & 258xyz^2 + 210xz^3 + 108y^4 + 140y^3z + 216y^2z^2 + 94yz^3 + 82z^4 + 45 - 73x^3 - 91x^2y + \\ & x^2z + 5x^2 - 86xy^2 + 43xyz - 4xy - 50xz^2 + 50xz + 67x - 39y^3 + 8y^2z - 49y^2 + 11yz^2 + \\ & 93yz - 14y - 99z^3 - 67z^2 + 68z; \end{aligned}$$

$$\begin{aligned} \text{Q36: } & 22x^4 + 46x^3y + 66x^3z + 151x^2y^2 + 161x^2z^2 + 110x^2yz + 126xy^3 + 148xy^2z + \\ & 180xyz^2 + 126xz^3 + 90y^4 + 84y^3z + 187y^2z^2 + 70yz^3 + 87z^4 + 21 - 84x^3 + 46x^2y + \\ & 59x^2z - 56x^2 - 83xy^2 - 91xyz + 92xy - 93xz^2 + 91xz - 54x + 10y^3 - 77y^2z - 63y^2 - \\ & 90yz^2 + 61yz - 3y - 82z^3 + 16z^2 - 40z; \end{aligned}$$

$$\begin{aligned} \text{Q37: } & 104x^4 + 88x^3y + 128x^3z + 136x^2y^2 + 239x^2z^2 + 208x^2yz + 36xy^3 + 120xy^2z + \\ & 152xyz^2 + 106xz^3 + 26y^4 + 56y^3z + 146y^2z^2 + 124yz^3 + 102z^4 - 17 - 79x^3 - 27x^2y + \\ & 32x^2z - 24x^2 - 46xy^2 + 12xyz + 81xy + 63xz^2 - 85xz - 36x - 35y^3 + 11y^2z + 90y^2 + \\ & 31yz^2 + 47yz - 50y - 54z^3 + 71z^2 - 71z; \end{aligned}$$

$$\begin{aligned} \text{Q38: } & 76x^4 + 172x^3y + 176x^3z + 285x^2y^2 + 247x^2z^2 + 360x^2yz + 204xy^3 + 342xy^2z + \\ & 420xyz^2 + 236xz^3 + 93y^4 + 182y^3z + 293y^2z^2 + 182yz^3 + 126z^4 + 6 + 76x^3 - 57x^2y - \\ & 80x^2z - 92x^2 + 81xy^2 + 77xyz - 87xy + 50xz^2 + 74xz - 60x + 19y^3 - 68y^2z + 78y^2 + \\ & 34yz^2 + 66yz - 53y + 59z^3 + 28z^2 + 38z; \end{aligned}$$

$$\begin{aligned} \text{Q39: } & 111x^4 + 126x^3y + 126x^3z + 256x^2y^2 + 221x^2z^2 + 162x^2yz + 122xy^3 + 214xy^2z + \\ & 186xyz^2 + 138xz^3 + 69y^4 + 78y^3z + 166y^2z^2 + 82yz^3 + 89z^4 - 15 - 50x^3 - 47x^2y + \\ & 75x^2z + 67x^2 - 79xy^2 + 53xyz + 63xy - 24xz^2 - 29xz - 4x - 21y^3 - 75y^2z + 93yz^2 + \\ & 23yz - 16y + 83z^3 - 17z^2 + 51z; \end{aligned}$$

$$\begin{aligned} \text{Q310: } & 59x^4 + 136x^3y + 56x^3z + 205x^2y^2 + 84x^2z^2 + 120x^2yz + 144xy^3 + 130xy^2z + \\ & 88xyz^2 + 14xz^3 + 80y^4 + 88y^3z + 91y^2z^2 + 38yz^3 + 11z^4 - 37 - 95x^3 - 47x^2y + 51x^2z + \\ & 47x^2 + 5xy^2 + 33xyz + 9xy + 26xz^2 - 55xz + 37x - 94y^3 - 65y^2z + 90y^2 - 38yz^2 - \\ & 46yz + 28y + 88z^3 + 64z^2 - 22z; \end{aligned}$$

$$\begin{aligned} \text{Q41: } & 96y^2zu + 380xuz^2 + 338xzu^2 + 476x^2zu + 408xyu^2 + 388xyz^2 + 456xy^2u + \\ & 470xy^2z + 464x^2yu + 420x^2yz + 208yuz^2 + 138yzu^2 + 240x^3y + 408x^3z + 444x^3u + \\ & 430x^2y^2 + 554x^2z^2 + 522x^2u^2 + 192xy^3 + 362xz^3 + 326xu^3 + 144y^3z + 192y^3u + 352y^2z^2 + \\ & 341y^2u^2 + 156yz^3 + 184yu^3 + 297z^2u^2 + 216x^4 + 143y^4 + 165z^4 + 132u^4 + 408xzyu - \\ & 61 - 93x - 50y - 62y^2u + 63xz^2 + 50xy^2 + 45xu^2 + 92y^3 + 43y^2z + 66yz^2 + z - 47u + \\ & 97x^2 + 49xy - 59xz - 8xu + 77y^2 - 5yz - 61yu + 31z^2 - 91u^2 + 79xyz + 56xyu + \\ & 57xzu + 54yzu - 55x^2y - 37x^2z - 35x^2u - 62zu + 99yu^2 - 18z^2u - 26zu^2 - 85x^3 - \\ & 12z^3 - 47u^3; \end{aligned}$$

$$\begin{aligned} \text{Q42: } & 760xyzu + 832x^2yz + 600x^2yu + 454x^2zu + 722xy^2z + 578xy^2u + 720xyz^2 + \\ & 576xyu^2 + 422xz^2u + 476xz^2u + 614y^2zu + 408x^3y + 442x^3z + 306x^3u + 604x^2y^2 + \\ & 619x^2z^2 + 514x^2u^2 + 344xy^3 + 354xz^3 + 252xu^3 + 456y^3z + 328y^3u + 617y^2z^2 + 486y^2u^2 + \\ & 436yz^3 + 288yu^3 + 616yzu^2 + 289x^4 + 195y^4 + 536yz^2u + 182z^4 + 160u^4 + 190z^3u + \\ & 396z^2u^2 + 188zu^3 - 61 - 93x - 50y + z - 47u + 97x^2 + 49xy - 59xz - 8xu + 77y^2 - \\ & 5yz - 61yu + 31z^2 - 62zu - 91u^2 + 79xyz + 56xyu + 57xzu + 54yzu - 55x^2y - 37x^2z - \\ & 35x^2u + 50xy^2 + 63xz^2 + 45xu^2 + 43y^2z - 62y^2u + 66yz^2 + 99yu^2 - 18z^2u - 26zu^2 - \\ & 85x^3 + 92y^3 - 12z^3 - 47u^3; \end{aligned}$$

$$\begin{aligned} \text{Q43: } & 380xyzu + 308x^2yz + 336x^2yu + 210x^2zu + 358xy^2z + 302xy^2u + 324xyz^2 + \\ & 296xyu^2 + 182xz^2u + 310xzu^2 + 336y^2zu + 196x^3y + 192x^3z + 86x^3u + 429x^2y^2 + 284x^2z^2 + \\ & 327x^2u^2 + 260xy^3 + 144xz^3 + 114xu^3 + 162y^3z + 362y^3u + 367y^2z^2 + 564y^2u^2 + 132yz^3 + \\ & 294yu^3 + 282yzu^2 + 130x^4 + 272y^4 + 316yz^2u + 90z^4 + 180u^4 + 138z^3u + 345z^2u^2 + \\ & 192zu^3 + 79 + 85x + 72y - 47z - 87u - x^2 + 23xy - 50xz - 53xu + 72y^2 - 86yz + 80yu - 91z^2 - \\ & 19zu - 72u^2 + 83xyz - 86xyu + 19xzu - 85yzu - 58x^2y - 90x^2z + 53x^2u + 94xy^2 - 84xz^2 + \\ & 88xu^2 + 78y^2z + 17y^2u - 99yz^2 + 30yu^2 - 29z^2u - 53zu^2 + 41x^3 + 49y^3 + 66z^3 + 68u^3. \end{aligned}$$

$$\begin{aligned} \text{Q44: } & 308xyz^2 + 518xy^2u + 400xy^2z + 556x^2zu + 548x^2yu + 406x^2yz + 322y^2zu + \\ & 632xzu^2 + 450xz^2u + 468xyu^2 + 324yzu^2 + 236yz^2u + 428xyzu + 316x^3y + 248x^3z + \\ & 520x^3u + 481x^2y^2 + 343x^2z^2 + 735x^2u^2 + 288xy^3 + 144xz^3 + 530xu^3 + 210y^3z + 108y^3u + \\ & 303y^2z^2 + 445y^2u^2 + 144yz^3 + 128yu^3 + 162z^3u + 344z^2u^2 + 298zu^3 + 240x^4 + 180y^4 + \\ & 81z^4 + 255u^4 + 4 - 61x - 17y - 36z^2u + 78yzu - 76xzu - 34xyu + 31xyz + 25u + \\ & 94yu^2 + 9y^2z + 29y^2u - 32yz^2 - 53x^2z + 22zu^2 - 37xy^2 - 88z - 23x^2 - 42xy - 65xz + \\ & 28xu - 66y^2 + 39yz + 68yu + 40z^2 + 5zu - 73u^2 - 66x^2y - 61x^2u + 88xz^2 + 25xu^2 + \\ & 43x^3 - 60y^3 - 98z^3 - 43u^3; \end{aligned}$$

$$\begin{aligned} \text{Q45: } & 440xyz^2 + 610xy^2u + 392xy^2z + 654x^2zu + 526x^2yu + 370x^2yz + 610y^2zu + \\ & 732xzu^2 + 744xz^2u + 628xyu^2 + 628yzu^2 + 596yz^2u + 720xyzu + 190x^3y + 250x^3z + \\ & 354x^3u + 399x^2y^2 + 468x^2z^2 + 598x^2u^2 + 242xy^3 + 300xz^3 + 456xu^3 + 242y^3z + 332y^3u + \\ & 474y^2z^2 + 562y^2u^2 + 260yz^3 + 364yu^3 + 444z^3u + 668z^2u^2 + 456zu^3 + 144x^4 + 195y^4 + \\ & 224z^4 + 224u^4 - 73 + 81x + 38y + 58z^2u + 57yzu + 11xzu + 40xyu + 61xyz - 51u - 48yu^2 - \\ & 28y^2z + 4y^2u + 10yz^2 - 55x^2z - 68zu^2 + 40xy^2 - 35z + 9x^2 - 78xy + 88xz + 30xu - \\ & 11y^2 - 82yz - 11yu - 94z^2 + 14zu - 9u^2 + 62x^2y + 25x^2u + 62xz^2 + xu^2 - 59x^3 - 5y^3 - \\ & 7z^3 - 14u^3; \end{aligned}$$

$$\begin{aligned} \text{Q61: } & 9x^4 + 2y^4 + 6z^4 + 4u^4 + 8v^4 + 7w^4 + 2x + 6y + 5z + 2w + 4x^2 + 8xy + 18xz + \\ & 6xu + 8xv + 2xw + 3y^2 + 14yz + 18yu + 18yv + 4yw + 4z^2 + 14zu + 12zv + 12zw + \\ & 4u^2 + 4uv + 12uw + 8v^2 + 6vw + 5w^2; \end{aligned}$$

$$\begin{aligned} \text{Q62: } & 4x^4 + y^4 + 8z^4 + 4u^4 + 6v^4 + 7w^4 + 4x^2 + 6xu + 6xw + 12yu + 12yv + 5z^2 + \\ & 6zv + 12zw + 4u^2 + 8uv + 6uw + 4v^2 + 10vw + 2w^2 + 8x + 7y + 7z + 8u + 6v + 2w; \end{aligned}$$

Q63: $9x^4 + 7y^4 + z^4 + 4u^4 + 9v^4 + 9w^4 + 8x^2 + 2xz + 6xu + 18xv + 18xw + 18yz + 10yu + 4yv + 12yw + 4z^2 + 2zu + 2zv + 16zw + 16uv + 2v^2 + 2vw + 8w^2 + 5x + 8y + 6z + 9u + 9v$;

Q64: $x^4 + 2y^4 + z^4 + 6u^4 + 2v^4 + w^4 + 8x + 7y + 6z + 7v + 4u + 6w + 4x^2 + 2xy + 8xz + 4xu + 8xv + 8xw + y^2 + 8yz + 2yv + 14yw + 4z^2 + 12zu + 12zv + 14zw + 6u^2 + 14uv + 18uw + 3v^2 + 3w^2$.

Acknowledgments. The authors thank Professor Sheng-Li Tan and Dr. Guohui Zhao for sharing their knowledge of algebraic varieties, and thank a referee for helpful comments.

REFERENCES

- [1] S. BELLAVIA, M. MACCONI, AND B. MORINI, *An affine scaling trust-region approach to bound-constrained nonlinear systems*, Appl. Numer. Math., 44 (2003), pp. 257–280.
- [2] A. CANTONI AND M. C. BLASIKIEWIEZ, *Derivative constrained broad band antenna array processors revisited*, in Proceedings of IASTED, 2nd International Symposium on Signal Processing and Its Applications, Gold Coast, Queensland, August 1990, pp. 319–326.
- [3] D. COX, J. LITTLE, AND D. O’SHEA, *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*, Springer-Verlag, New York, 1992.
- [4] A. DURFEE, N. KRONENFELD, H. MUNSON, J. ROY, AND I. WESTBY, *Counting critical points of real polynomials in two variables*, Amer. Math. Monthly, 100 (1993), pp. 255–271.
- [5] M. H. ER AND A. CANTONI, *Techniques in robust broad band beamforming*, in Control and Dynamic Systems, C. T. Leondes, ed., Academic Press, San Diego, 1992.
- [6] C. A. FLOUDAS, *Deterministic Global Optimization: Theory, Methods and Applications*, Kluwer, Dordrecht, The Netherlands, 2000.
- [7] A. FRIEDLANDER AND J. M. MARTÍNEZ, *On the maximization of a concave quadratic function with box constraints*, SIAM J. Optim., 4 (1994), pp. 177–192.
- [8] A. FRIEDLANDER, J. M. MARTÍNEZ, AND S. A. SANTOS, *A new trust region algorithm for bound constrained minimization*, Appl. Math. Optim., 30 (1994), pp. 235–266.
- [9] S. A. GABRIEL AND J. S. PANG, *A trust region method for constrained nonsmooth equations*, in Large Scale Optimization - State of the Art, W. W. Hager, D. W. Hearn, and P. M. Pardalos, eds., Kluwer, Dordrecht, The Netherlands, 1994, pp. 155–181.
- [10] C. G. GIBSON, *Elementary Geometry of Algebraic Curves*, Cambridge University Press, Cambridge, UK, 1998.
- [11] C. KANZOW, *An active-set type Newton method for constrained nonlinear equations*, in Complementarity: Applications, Algorithms and Extensions, M. C. Ferris, O. L. Mangasarian, and J.-S. Pang, eds., Kluwer, Dordrecht, The Netherlands, 2001, pp. 179–200.
- [12] M. KOJIMA, S. KIM, AND H. WAKI, *A general framework for convex relaxation of polynomial optimization problems over cones*, in Proceedings of the Fourteenth RAMP Symposium, A. Tamura and H. Ito, eds., RAMP, Kyoto, 2002, pp. 115–132.
- [13] D. N. KOZAKEVICH, J. M. MARTÍNEZ, AND S. A. SANTOS, *Solving nonlinear systems of equations with simple constraints*, Math. Appl. Comput., 16 (1997), pp. 215–235.
- [14] J. B. LASSERRE, *Global optimization with polynomials and the problem of moments*, SIAM J. Optim., 11 (2001), pp. 796–817.
- [15] J. L. NAZARETH AND L. QI, *Globalization of Newton’s method for solving nonlinear equations*, J. Numer. Algebra Appl., 3 (1996), pp. 239–249.
- [16] Y. NESTEROV, *Squared functional systems and optimization problems*, in High Performance Optimization, H. Frenk, K. Roos, T. Terlaky, and S. Zhang, eds., Kluwer, Dordrecht, The Netherlands, 2000, pp. 405–440.
- [17] L. QI, *Extrema of a real polynomial*, J. Global Optim., to appear.
- [18] L. QI AND K. L. TEO, *Multivariate polynomial minimization and its application in signal processing*, J. Global Optim., 26 (2003), pp. 419–433.
- [19] L. QI, X. J. TONG, AND D. H. LI, *An active-set projected trust region algorithm for box constrained nonsmooth equations*, J. Optim. Theory Appl., 120 (2004), pp. 601–625.
- [20] N. Z. SHOR, *Nondifferentiable Optimization and Polynomial Problems*, Kluwer, Dordrecht, The Netherlands, 1998.
- [21] E. SHUSTIN, *Critical points of real polynomials, subdivisions of Newton polyhedra and topology of real algebraic hypersurfaces*, Amer. Math. Soc. Transl., 173 (1996), pp. 203–223.

- [22] I. THNG, A. CANTONI, AND Y. H. LEUNG, *Derivative constrained optimum broad band antenna arrays*, IEEE Trans. Signal Process., 41 (1993), pp. 2376–2388.
- [23] I. THNG, A. CANTONI, AND Y. H. LEUNG, *Analytical solutions to the optimization of a quadratic cost function subject to linear and quadratic equality constraints*, Appl. Math. Optim., 34 (1996), pp. 161–182.
- [24] M. ULBRICH, *Nonmonotone trust-region methods for bound-constrained semismooth equations with applications to nonlinear mixed complementarity problem*, SIAM J. Optim., 11 (2001), pp. 889–917.
- [25] A. Y. UTESHEV AND T. M. CHERKASOV, *The search for the maximum of a polynomial*, J. Symbolic Comput., 25 (1998), pp. 587–618.
- [26] H. WHITNEY, *Elementary structure of real algebraic varieties*, Ann. Math., 66 (1957), pp. 545–556.

A GLOBAL ALGORITHM FOR NONLINEAR SEMIDEFINITE PROGRAMMING*

RAFAEL CORREA[†] AND HECTOR RAMIREZ C.[‡]

Abstract. In this paper we propose a global algorithm for solving nonlinear semidefinite programming problems. This algorithm, inspired by the classic SQP (sequentially quadratic programming) method, modifies the S-SDP (sequentially semidefinite programming) local method by using a nondifferentiable merit function combined with a line search strategy.

Key words. nonlinear semidefinite programming, sequentially programming, global convergence

AMS subject classification. 90C22

DOI. 10.1137/S1052623402417298

1. Introduction. We consider the nonlinear programming problem

$$\begin{aligned} & \text{minimize} && f(x) \\ \text{(P)} & \text{subject to} && \mathcal{A}(x) \preceq 0, \\ & && h(x) = 0, \end{aligned}$$

where $x \in \mathbb{R}^n$, $\mathcal{A}$ is a smooth function whose values are symmetric matrices, $\preceq$ denotes the negative semidefinite order (that is, $A \preceq B$ if and only if $A - B$ is a negative semidefinite matrix), h is a smooth vector function with values in $\mathbb{R}^p$, and f is the smooth objective function. The smoothness of all these functions is specified at each statement.

This problem becomes interesting when the linear matrix formulation [23]

$$\begin{aligned} \text{(LMI)} & \text{minimize} && f(x) = c^T x \\ & \text{subject to} && \mathcal{A}(x) = A_0 + \sum_{i=1}^m x_i A_i \preceq 0 \end{aligned}$$

does not give a satisfactory model for certain problems, particularly those from control theory [1, 5, 8, 9].

This paper is organized as follows. In section 2 the optimality and constraint qualification conditions for problem (P) are presented. The results contained in this section are adaptations of known results (see, for example, [17, 21]). Here only the optimality conditions that are useful in our context are discussed. Other conditions can be found in [2, 6]. In section 3 we demonstrate some exactness results associated with the Lagrangian, the augmented Lagrangian and the Han penalty function. The first one is well known and we review it to make our exposition self-contained. The exactness of the augmented Lagrangian and the Han penalty function are extensions of the corresponding classical mathematical programming results [3, 15]. In section 4 we propose a global S-SDP (sequentially semidefinite programming) algorithm and prove

*Received by the editors November 5, 2002; accepted for publication (in revised form) September 6, 2003; published electronically December 9, 2004. This research was supported by a FONDAP-CONICYT grant.

<http://www.siam.org/journals/siopt/15-1/41729.html>

[†]Centre for Mathematical Modelling, UMR 2071, University of Chile-CNRS, Casilla 170-3, Santiago 3, Chile (rcorrea@dim.uchile.cl).

[‡]Centre for Mathematical Modelling, UMR 2071, University of Chile-CNRS and Department of Mathematical Engineering, University of Chile, Casilla 170-3, Santiago 3, Chile (hramirez@dim.uchile.cl).

its convergence. The convergence of a local S-SDP algorithm has been proved by Fares, Noll, and Apkarian [9]. Other works concerning the global convergence of methods for solving optimization programs with nonlinear matrix inequalities constraints are [8, 16].

1.1. Notations. Throughout we denote by S^m the set of all symmetric matrices of dimension m , by S_+^m the set of all symmetric positive semidefinite matrices, and by S_{++}^m the set of all symmetric positive definite matrices. The sets S_-^m and S_{--}^m are defined similarly. For all these sets of matrices we use the trace product $\langle A, B \rangle = \text{Tr}(AB)$ and the Frobenius norm $\|A\|_{Fr} = \sqrt{\text{Tr}(A^2)}$. For a given matrix A , $\lambda_j(A)$ denotes its j th eigenvalue in nonincreasing order and A_+ denotes the matrix defined by

$$(1.1) \quad A_+ := P \text{diag}((\lambda_1)_+, \dots, (\lambda_m)_+) P^T,$$

where $(\lambda)_+ = \max\{0, \lambda\}$ and P is the matrix in the spectral decomposition $A = P \text{diag}(\lambda_1, \dots, \lambda_m) P^T$. It is easy to see that A_+ is the orthogonal projection of A on S_+^m .

Given a matrix-valued function $\mathcal{A}(\cdot)$ we will use the notation

$$D\mathcal{A}(x_*) = \left(\frac{\partial \mathcal{A}(x_*)}{\partial x_i} \right)_{i=1}^n = \left(\frac{\partial \mathcal{A}(x_*)}{\partial x_1}, \dots, \frac{\partial \mathcal{A}(x_*)}{\partial x_n} \right)^T$$

for its differential operator evaluated at x_* . This notation comes from the fact that

$$(1.2) \quad D\mathcal{A}(x_*)y = \sum_{i=1}^n y_i \frac{\partial \mathcal{A}(x_*)}{\partial x_i} \quad \forall y \in \mathbb{R}^n.$$

Finally, if we define the linear operator V from $\mathbb{R}^n$ to S^m by $Vy = \sum_{i=1}^n y_i V_i$, where $V_i \in S^m$ for all $i \in \{1, \dots, n\}$, we have for the adjoint operator V^* the formula

$$(1.3) \quad V^*Z = (\text{Tr}(V_1 Z), \dots, \text{Tr}(V_n Z))^T \quad \forall Z \in S^m.$$

2. Optimality conditions review. In this section we state the first- and second-order optimality conditions for (P) and discuss their implications. To this end, we consider the Lagrangian $L : \mathbb{R}^n \times S^m \times \mathbb{R}^p \rightarrow \mathbb{R}$ of problem (P) defined by

$$(2.1) \quad L(x, Z, \lambda) = f(x) + \text{Tr}(Z\mathcal{A}(x)) + \lambda^T h(x).$$

2.1. First order optimality conditions. The Karush–Kuhn–Tucker necessary optimality conditions for a feasible point x_* of (P) are given by the existence of $Z_* \in S^m$ and $\lambda_* = (\lambda_{*1}, \dots, \lambda_{*p})^T \in \mathbb{R}^p$ such that

$$(KKT) \quad \begin{aligned} \nabla f(x_*) + D\mathcal{A}(x_*)^* Z_* + \sum_{j=1}^p \lambda_{*j} \nabla h_j(x_*) &= 0, \\ \text{Tr}(Z_* \mathcal{A}(x_*)) &= 0, \\ Z_* &\succeq 0. \end{aligned}$$

The pair (Z_*, λ_*) is called the (KKT)-multiplier associated with x_* . The complementarity condition $\text{Tr}(Z_* \mathcal{A}(x_*)) = 0$ has the following two useful equivalent forms:

$$(2.2) \quad \lambda_j(Z_*) \lambda_j(\mathcal{A}(x_*)) = 0 \quad \forall j \in \{1, \dots, m\}$$

and

$$(2.3) \quad Z_* \mathcal{A}(x_*) = 0.$$

Both forms are easily obtained from the *Von Neumann–Theobald inequality*:

$$(2.4) \quad \text{Tr}(AB) \leq \sum_{j=1}^m \lambda_j(A) \lambda_j(B),$$

where the equality holds if and only if there is a matrix P such that $P^{-1}AP$ and $P^{-1}BP$ are diagonal [22, 24].

Condition (2.2) allows us to define the *strict complementarity condition* in (KKT) as follows:

$$(2.5) \quad \lambda_j(Z_*) = 0 \quad \text{if and only if} \quad \lambda_j(\mathcal{A}(x_*)) < 0 \quad \forall j \in \{1, \dots, m\}.$$

As is well known, the (KKT) conditions are not a consequence of the optimality of x_* , and to ensure this consequence, we must assume an extra condition. In this paper, we will use *Robinson's constraint qualification condition* [18]

$$(2.6) \quad 0 \in \text{int} \left\{ \begin{pmatrix} \mathcal{A}(x_*) \\ h(x_*) \end{pmatrix} + \begin{pmatrix} D\mathcal{A}(x_*) \\ \nabla h(x_*) \end{pmatrix} \mathbb{R}^n - \begin{pmatrix} S^m \\ \{0\} \end{pmatrix} \right\},$$

where $\text{int } C$ denotes the topological interior of the set C . A direct consequence of [12, Chapter 3, Prop. 2.1.12] is the equivalence between condition (2.6) and the *Mangasarian–Fromovitz constraint qualification condition*

$$(2.7a) \quad \{\nabla h_j(x_*)\} \text{ is linearly independent, and}$$

$$(2.7b) \quad \exists \bar{d} \in \mathbb{R}^n \text{ s. t. } \begin{cases} \nabla h(x_*) \bar{d} = 0 \\ \text{and } \mathcal{A}_*(\bar{d}) \prec 0, \end{cases}$$

where $\mathcal{A}_* : \mathbb{R}^n \rightarrow S^m$ is the linear affine function defined by $\mathcal{A}_*(y) := \mathcal{A}(x_*) + \sum_{i=1}^n y_i \frac{\partial \mathcal{A}(x_*)}{\partial x_i}$. It can be shown that under (2.6) the set of (KKT) Lagrange multipliers is nonempty and also bounded [14].

We will also consider the *transversality condition* which asks that the function $\psi : \mathbb{R}^n \rightarrow \mathbb{R}^p \times S_r$ defined by

$$(2.8) \quad \psi(d) := \left((\nabla h(x_*)d)^T, N^T \sum_{i=1}^n d_i \frac{\partial \mathcal{A}(x_*)}{\partial x_i} N \right)^T$$

be surjective, where

$$(2.9) \quad N = [v_1 \dots v_r]$$

is the matrix whose columns v_i are an orthonormal basis of $\text{Ker } \mathcal{A}(x_*)$. We set $N = 0$ if $\text{Ker } \mathcal{A}(x_*) = \{0\}$. This condition has originally been defined in the context of smooth manifolds [10] and implies the Robinson's constraint qualification condition (2.7); moreover, (2.8) guarantees the uniqueness of the (KKT)-multiplier. Unfortunately, this condition can be very strong, because it forces $n \geq p + r(r+1)/2$, where $r = \dim[\text{Ker } \mathcal{A}(x_*)]$.

It is clear that the transversality condition (2.8) cannot hold when the matrix $\mathcal{A}(x_*)$ has a diagonal block structure. Indeed, in this case the multiplier Z_* is not

unique, and therefore the transversality condition does not hold. This difficulty can be easily avoided if we assume the transversality condition for each block of $\mathcal{A}(x_*)$. For example, if $\mathcal{A}(x_*)$ has two diagonal block structures with sizes m_1 and m_2 , then the mapping ψ should be considered into the cross-product space $S^{m_1} \times S^{m_2}$ rather than the larger space $S^{m_1+m_2}$. For simplicity of notation we only consider the case where $\mathcal{A}(x_*)$ is a one-block matrix. More details about the transversality condition in the semidefinite programming context can be seen in [21] and the references within.

2.2. Second-order sufficient conditions. In this section we introduce only the second-order sufficient conditions that will be used in this paper as well as results that involve the transversality condition (2.8). We assume that f , h , and $\mathcal{A}$ are twice differentiable at x_* .

Given a set $B \subseteq \mathbb{R}^m$ we define

$$(2.10) \quad S_-^m(B) := \{M \in S^m : w^T M w \leq 0 \quad \forall w \in B\}.$$

PROPOSITION 1. *A sufficient condition to obtain the isolated optimality of x_* for problem (P) is the existence of $(Z_*, \lambda_*) \in S^m \times \mathbb{R}^p$ such that (x_*, Z_*, λ_*) satisfies (KKT) and*

$$(2.11) \quad d^T \nabla_{xx}^2 L(x_*, Z_*, \lambda_*) d > 0$$

for all nonzero vectors $d \in C(x_*)$, where

$$(2.12) \quad C(x_*) = \left\{ d \in \mathbb{R}^n : \sum_{i=1}^n d_i \frac{\partial \mathcal{A}(x_*)}{\partial x_i} \in S_-^m(\text{Ker } \mathcal{A}(x_*)), \nabla h(x_*) d = 0 \text{ and } \nabla f(x_*)^T d = 0 \right\}$$

is a cone of critical directions for problem (P) at the point x_* .

Proof. See, for example, [19, Thm. 2.2] and also note that $T_{S_-^m}(\mathcal{A}(x_*)) = S_-^m(\text{Ker } \mathcal{A}(x_*))$. $\square$

Remark. Condition (2.11) can be far from necessary. For instance, in the problem (LMI), mentioned in the introduction, we always have $\nabla_{xx}^2 L = 0$; thus, if $C(x_*) \neq \{0\}$, condition (2.11) never holds. This is because condition (2.11) does not consider the geometry of S_-^m . This kind of problem was the motivation for works such as [2, 6, 21] in the 1990s. We will just consider the nonlinear problem (P), where the algorithm S-SDP makes sense.

Let us now define a larger cone of critical directions $C'(x_*, Z_*)$, which considers the (KKT)-multiplier Z_* associated with the matrix inequality $\mathcal{A}(x) \preceq 0$, as follows:

$$(2.13) \quad C'(x_*, Z_*) := \left\{ d \in \mathbb{R}^n : \text{Im } Z_* \subseteq \text{Ker } Pr \sum_{i=1}^n d_i \frac{\partial \mathcal{A}(x_*)}{\partial x_i} \text{ and } \nabla h(x_*) d = 0 \right\},$$

where Pr is the orthogonal projection in $\mathbb{R}^m$ over $\text{Ker } \mathcal{A}(x_*)$. Note that $Pr = NN^T$ with N defined in (2.9).

The next proposition relates both cones of critical directions and the function ψ used in the transversality condition.

PROPOSITION 2. *Let x_* be a solution of (P) and (Z_*, λ_*) be a (KKT)-multiplier. Let us also consider the function ψ , defined in (2.8), and the cones of critical directions defined above. Then*

$$(2.14) \quad \text{Ker } \psi \subseteq C(x_*) \subseteq C'(x_*, Z_*),$$

with equality when the strict complementarity condition (2.5) holds.

Proof. First, note that we can write (2.3) in the equivalent form

$$(2.15) \quad Z_* = N\phi_*N^T,$$

with $\phi_* \in S_r^+$ and $r = \dim \text{Ker } \mathcal{A}(x_*)$. Then, to prove the first inclusion in (2.14), it is sufficient to show that $\nabla f(x_*)^T d = 0$. This comes from the first equation in (KKT) and the equality

$$(2.16) \quad \text{Tr} \left(Z_* \sum_{i=1}^n d_i \frac{\partial \mathcal{A}(x_*)}{\partial x_i} \right) = \text{Tr} \left(\phi_* N^T \sum_{i=1}^n d_i \frac{\partial \mathcal{A}(x_*)}{\partial x_i} N \right).$$

For the second inclusion, if $d \in C(x_*)$ then $\nabla f(x_*)^T d = 0$ and $\nabla h(x_*)d = 0$, and we obtain from (2.16) and the first equation in (KKT) that $\text{Tr}(\phi_* N^T \sum_{i=1}^n d_i \frac{\partial \mathcal{A}(x_*)}{\partial x_i} N) = 0$. Since $\sum_{i=1}^n d_i \frac{\partial \mathcal{A}(x_*)}{\partial x_i} \in S_-^m(\text{Ker } \mathcal{A}(x_*))$, we see that $N^T \sum_{i=1}^n d_i \frac{\partial \mathcal{A}(x_*)}{\partial x_i} N \in S_r^-$, and using (2.4), we deduce from the last equality that

$$(2.17) \quad N^T \sum_{i=1}^n d_i \frac{\partial \mathcal{A}(x_*)}{\partial x_i} N \phi_* = 0,$$

which is equivalent to

$$(2.18) \quad Pr \sum_{i=1}^n d_i \frac{\partial \mathcal{A}(x_*)}{\partial x_i} Z_* = N N^T \sum_{i=1}^n d_i \frac{\partial \mathcal{A}(x_*)}{\partial x_i} N \phi_* N^T = 0,$$

and we conclude that $d \in C'(x_*, Z_*)$.

If, in addition, we assume the strict complementarity condition (2.5), we have that ϕ_* is nonsingular, and from the equivalence between (2.18) and (2.17) we deduce the converse inclusion $C'(x_*, Z_*) \subseteq \text{Ker } \psi$. $\square$

A direct consequence of propositions 1 and 2 is the following stronger second-order sufficient condition for optimality.

PROPOSITION 3. *Under the hypotheses of Proposition 1, where the critical cone $C(x_*)$ is replaced by $C'(x_*, Z_*)$, the point x_* is an isolated local minimum of (P).*

3. Exact penalty functions. A pair (x_*, y_*) in the product set $X \times Y$ is said to be a *saddle-point* of the function $\varphi : X \times Y \rightarrow \mathbb{R}$ on $X \times Y$ if

$$\varphi(x_*, y) \leq \varphi(x_*, y_*) \leq \varphi(x, y_*) \quad \forall x \in X, \forall y \in Y.$$

We say that a function $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}$ is an *exact penalty function* for a local minimum x_* of (P) if x_* is a local minimum of Φ too.

In this section we study different penalty functions associated with problem (P) and state necessary and sufficient conditions for exactness. A general approach for the study of exact penalty functions can be found in [4, Sect. 3.4.2].

3.1. The Lagrangian function in the convex case. Let us consider the particular case of (P) when $h : \mathbb{R}^n \rightarrow \mathbb{R}^p$ is an affine function $h(x) = h_0 + Hx$, with $H \in \mathbb{R}^{p \times n}$ and $h_0 \in \mathbb{R}^p$, $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex, and $\mathcal{A}(\cdot)$ is convex in the sense of the semidefinite order, that is,

$$\mathcal{A}(tx + (1-t)y) \preceq t\mathcal{A}(x) + (1-t)\mathcal{A}(y) \quad \forall t \in [0, 1], \forall x, y \in \mathbb{R}^n.$$

With these assumptions, (P) will be denoted by (P_C) (*convex problem*). For (x_*, Z_*, λ_*) satisfying (KKT), it can be shown that the function $L(\cdot, Z_*, \lambda_*)$, defined in (2.1), is an exact penalty function for (P_C) . This is an immediate consequence of the fact that (x_*, Z_*, λ_*) is a saddle-point of the Lagrangian function on $\mathbb{R}^n \times S_+^m \times \mathbb{R}^p$ (see, for example, [25, Thm. 4.1.3]). However, it is known that the Lagrangian is no longer an exact penalty function in the nonconvex case, which is the reason other penalty functions are introduced to obtain exactness results for our general problem (P).

3.2. The augmented Lagrangian. We define the augmented Lagrangian function L_σ associated with problem (P) as

$$(3.1) \quad \begin{aligned} L_\sigma(x, Z, \lambda) = & f(x) + \lambda^T h(x) + \frac{\sigma}{2} \|h(x)\|^2 \\ & + \text{Tr}(Z[\mathcal{A}(x) + (-\sigma^{-1}Z - \mathcal{A}(x))_+]) + \frac{\sigma}{2} \|\mathcal{A}(x) + (-\sigma^{-1}Z - \mathcal{A}(x))_+\|_{Fr}^2, \end{aligned}$$

where $\sigma > 0$ is the penalty parameter. In [20], L_σ is called the *proximal augmented Lagrangian*.

If (x_*, Z_*, λ_*) is any point satisfying (KKT), from (2.2) it can be shown that

$$(3.2) \quad L_\sigma(x_*, Z_*, \lambda_*) = f(x_*).$$

In the next theorem we prove that $L_\sigma(\cdot, Z_*, \lambda_*)$ is an exact penalty function when σ is sufficiently large.

THEOREM 4. *Let us assume that f , h , and $\mathcal{A}$ are twice differentiable at x_* and that (x_*, Z_*, λ_*) satisfies (KKT) conditions and the second-order sufficient condition (2.11). Then there is a neighborhood V of x_* and a real $\bar{\sigma} > 0$ such that for all $\sigma \geq \bar{\sigma}$, (x_*, Z_*, λ_*) is a saddle-point of L_σ on $V \times (S^m \times \mathbb{R}^p)$. Moreover,*

$$L_\sigma(x, Z_*, \lambda_*) > L_\sigma(x_*, Z_*, \lambda_*) \geq L_\sigma(x_*, Z, \lambda)$$

for all $(x, Z, \lambda) \in V \times S^m \times \mathbb{R}^p$ with $x \neq x_*$.

Proof. Since the operator $(\cdot)_+$, defined in (1.1), is the projection on S_+^m , we have that

$$(3.3) \quad \|(-\sigma^{-1}Z - \mathcal{A}(x))_+ - (-\sigma^{-1}Z - \mathcal{A}(x))\|_{Fr}^2 \leq \|W - (-\sigma^{-1}Z - \mathcal{A}(x))\|_{Fr}^2$$

for all $W \in S_+^m$, and then

$$(3.4) \quad \begin{aligned} \frac{\sigma}{2} \|(-\sigma^{-1}Z - \mathcal{A}(x))_+ + \mathcal{A}(x)\|_{Fr}^2 + \text{Tr}(Z[(-\sigma^{-1}Z - \mathcal{A}(x))_+ + \mathcal{A}(x)]) \\ \leq \text{Tr}(Z[W + \mathcal{A}(x)]) + \frac{\sigma}{2} \|W + \mathcal{A}(x)\|_{Fr}^2 \end{aligned}$$

for all $W \in S_+^m$. Taking $x = x_*$ and $W = -\mathcal{A}(x_*)$ (which belongs to S_+^m), we get

$$\frac{\sigma}{2} \|(-\sigma^{-1}Z - \mathcal{A}(x_*))_+ + \mathcal{A}(x_*)\|_{Fr}^2 + \text{Tr}(Z[(-\sigma^{-1}Z - \mathcal{A}(x_*))_+ + \mathcal{A}(x_*)]) \leq 0;$$

hence

$$L_\sigma(x_*, Z, \lambda) \leq f(x_*) = L_\sigma(x_*, Z_*, \lambda_*) \quad \forall Z \in S^m, \forall \lambda \in \mathbb{R}^p, \forall \sigma > 0.$$

Let us now prove the second inequality. Let $\bar{B}_\varepsilon(x_*)$ be a closed ball with center x_* and radius ε such that $f(x) > f(x_*)$ for all feasible points $x \in \bar{B}_\varepsilon(x_*)$, $x \neq x_*$

x_* . We prove that for all $\sigma > 0$ sufficiently large, x_* is the unique point satisfying $\inf_{x \in \bar{B}_\varepsilon(x_*)} L_\sigma(x, Z_*, \lambda_*) = f(x_*)$. For this purpose, we define the problem:

$$(3.5) \quad \psi_\sigma := \inf_{\substack{(x, W) \in \bar{B}_\varepsilon(x_*) \times S^m \\ \mathcal{A}(x) \preceq W}} \left\{ f(x) + \text{Tr}(Z_* W) + \lambda_*^T h(x) + \frac{\sigma}{2} (\|h(x)\|^2 + \|W\|_{Fr}^2) \right\},$$

and from inequality (3.4) we can deduce that

$$(3.6) \quad \psi_\sigma = \inf_{x \in \bar{B}_\varepsilon(x_*)} L_\sigma(x, Z_*, \lambda_*).$$

To conclude, we show that $(x_*, 0, Z_*, 0)$ is a point that satisfies the Karush–Kuhn–Tucker and the second-order sufficient conditions for the optimization problem (3.5). The Lagrangian associated with minimization problem (3.5) is

$$(3.7) \quad \begin{aligned} \tilde{L}(x, W, \Omega, \alpha) := & f(x) + \text{Tr}(Z_* W) + \lambda_*^T h(x) + \frac{\sigma}{2} \|h(x)\|^2 \\ & + \frac{\sigma}{2} \|W\|_{Fr}^2 + \frac{\alpha}{2} (\|x - x_*\|^2 - \varepsilon^2) + \text{Tr}(\Omega(\mathcal{A}(x) - W)), \end{aligned}$$

and the (KKT) conditions are

$$\begin{aligned} \nabla f(x) + \nabla h(x)^T \lambda_* + \sigma \nabla h(x)^T h(x) + \alpha(x - x_*) + D\mathcal{A}(x)^* \Omega &= 0, \\ Z_* + \sigma W &= \Omega, \\ \frac{\alpha}{2} (\|x - x_*\|^2 - \varepsilon^2) &= 0, \\ \text{Tr}(\Omega(\mathcal{A}(x) - W)) &= 0, \\ \alpha \geq 0, \quad \Omega \succeq 0, \\ \|x - x_*\| - \varepsilon &\leq 0, \\ \mathcal{A}(x) - W &\preceq 0. \end{aligned}$$

It can be easily seen that $(x, W, \Omega, \alpha) = (x_*, 0, Z_*, 0)$ satisfies all these conditions. In what follows we will state the second-order sufficient condition

$$(3.8) \quad (d^T, U) \nabla^2 \tilde{L}(x_*, 0, Z_*, 0) \begin{pmatrix} d \\ U \end{pmatrix} > 0,$$

for any nonzero vector $(d, U) \in \tilde{C}(x_*, 0)$.

The Hessian of $\tilde{L}$ with respect to the variables (x, W) at $(x_*, 0, Z_*, 0)$ is given by

$$(3.9) \quad \nabla_{(x, W)}^2 \tilde{L}(x_*, 0, Z_*, 0) = \left(\begin{array}{c|c} \nabla^2 f(x_*) + \sum_{j=1}^p \lambda_{*j} \nabla^2 h_j(x_*) + \mathcal{H} & 0 \\ \hline + \sigma \sum_{j=1}^p h_j(x_*) \nabla^2 h_j(x_*) + \sigma \nabla h(x_*)^T \nabla h(x_*) & \sigma I_m \end{array} \right),$$

with $\mathcal{H} := [\text{Tr}(Z_* \frac{\partial^2 \mathcal{A}(x_*)}{\partial x_i \partial x_j})]_{i,j}$, and the cone of critical directions for problem (3.5) is

$$\begin{aligned} \tilde{C}(x_*, 0) = & \left\{ (d, U) \in \mathbb{R}^n \times S^m : \nabla f(x_*)^T d + \text{Tr}(Z_* U) + \lambda_*^T \nabla h(x_*) d = 0, \right. \\ & \left. \sum_{i=1}^n d_i \frac{\partial \mathcal{A}(x_*)}{\partial x_i} - U \in S_-^m(\text{Ker } \mathcal{A}(x_*)) \right\}. \end{aligned}$$

Thus, condition (3.8) can be written as

$$(3.10) \quad d^T \nabla^2 f(x_*)d + \sum_{j=1}^p \lambda_{*j} d^T \nabla^2 h_j(x_*)d + d^T \mathcal{H}d + \sigma \|\nabla h(x_*)d\|^2 + \sigma \|U\|_{Fr}^2 > 0$$

for any nonzero vector $(d, U) \in \tilde{C}(x_*, 0)$.

The case when $d = 0$ and $U \neq 0$ is trivial. Another easy case is when we take $(d, U) \in \tilde{C}(x_*, 0)$ satisfying either $\|U\|_{Fr} > \delta \|d\|$ or $\|\nabla h(x_*)d\| > \delta \|d\|$ for some fixed $\delta > 0$; indeed

$$\begin{aligned} d^T \nabla^2 f(x_*)d + \sum_{j=1}^p \lambda_{*j} d^T \nabla^2 h_j(x_*)d + d^T \mathcal{H}d + \sigma (\|\nabla h(x_*)d\|^2 + \|U\|_{Fr}^2) \\ > d^T \nabla^2 f(x_*)d + \sum_{j=1}^p \lambda_{*j} d^T \nabla^2 h_j(x_*)d + d^T \mathcal{H}d + \sigma \delta^2 \|d\|^2 \\ \geq -\|\nabla^2 f(x_*) + \sum_{j=1}^p \lambda_{*j} \nabla^2 h_j(x_*) + \mathcal{H}\| \|d\|^2 + \sigma \delta^2 \|d\|^2, \end{aligned}$$

and (3.10) is verified by taking $\sigma \geq \sigma_\delta := \frac{1}{\delta^2} \|\nabla^2 f(x_*) + \sum_{j=1}^p \lambda_{*j} \nabla^2 h_j(x_*) + \mathcal{H}\|$.

Finally, we show that such a $\delta > 0$ always exists. We proceed by contradiction. Let us suppose that there is a sequence $\{(d_k, U_k)\}$ in $\tilde{C}(x_*, 0)$ such that

$$(3.11) \quad \|U_k\|_{Fr} \leq \frac{1}{k} \|d_k\|,$$

$$(3.12) \quad \|\nabla h(x_*)d_k\| \leq \frac{1}{k} \|d_k\|$$

and

$$(3.13) \quad d_k^T \nabla^2 f(x_*)d_k + \sum_{j=1}^p \lambda_{*j} d_k^T \nabla^2 h_j(x_*)d_k + d_k^T \mathcal{H}d_k \leq 0 \quad \forall k.$$

If we divide (3.13) by $\|d_k\|^2$ and suppose that $\frac{d_k}{\|d_k\|} \rightarrow \hat{d}$, by taking the limit in this inequality we get

$$(3.14) \quad \hat{d}^T \nabla_{xx}^2 L(x_*, Z_*, \lambda_*) \hat{d} \leq 0,$$

which means, by Proposition 1, that $\hat{d} \notin C(x_*)$.

On the other hand, since $(d_k, U_k) \in \tilde{C}(x_*, 0)$, we have that

$$v^T \sum_{i=1}^n d_{ki} \frac{\partial \mathcal{A}(x_*)}{\partial x_i} v \leq v^T U_k v \quad \forall v \in \text{Ker } \mathcal{A}(x_*) \text{ and } \forall k,$$

and using the fact that $\|U_k\|_{Fr} \geq \frac{v^T U_k v}{\|v\|^2}$, for all $v \neq 0$, together with (3.11) we obtain

$$v^T \sum_{i=1}^n d_{ki} \frac{\partial \mathcal{A}(x_*)}{\partial x_i} v \leq \frac{1}{k} \|d_k\| \|v\|^2 \quad \forall v \in \text{Ker } \mathcal{A}(x_*) \text{ and } \forall k,$$

which implies that

$$\sum_{i=1}^n \hat{d}_i \frac{\partial \mathcal{A}(x_*)}{\partial x_i} \in S_-^m(\text{Ker } \mathcal{A}(x_*)).$$

The equality $\nabla h(x_*)\hat{d} = 0$ follows directly from (3.12) and since $\nabla f(x_*)^T d_k = -\text{Tr}(Z_* U_k) - \lambda_*^T \nabla h(x_*) d_k$, from (3.11) and (3.12) we obtain that $\nabla f(x_*)^T \hat{d} = 0$. In this way we conclude that $\hat{d} \in C(x_*)$, which contradicts (3.14). $\square$

3.3. The Han penalty function. We now define another penalty function associated with problem (P), which will be a key issue in the global algorithm that we will describe in section 4. For $\sigma > 0$ we define

$$(3.15) \quad \theta_\sigma(x) = f(x) + \sigma(\lambda_1(\mathcal{A}(x))_+ + \|h(x)\|).$$

This function comes from the Han penalty function in mathematical programming [3, 11]. In the rest of this section we will prove some properties of θ_σ and its exactness.

In order to compute the directional derivative $\theta'_\sigma(x; d)$, we start by recalling a particular chain rule.

LEMMA 5. *Let $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be a function with directional derivative $\varphi'(x; d) = \lim_{t \rightarrow 0^+} t^{-1}(\varphi(x + td) - \varphi(x))$, and let $\phi : \mathbb{R}^m \rightarrow \mathbb{R}^p$ be a Lipschitz function in a neighborhood of $\varphi(x)$ with directional derivative $\phi'(\varphi(x); \varphi'(x; d))$. Then, the function $(\phi \circ \varphi)$ has a directional derivative at x in the direction d given by*

$$(3.16) \quad (\phi \circ \varphi)'(x; d) = \phi'(\varphi(x); \varphi'(x; d)).$$

Proof. By using the usual notation $o(t)$ for a function verifying $\lim_{t \rightarrow 0} t^{-1}o(t) = 0$, we can write for $t > 0$

$$\begin{aligned} t^{-1}[(\phi \circ \varphi)(x + td) - (\phi \circ \varphi)(x)] &= t^{-1}[\phi(\varphi(x) + t\varphi'(x; d) + o(t)) - (\phi \circ \varphi)(x)] \\ &= t^{-1}[\phi(\varphi(x) + t\varphi'(x; d)) - \phi(\varphi(x))] + t^{-1}o(t), \end{aligned}$$

and we can conclude by taking the limit when $t \rightarrow 0^+$. $\square$

As a consequence of this result we give in the next lemma the directional derivative of the penalty function θ_σ .

LEMMA 6. *If f , h , and $\mathcal{A}$ in (3.15) have directional derivatives at x in the direction d , where x is a feasible point for (P), then θ_σ also has a directional derivative that can be characterized by*

$$\theta'_\sigma(x; d) = f'(x; d) + \sigma(\lambda_1(N^T \mathcal{A}'(x; d)N)_+ + \|h'(x; d)\|),$$

where N is the matrix defined in (2.9).

Proof. Let x be a feasible point. From Lemma 5, we have that

$$\begin{aligned} \theta'_\sigma(x; d) &= f'(x; d) + \sigma([\lambda_1(\mathcal{A}(\cdot))_+]'(x; d) + \|h'(x; d)\|) \\ &= f'(x; d) + \sigma([\lambda_1(\cdot)_+]'(\mathcal{A}(x); \mathcal{A}'(x; d)) + \|h'(x; d)\|). \end{aligned}$$

For $A \in S_-^m$ and $B \in S^m$, we easily check that

$$[\lambda_1(\cdot)_+]'(A; B) = \begin{cases} 0 & \text{if } \lambda_1(A) < 0, \\ \lambda'_1(A; B)_+ & \text{if } \lambda_1(A) = 0, \end{cases}$$

and using formula (2.8) in [7] for the calculus of the directional derivative of $\lambda_1(A) = \max\{x^T Ax : \|x\| = 1\}$, we obtain that $\lambda'_1(A; B) = \max\{x^T Bx : \|x\| = 1 \text{ and } x^T Ax = \lambda_1(A)\}$. Then, if $\lambda_1(A) = 0$, we can write $\lambda'_1(A; B) = \lambda_1(N^T BN)$ where N is a matrix whose columns are an orthonormal base of $\text{Ker } A$.

We conclude replacing A by $\mathcal{A}(x)$, B by $\mathcal{A}'(x; d)$ and recalling that N is the matrix 0 when $\lambda_1(A) < 0$. $\square$

Remark. If f , h , and $\mathcal{A}$ are differentiable at x , then

$$\theta'_\sigma(x; d) = \nabla f(x)d + \sigma \left(\lambda_1 \left(N^T \sum_{i=1}^n d_i \frac{\partial \mathcal{A}(x)}{\partial x_i} N \right)_+ + \|\nabla h(x)d\| \right).$$

In the following proposition we give a lower bound for the parameter σ in order to obtain the exactness of θ_σ .

PROPOSITION 7. *If x_* is a feasible point of (P) and θ_σ has a (strict) local minimum at x_* , then x_* is a (strict) local minimum of (P). Furthermore, if f , h , and $\mathcal{A}$ are differentiable at x_* and if the transversality condition (2.8) is verified, then $\sigma \geq \max\{\text{Tr}(Z_*), \|\lambda_*\|\}$.*

Proof. If x_* is a local minimum of θ_σ , there is a neighborhood V of x_* such that for all $x \in V$ we have that $\theta_\sigma(x_*) \leq \theta_\sigma(x)$, and since x_* is feasible we obtain

$$\begin{aligned} f(x_*) &= \theta_\sigma(x_*) \leq \theta_\sigma(x) \quad \forall x \in V \\ &= f(x) \quad \forall x \in V, x \text{ feasible,} \end{aligned}$$

which means that x_* is a local minimum of (P). When the minimum x_* is strict, the proof is identical.

Now, let us assume that f , h , and $\mathcal{A}$ are differentiable at x_* and that the transversality condition holds. Since x_* is a local minimum of θ_σ , we have that $\theta'_\sigma(x_*; d) \geq 0$ for all directions d , and using Lemma 6 we can write

$$(3.17) \quad 0 \leq \nabla f(x_*)^T d + \sigma \left(\lambda_1 \left(N^T \sum_{i=1}^n d_i \frac{\partial \mathcal{A}(x_*)}{\partial x_i} N \right)_+ + \|\nabla h(x_*)d\| \right) \quad \forall d \in \mathbb{R}^n.$$

Let us first consider the case when $\text{Ker } \mathcal{A}(x_*) = \{0\}$. This implies $Z_* = 0$ and $N = 0$; hence from inequality (3.17) and the first equation in (KKT), we see that $\sigma \geq \frac{\lambda_*^T \nabla h(x_*)d}{\|\nabla h(x_*)d\|}$ for all nonzero $d \in \mathbb{R}^n$. The surjectivity of $\nabla h(x_*)$ shows that $\sigma \geq \|\lambda_*\|$.

Let us suppose now that $\text{Ker } \mathcal{A}(x_*) \neq \{0\}$. From the first equation in (KKT), inequality (3.17) and equality (2.15), we can write for all $d \in \mathbb{R}^n$

$$\begin{aligned} \sigma \left(\|\nabla h(x_*)d\| + \lambda_1 \left(N^T \sum_{i=1}^n d_i \frac{\partial \mathcal{A}(x_*)}{\partial x_i} N \right)_+ \right) &\geq \lambda_*^T \nabla h(x_*)d + \text{Tr} \left(Z_* \sum_{i=1}^n d_i \frac{\partial \mathcal{A}(x_*)}{\partial x_i} \right) \\ &= \lambda_*^T \nabla h(x_*)d + \text{Tr} \left(\phi_* N^T \sum_{i=1}^n d_i \frac{\partial \mathcal{A}(x_*)}{\partial x_i} N \right). \end{aligned}$$

Then the transversality condition (surjectivity of ψ) allows us to say that

$$(3.18) \quad \sigma(\|v\| + \lambda_1(W)_+) \geq \lambda_*^T v + \text{Tr}(\phi_* W)$$

for all $(v, W) \in \mathbb{R}^p \times S_r$, and we can conclude from the inequality

$$\|W\|_2 := \sqrt{\max\{\lambda^2 : \lambda \text{ is an eigenvalue of } W\}} \geq \lambda_1(W)_+$$

and the equality

$$\|(\lambda_*, \phi_*)\|_D := \sup\{|\lambda_*^T v + \text{Tr}(\phi_* W)| : \|v\| + \|W\|_2 = 1\} = \max\{\|\lambda_*\|, \text{Tr}(Z_*)\}. \quad \square$$

We conclude this section establishing sufficient conditions for exactness of the Han penalty function θ_σ . In Proposition 9 we consider the convex case and in Theorem 10 the general one.

The following useful lemma is a direct consequence of inequality (2.4).

LEMMA 8. *If $Z \succeq 0$ and $\sigma \geq \max\{\text{Tr}(Z), \|\lambda\|\}$, then $L(\cdot, Z, \lambda) \leq \theta_\sigma(\cdot)$.*

PROPOSITION 9. *Let us consider the convex problem (P_C) , defined in section 3.1, and let us suppose that f , h , and $\mathcal{A}$ are differentiable at a solution x_* of (P_C) . Then, if (Z_*, λ_*) are (KKT)-multipliers associated with x_* and $\sigma \geq \max\{\text{Tr}(Z_*), \|\lambda_*\|\}$, we have that θ_σ has a global minimum in x_* .*

Proof. Let us suppose that (x_*, Z_*, λ_*) satisfies the (KKT) conditions. For the convex problem (P_C) , it can be easily seen that $L(\cdot, Z_*, \lambda_*)$ has a global minimum at x_* , that is, $\theta_\sigma(x_*) = f(x_*) = L(x_*, Z_*, \lambda_*) \leq L(x, Z_*, \lambda_*)$ for all x . If $\sigma \geq \max\{\text{Tr}(Z_*), \|\lambda_*\|\}$, from Lemma 8 we have $L(\cdot, Z_*, \lambda_*) \leq \theta_\sigma(\cdot)$, which leads to the desired result. $\square$

THEOREM 10. *Let us suppose that f , h , and $\mathcal{A}$ are differentiable at x_* . Let (x_*, Z_*, λ_*) be a point that satisfies the (KKT) conditions and the second-order sufficient condition (2.11). If $\sigma > \max\{\text{Tr}(Z_*), \|\lambda_*\|\}$, then θ_σ has a strict local minimum in x_* .*

Proof. Taking $Z = Z_*$ and $W = (-\mathcal{A}(x))_+ = -(\mathcal{A}(x))_-$ in (3.4) we have that

$$(3.19) \quad \frac{r}{2} \left\| \left(-\frac{Z_*}{r} - \mathcal{A}(x) \right)_+ + \mathcal{A}(x) \right\|_{F_r}^2 + \text{Tr} \left(Z_* \left[\left(-\frac{Z_*}{r} - \mathcal{A}(x) \right)_+ + \mathcal{A}(x) \right] \right) \\ \leq \text{Tr}(Z_* \mathcal{A}(x)_+) + \frac{r}{2} \|\mathcal{A}(x)_+\|_{F_r}^2.$$

Hence, using the Cauchy-Schwarz inequality and the Von Neumann-Theobald inequality (2.4), we obtain

$$L_r(x, Z_*, \lambda_*) \leq f(x) + \lambda_*^T h(x) + \frac{r}{2} \|h(x)\|^2 + \text{Tr}(Z_* \mathcal{A}(x)_+) + \frac{r}{2} \|\mathcal{A}(x)_+\|_{F_r}^2 \\ \leq f(x) + \|h(x)\| \left(\|\lambda_*\| + \frac{r}{2} \|h(x)\| \right) + \lambda_1(\mathcal{A}(x)_+) \left(\text{Tr}(Z_*) + \frac{r}{2} \sum_{j=1}^m \lambda_j(\mathcal{A}(x)_+) \right).$$

The last inequality follows from $\lambda_1(\mathcal{A}(x)_+) = \lambda_1(\mathcal{A}(x))_+$. Since $\sigma > \max\{\text{Tr}(Z_*), \|\lambda_*\|\}$, for any fixed $r > 0$ there is a neighborhood V_r of x_* such that

$$L_r(x, Z_*, \lambda_*) \leq f(x) + \sigma(\|h(x)\| + \lambda_1(\mathcal{A}(x))_+) = \theta_\sigma(x) \quad \forall x \in V_r.$$

From Theorem 4, we know that there is an $\bar{r} > 0$ and a neighborhood $\bar{V}$ of x_* where x_* is a strict minimum of $L_{\bar{r}}(\cdot, Z_*, \lambda_*)$. This implies that x_* is a strict minimum of θ_σ on $\bar{V} \cap V_{\bar{r}}$. $\square$

4. Sequentially semidefinite programming. In this section we propose a global S-SDP algorithm for solving problem (P). This algorithm is inspired by the classical sequentially quadratic programming (SQP). We begin by recalling the local S-SDP algorithm proposed in [9] and its convergence theorem.

Given an initial point (x_0, Z_0, λ_0) close to a point (x_*, Z_*, λ_*) that satisfies the (KKT) conditions, we generate a sequence (x_k, Z_k, λ_k) by solving the linearized problem:

$$\begin{aligned} & \text{minimize}_{d \in \mathbb{R}^n} && \nabla f(x_k)^T d + \frac{1}{2} d^T M_k d \\ (T_k) \quad & \text{subject to} && \mathcal{A}_k(d) \preceq 0, \\ & && h(x_k) + \nabla h(x_k) d = 0, \end{aligned}$$

where $\mathcal{A}_k(d) := \mathcal{A}(x_k) + \sum_{i=1}^n d_i \frac{\partial \mathcal{A}(x_k)}{\partial x_i}$ and the matrix M_k replaces the Hessian $\nabla_{xx}^2 L(x_k, Z_k, \lambda_k)$, emulating the so-called quasi-Newton methods.

If d_k is the solution of (T_k) , we define $x_{k+1} = x_k + d_k$. The point $(d_k, Z_{k+1}, \lambda_{k+1})$ is obtained from the (KKT) conditions for the minimization problem (T_k) , that is,

$$(4.1a) \quad \nabla f(x_k) + D\mathcal{A}(x_k)^* Z_{k+1} + \nabla h(x_k)^T \lambda_{k+1} + M_k d_k = 0,$$

$$(4.1b) \quad \mathcal{A}_k(d_k) \preceq 0,$$

$$(4.1c) \quad h(x_k) + \nabla h(x_k) d_k = 0,$$

$$(4.1d) \quad Z_{k+1} \succeq 0,$$

$$(4.1e) \quad \text{Tr}(Z_{k+1} \mathcal{A}_k(d_k)) = 0.$$

These conditions will be called (KKT_k) in what follows.

THEOREM 11. *Let (x_*, Z_*, λ_*) be a point satisfying the (KKT) conditions and the second-order sufficient condition (2.11). Suppose that $(D\mathcal{A}(x_*), \nabla h(x_*)^T)^T$ has full rank and that $M_k \rightarrow \nabla_{xx}^2 L(x_*, Z_*, \lambda_*)$. Then there is $\delta > 0$ such that if $\|x_0 - x_*\| < \delta$, $\|(Z_0, \lambda_0) - (Z_*, \lambda_*)\| < \delta$ and $\|M_k - \nabla_{xx}^2 L(x_*, Z_*, \lambda_*)\| < \delta$ for all k , the sequence (x_k, Z_k, λ_k) generated by the algorithm S-SDP is well defined and converges superlinearly to (x_*, Z_*, λ_*) . The convergence is even quadratic if $M_k - \nabla_{xx}^2 L(x_*, Z_*, \lambda_*) = O(\|x_k - x_*\| + \|(Z_k, \lambda_k) - (Z_*, \lambda_*)\|)$.*

Our purpose here is to extend the S-SDP algorithm to obtain global convergence. For this, we consider the Han penalty function, defined in (3.15), and an *Armijo line search*.

In the following proposition we prove that the solution d_k of (T_k) is a descent direction for θ_σ at the point x_k when M_k is positive definite and σ is sufficiently large.

PROPOSITION 12. *Suppose that f , h , and $\mathcal{A}$ are C^1 functions and that their derivatives are locally Lipschitz at x_k . Using the penalty function θ_σ , defined in (3.15), if the point $(d_k, Z_{k+1}, \lambda_{k+1})$ verifies the (KKT_k) conditions, written in (4.1), then*

$$\begin{aligned} (4.2) \quad \theta'_\sigma(x_k; d_k) &\leq \nabla f(x_k)^T d_k - \sigma(\lambda_1(\mathcal{A}(x_k))_+ + \|h(x_k)\|) \\ &= -d_k^T M_k d_k + \text{Tr}(Z_{k+1} \mathcal{A}(x_k)) + \lambda_{k+1}^T h(x_k) - \sigma(\lambda_1(\mathcal{A}(x_k))_+ + \|h(x_k)\|). \end{aligned}$$

Furthermore, if $\sigma \geq \max\{\text{Tr}(Z_{k+1}), \|\lambda_{k+1}\|\}$ we obtain

$$(4.3) \quad \theta'_\sigma(x_k; d_k) \leq -d_k^T M_k d_k.$$

Proof. Let us fix $t \in [0, 1]$. By (4.1c) we have that

$$\|h(x_k) + t \nabla h(x_k) d_k\| = (1 - t) \|h(x_k)\|,$$

and by the convexity of $\lambda_1(\cdot)_+$ and (4.1b) we obtain

$$\begin{aligned}\lambda_1(\mathcal{A}_k(td_k))_+ &= \lambda_1\left(\mathcal{A}(x_k) + t \sum_{i=1}^n d_{ki} \frac{\partial \mathcal{A}(x_k)}{\partial x_i}\right)_+ \\ &\leq (1-t)\lambda_1(\mathcal{A}(x_k))_+ + t\lambda_1(\mathcal{A}_k(d_k))_+ = (1-t)\lambda_1(\mathcal{A}(x_k))_+.\end{aligned}$$

From these relations we have that

$$(4.4a) \quad \|\cdot\|'(h(x_k); \nabla h(x_k)d_k) = -\|h(x_k)\|,$$

$$(4.4b) \quad (\lambda_1(\cdot)_+)' \left(\mathcal{A}(x_k); \sum_{i=1}^n d_{ki} \frac{\partial \mathcal{A}(x_k)}{\partial x_i} \right) \leq -\lambda_1(\mathcal{A}(x_k))_+.$$

and applying Lemma 5 we get

$$\theta'_\sigma(x_k; d_k) \leq \nabla f(x_k)^T d_k - \sigma(\lambda_1(\mathcal{A}(x_k))_+ + \|h(x_k)\|),$$

and from (4.1a), (4.1c), and (4.1e) we obtain

$$\theta'_\sigma(x_k; d_k) \leq \text{Tr}(Z_{k+1}\mathcal{A}(x_k)) + h(x_k)^T \lambda_{k+1} - \sigma(\|h(x_k)\| + \lambda_1(\mathcal{A}(x_k))_+) - d_k^T M_k d_k.$$

Finally, if $\sigma \geq \max\{\text{Tr}(Z_{k+1}), \|\lambda_{k+1}\|\}$, the Cauchy–Schwarz inequality and the Von Neumann–Theobald inequality (2.4) lead to the result. $\square$

We are now ready to describe the iteration k of the global algorithm for solving problem (P). We suppose that x_k is known and that M_k is positive definite.

Step 1. Compute a point $(d_k, Z_{k+1}, \lambda_{k+1})$ satisfying (KKT_k) in (4.1).

Step 2. Compute σ_k satisfying $\sigma_k \geq \max\{\text{Tr}(Z_{k+1}), \|\lambda_{k+1}\|\}$ in such a way that the sequence $\{\sigma_k\}$ satisfies the following properties:

$$(4.5) \quad \begin{aligned} &(a) \quad \sigma_k \geq \max\{\text{Tr}(Z_{k+1}), \|\lambda_{k+1}\|\} + \bar{\sigma}. \\ &(b) \quad \text{For all } k \geq k_1, \\ &\quad \text{if } \sigma_{k-1} \geq \max\{\text{Tr}(Z_{k+1}), \|\lambda_{k+1}\|\} + \bar{\sigma}, \text{ then } \sigma_k = \sigma_{k-1}. \\ &(c) \quad \text{If } \{\sigma_k\} \text{ is bounded, then } \sigma_k \text{ is modified just finitely many times,} \end{aligned}$$

where $k_1 \in \mathbb{N}$ and $\bar{\sigma} > 0$ are fixed parameters. A simple way to update σ_k verifying (4.5) is defining $\sigma_k = \max\{1.5\sigma_{k-1}, \max\{\text{Tr}(Z_{k+1}), \|\lambda_{k+1}\|\} + \bar{\sigma}\}$ when (b) fails.

Step 3. The step length α_k is computed by using an *Armijo search rule*; that is, α_k is an approximation of the maximum $\alpha \in (0, 1]$ which verifies

$$(4.6) \quad \theta_{\sigma_k}(x_k + \alpha d_k) \leq \theta_{\sigma_k}(x_k) + w\alpha\Delta_k,$$

where $0 < w < 1$ and Δ_k is the upper bound of $\theta'_{\sigma_k}(x_k; d_k)$ given in (4.2). More precisely, α_k can be computed as follows:

Step 0. $j := 0, \quad r_j := 1.$

Step 1. If (4.6) is satisfied with $\alpha = r_j$, then $\alpha_k = r_j$ and stop the line search.

Step 2. If not, take $r_{j+1} = \beta r_j$, increase j by one and go to Step 1 with $\beta \in (0, 1)$ a fixed constant.

Step 4. Define $x_{k+1} = x_k + \alpha_k d_k$.

Remark. The existence of the Armijo step α satisfying (4.6) is a consequence of inequality (4.3) and the fact that M_k is chosen positive definite.

THEOREM 13. *Let us suppose that f , h , and $\mathcal{A}$ are C^1 functions and that their derivatives are Lipschitz. If we consider the global algorithm described in Steps 1 to 4 and suppose that the matrices M_k are chosen positive definite such that the sequence $\{M_k\}$ is bounded together with the sequence $\{M_k^{-1}\}$, then one of the following situations occurs for the sequence $\{(x_k, Z_{k+1}, \lambda_{k+1})\}$:*

1. *The sequences $\{\sigma_k\}$ and $\{(Z_{k+1}, \lambda_{k+1})\}$ are unbounded.*
2. *There is an index k_2 such that σ_k is constant for all $k \geq k_2$. In this case one of the following situations occurs:*
 - (a) $\theta_{\sigma_k}(x_k) \rightarrow -\infty$, or
 - (b) $\nabla_x L(x_k, Z_{k+1}, \lambda_{k+1}) \rightarrow 0$, $h(x_k) \rightarrow 0$, $\lambda_1(\mathcal{A}(x_k))_+ \rightarrow 0$, and $\text{Tr}(Z_{k+1}\mathcal{A}(x_k)) \rightarrow 0$.

Proof. 1. The equivalence between the unboundedness of $\{\sigma_k\}$ and $\{(Z_{k+1}, \lambda_{k+1})\}$ is direct from (4.5a) and (4.5b).

2. Let us suppose that $\{\sigma_k\}$ is bounded. By (4.5c) we know that there is an index k_2 such that $\sigma_k = \sigma := \sigma_{k_2}$ for all $k \geq k_2$.

To conclude we prove that if (a) is not true, then (b) holds. From (4.6), with $\alpha = \alpha_k$, we know that the sequence $\{\theta_{\sigma_k}(x_k)\}$ is decreasing for all $k \geq k_2$, and then $\theta_{\sigma_k}(x_k) \geq C$ for some constant C , obtaining again from (4.6) the limit $\alpha_k \Delta_k \rightarrow 0$.

All limits in (b) are consequences of the existence of $\bar{\alpha} > 0$ such that $\alpha_k \geq \bar{\alpha}$ for all $k \geq k_2$, which implies from the limit above that $\Delta_k \rightarrow 0$. Indeed, inequalities

$$\begin{aligned} \Delta_k &\leq -d_k^T M_k d_k + \lambda_1(\mathcal{A}(x_k))_+ \text{Tr}(Z_{k+1}) + \|\lambda_{k+1}\| \|h(x_k)\| - \sigma_k (\lambda_1(\mathcal{A}(x_k))_+ + \|h(x_k)\|) \\ &\leq -d_k^T M_k d_k + (\sigma_k - \bar{\sigma}) (\lambda_1(\mathcal{A}(x_k))_+ + \|h(x_k)\|) - \sigma_k (\lambda_1(\mathcal{A}(x_k))_+ + \|h(x_k)\|) \end{aligned}$$

prove that

$$(4.7) \quad \Delta_k \leq -d_k^T M_k d_k - \bar{\sigma} (\lambda_1(\mathcal{A}(x_k))_+ + \|h(x_k)\|) \leq 0,$$

which implies that if $\Delta_k \rightarrow 0$, then

$$\lambda_1(\mathcal{A}(x_k))_+ \rightarrow 0 \quad \text{and} \quad h(x_k) \rightarrow 0.$$

Inequality (4.7) also shows that $-d_k^T M_k d_k \rightarrow 0$, and together with the fact that $\{M_k^{-1}\}$ is bounded, it is easy to conclude that $d_k \rightarrow 0$. This, together with (4.1a) and the boundedness of $\{M_k\}$, implies that

$$\nabla_x L(x_k, Z_{k+1}, \lambda_{k+1}) \rightarrow 0.$$

Finally, by definition of Δ_k we know that

$$\nabla f(x_k)^T d_k = \Delta_k + \sigma_k (\lambda_1(\mathcal{A}(x_k))_+ + \|h(x_k)\|) \rightarrow 0,$$

and from (4.1a) we see that

$$\text{Tr} \left(Z_{k+1} \sum_{i=1}^n d_{ki} \frac{\partial \mathcal{A}(x_k)}{\partial x_i} \right) + \lambda_{k+1}^T \nabla h(x_k) d_k = -d_k^T M_k d_k - \nabla f(x_k)^T d_k \rightarrow 0.$$

This limit and the equalities (4.1e) and (4.1c), together with the boundedness of $\{\lambda_{k+1}\}$, allow us to write

$$\begin{aligned} \lim_{k \rightarrow +\infty} \text{Tr}(Z_{k+1}\mathcal{A}(x_k)) &= \lim_{k \rightarrow +\infty} -\text{Tr} \left(Z_{k+1} \sum_{i=1}^n d_{ki} \frac{\partial \mathcal{A}(x_k)}{\partial x_i} \right) \\ &= \lim_{k \rightarrow +\infty} \lambda_{k+1}^T \nabla h(x_k) d_k = \lim_{k \rightarrow +\infty} -\lambda_{k+1}^T h(x_k) = 0. \end{aligned}$$

Let us prove now that $\alpha_k \geq \bar{\alpha} > 0$. If $\alpha_k < 1$, by the Armijo search rule, there is an $r_j \in (0, 1]$ such that $\alpha_k = \beta r_j$ and

$$(4.8) \quad \theta_{\sigma_k}(x_k + r_j d_k) > \theta_{\sigma_k}(x_k) + w r_j \Delta_k.$$

Let us consider the first-order Taylor expansion

$$\begin{aligned} f(x_k + r_j d_k) &= f(x_k) + r_j \nabla f(x_k)^T d_k + O(r_j^2 \|d_k\|^2), \\ h(x_k + r_j d_k) &= h(x_k) + r_j \nabla h(x_k)^T d_k + O(r_j^2 \|d_k\|^2), \\ &= (1 - r_j)h(x_k) + r_j(h(x_k) + \nabla h(x_k)^T d_k) + O(r_j^2 \|d_k\|^2), \\ \mathcal{A}(x_k + r_j d_k) &= \mathcal{A}(x_k) + r_j \sum_{i=1}^n d_{ki} \frac{\partial \mathcal{A}(x_k)}{\partial x_i} + O(r_j^2 \|d_k\|^2) \\ &= (1 - r_j)\mathcal{A}(x_k) + r_j \left(\mathcal{A}(x_k) + \sum_{i=1}^n d_{ki} \frac{\partial \mathcal{A}(x_k)}{\partial x_i} \right) + O(r_j^2 \|d_k\|^2). \end{aligned}$$

Since $r_j \leq 1$, from the convexity of $\lambda_1(\cdot)_+$ and the relations (4.1b) and (4.1c) we obtain

$$\begin{aligned} \|h(x_k + r_j d_k)\| &= (1 - r_j)\|h(x_k)\| + O(r_j^2 \|d_k\|^2), \\ \lambda_1(\mathcal{A}(x_k + r_j d_k))_+ &\leq (1 - r_j)\lambda_1(\mathcal{A}(x_k))_+ + O(r_j^2 \|d_k\|^2), \end{aligned}$$

which imply from (4.8) the inequality

$$\theta_{\sigma_k}(x_k) + r_j \Delta_k + C_1 r_j^2 \|d_k\|^2 > \theta_{\sigma_k}(x_k) + w r_j \Delta_k,$$

that is, $-(1 - w)r_j \Delta_k < C_1 r_j^2 \|d_k\|^2$ for some constant $C_1 > 0$. Due to inequality (4.7), the boundedness of $\{M_k^{-1}\}$, and the fact that M_k is positive definite, we see that $\Delta_k \leq -C_2 \|d_k\|^2$ for some constant $C_2 > 0$. The last two inequalities show that

$$r_j \geq \frac{C_2}{C_1} (1 - w) > 0,$$

and the proof is complete. $\square$

Remark. The notion of “global convergence” characterized by the situations 1 and 2 in Theorem 13 is fairly standard. However, it should be noted that the “pathological” situations 1 and 2(a) can happen.

Acknowledgments. The authors are grateful to Frédéric Bonnans and Dominik Noll for fruitful discussions concerning the subject of this paper.

REFERENCES

- [1] P. APKARIAN, D. NOLL, AND D. TUAN, *A Prototype Primal-Dual LMI-Interior Algorithm for Nonconvex Robust Control Problems*, Technical report, 01–08, MIP, UMR 5640, <http://mip.ups-tlse.fr/publi/publi.html> (2001).
- [2] J. F. BONNANS, R. COMINETTI, AND A. SHAPIRO, *Second-order optimality conditions based on parabolic second-order tangent sets*, SIAM J. Optim., 9 (1999), pp. 466–492.
- [3] J. F. BONNANS, J.-C. GILBERT, C. LEMARÉCHAL, AND C. SAGASTIZÁBAL, *Numerical Optimization*, Springer-Verlag, Berlin, 2002.
- [4] J. F. BONNANS AND A. SHAPIRO, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, New York, 2000.

- [5] S. BOYD, L. EL GHAOU, E. FERON, AND V. BALAKRISHNAN, *Linear Matrix Inequalities in System and Control Theory*, SIAM Stud. Appl. Math. 15, SIAM, Philadelphia, 1994.
- [6] R. COMINETTI, *Metric regularity, tangent sets and second-order optimality conditions*, Appl. Math. Optim., 21 (1990), pp. 265–287.
- [7] R. CORREA AND A. SEEGER, *Directional derivatives of minimax functions*, Nonlinear Anal., 9 (1985), pp. 13–22.
- [8] B. FARES, P. APKARIAN, AND D. NOLL, *An augmented Lagrangian method for a class of LMI-Constrained problems in robust control theory*, Int. J. Control, 74 (2001), pp. 348–360.
- [9] B. FARES, D. NOLL, AND P. APKARIAN, *Robust control via sequential semidefinite programming*, SIAM J. Control Optim., 40 (2002), pp. 1791–1820.
- [10] M. GOLUBITSKY AND V. GUILLEMIN, *Stable Mappings and Their Singularities*, Springer-Verlag, New York, 1973.
- [11] S. P. HAN, *A globally convergent method for nonlinear programming*, J. Optim. Theory Appl., 22 (1977), pp. 297–309.
- [12] J.-B. HIRIART-URRUTY AND C. LEMARÉCHAL, *Convex Analysis and Minimization Algorithms I*, Springer-Verlag, Berlin, 1993.
- [13] R. HORN AND C. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.
- [14] S. KURCYUSZ, *On the existence and nonexistence of Lagrange multipliers in Banach spaces*, J. Optim. Theory Appl., 20 (1976), pp. 81–110.
- [15] J. NOCEDAL AND S. J. WRIGHT, *Numerical Optimization*, Springer Ser. Oper. Res., Springer-Verlag, New York, 1999.
- [16] D. NOLL, M. TORKI, AND P. APKARIAN, *Partially augmented Lagrangian method for matrix inequality constraints*, MIP, Université Paul Sabatier Toulouse 3 - CNRS - INSA Toulouse, France, UMR 5640, 2002.
- [17] S. M. ROBINSON, *First-order conditions for general nonlinear optimization*, SIAM J. Appl. Math., 30 (1976), pp. 597–607.
- [18] S. M. ROBINSON, *Stability theorems for systems of inequalities, Part II: Differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.
- [19] S. M. ROBINSON, *Generalized equations and their solutions, Part II: Applications to nonlinear programming*, Math Programming Stud., 19 (1982), pp. 200–221.
- [20] R. T. ROCKAFELLAR AND R. J.-B. WETS, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
- [21] A. SHAPIRO, *First and second-order Analysis of nonlinear semidefinite programs*, Math. Programming, 77 (1997), pp. 301–320.
- [22] C. M. THEOBALD, *An inequality for the trace of the product of two symmetric matrices*, Math. Proc. Cambridge Philos. Soc., 77 (1975), pp. 77–265.
- [23] L. VANDENBERGHE AND S. BOYD, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.
- [24] J. VON NEUMANN, *Some matrix inequalities and metrization of matrix-space*, Tomsk University Review, 1 (1937), pp. 286–300. In *Collected Works*, Vol. IV, Pergamon, Oxford, 1962, pp. 205–218.
- [25] H. WOLKOWITZ, R. SAIGAL, AND L. VANDENBERGHE, EDS., *Handbook of Semidefinite Programming: Theory, Algorithms and Applications*, Kluwer International Series in Operations Research and Management Science, Kluwer, Boston, 2000.

A NEW ITERATION-COMPLEXITY BOUND FOR THE MTY PREDICTOR-CORRECTOR ALGORITHM*

RENATO D. C. MONTEIRO[†] AND TAKASHI TSUCHIYA[‡]

Abstract. In this paper we present a new iteration-complexity bound for the Mizuno–Todd–Ye predictor-corrector (MTY P-C) primal-dual interior-point algorithm for linear programming. The analysis of the paper is based on the important notion of crossover events introduced by Vavasis and Ye. For a standard form linear program $\min\{c^T x : Ax = b, x \geq 0\}$ with decision variable $x \in \mathbb{R}^n$, we show that the MTY P-C algorithm, started from a well-centered interior-feasible solution with duality gap $n\mu_0$, finds an interior-feasible solution with duality gap less than $n\eta$ in $\mathcal{O}(T(\mu_0/\eta) + n^{3.5} \log(\bar{\chi}_A^*))$ iterations, where $T(t) \equiv \min\{n^2 \log(\log t), \log t\}$ for all $t > 0$ and $\bar{\chi}_A^*$ is a scaling invariant condition number associated with the matrix A . More specifically, $\bar{\chi}_A^*$ is the infimum of all the conditions numbers $\bar{\chi}_{AD}$, where D varies over the set of positive diagonal matrices. Under the setting of the Turing machine model, our analysis yields an $\mathcal{O}(n^{3.5} L_A + \min\{n^2 \log L, L\})$ iteration-complexity bound for the MTY P-C algorithm to find a primal-dual optimal solution, where L_A and L are the input sizes of the matrix A and the data (A, b, c) , respectively. This contrasts well with the classical iteration-complexity bound for the MTY P-C algorithm, which depends linearly on L instead of $\log L$.

Key words. interior-point algorithms, primal-dual algorithms, path-following, central path, layered least squares steps, condition number, polynomial complexity, crossover events, scale-invariance, predictor-corrector, affine scaling, strongly polynomial, linear programming

AMS subject classifications. 65K05, 68Q25, 90C05, 90C51, 90C60

DOI. 10.1137/S1052623402416803

1. Introduction. We consider the linear programming (LP) problem

$$(1.1) \quad \begin{array}{ll} \text{minimize}_x & c^T x \\ \text{subject to} & Ax = b, \quad x \geq 0, \end{array}$$

and its associated dual problem

$$(1.2) \quad \begin{array}{ll} \text{maximize}_{(y,s)} & b^T y \\ \text{subject to} & A^T y + s = c, \quad s \geq 0, \end{array}$$

where $A \in \mathbb{R}^{m \times n}$, $c \in \mathbb{R}^n$, and $b \in \mathbb{R}^m$ are given, and the vectors $x, s \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ are the unknown variables.

Karmarkar in his seminal paper [6] proposed the first polynomially convergent interior-point method with an $\mathcal{O}(nL)$ iteration-complexity bound, where L is the size of the LP instance (1.1). The first path-following interior-point algorithm was proposed by Renegar in his breakthrough paper [20]. Renegar’s method closely follows the primal central path and exhibits an $\mathcal{O}(\sqrt{n}L)$ iteration-complexity bound.

*Received by the editors October 31, 2002; accepted for publication (in revised form) July 2, 2004; published electronically December 30, 2004.

<http://www.siam.org/journals/siopt/15-2/41680.html>

[†]School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332 (monteiro@isye.gatech.edu). This author was supported in part by NSF grants CCR-9902010, CCR-0203113, and INT-9910084 and by ONR grant N00014-03-1-0401.

[‡]The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-Ku, Tokyo, 106-8569, Japan (tsuchiya@sun312.ism.ac.jp). This author was supported in part by Japan-U.S. Joint Research Projects of Japan Society for the Promotion of Science “Algorithms for linear programs over symmetric cones” and the Grants-in-Aid for Scientific Research (C) 08680478 and (C) 15510144 of the Ministry of Science, Technology, Education and Culture of Japan.

The first path-following algorithm which simultaneously generates iterates in both the primal and dual spaces was proposed by Kojima, Mizuno, and Yoshise [7] and Tanabe [22], based on ideas suggested by Megiddo [10]. In contrast to Renegar's algorithm, Kojima et al.'s algorithm has an $\mathcal{O}(nL)$ iteration-complexity bound. A primal-dual path-following algorithm with an $\mathcal{O}(\sqrt{n}L)$ iteration-complexity bound was subsequently obtained by Kojima, Mizuno, and Yoshise [8] and Monteiro and Adler [15, 16] independently. Following these developments, many other primal-dual interior-point algorithms for LP have been proposed.

An outstanding open problem in optimization is whether there exists a strongly polynomial algorithm for LP, that is, one whose complexity is bounded by a polynomial of m and n only. A major effort in this direction is due to Tardos [23], who developed a polynomial-time algorithm whose complexity is bounded by a polynomial of m , n , and L_A , where L_A denotes the size of the (integral) matrix A . Such an algorithm gives a strongly polynomial method for the important class of LP problems where the entries of A are either 1, -1 , or 0, e.g., LP formulations of network flow problems. Tardos's algorithm consists of solving a sequence of LP problems of small sizes by a standard polynomially convergent LP method and using their solutions to obtain the solution of the original LP problem.

The development of a method which works entirely in the context of the original LP problem and whose complexity is also bounded by a polynomial of m , n , and L_A is due to Vavasis and Ye [30, 31]. Their method is a primal-dual path-following interior-point algorithm similar to the ones mentioned above except that from time to time it uses a crucial step, namely, the layered least squares (LLS) direction. They showed that their method has an $\mathcal{O}(n^{3.5} \log(\bar{\chi}_A + n))$ iteration-complexity bound, where $\bar{\chi}_A$ is a condition number associated with A which has the property that $\log \bar{\chi}_A = \mathcal{O}(L_A)$ whenever A is integral. The number $\bar{\chi}_A$ was first introduced implicitly by Dikin [1] in the study of primal affine scaling (AS) algorithms and was later studied by several researchers including Vanderbei and Lagarias [29], Todd [24], and Stewart [21]. Properties of $\bar{\chi}_A$ are studied in [4, 27, 28].

The complexity analysis of Vavasis and Ye's algorithm is based on the notion of a crossover event, a combinatorial event concerning the central path. Intuitively, a crossover event occurs between two variables when one of them is larger than the other at a point in the central path and then becomes smaller asymptotically as the optimal solution set is approached. Vavasis and Ye showed that there can be at most $n(n-1)/2$ crossover events and that a distinct crossover event occurs every $\mathcal{O}(n^{1.5} \log(\bar{\chi}_A + n))$ iterations, from which they deduced the overall $\mathcal{O}(n^{3.5} \log(\bar{\chi}_A + n))$ iteration-complexity bound. In [14], an LP instance is given in which the number of crossover events is $\Theta(n^2)$.

One disadvantage of Vavasis and Ye's method is that it requires explicit knowledge of $\bar{\chi}_A$ in order to determine a partition of the variables into layers used in the computation of the LLS step. This difficulty was remedied in a variant proposed by Megiddo, Mizuno, and Tsuchiya [11] which does not require explicit knowledge of the number $\bar{\chi}_A$. They observed that at most n types of partitions arise as $\bar{\chi}_A$ varies from 1 to ∞ , and that one of these can be used to compute the LLS step. Based on this idea, they developed a variant which computes the LLS steps for all these partitions and picks the one that yields the greatest duality gap reduction at the current iteration. Moreover, using the argument that, once the first LLS step is computed, the other ones can be cheaply computed by performing rank-one updates, they show that the overall complexity of their algorithm is exactly the same as Vavasis and Ye's algorithm.

Another approach that also remedies the above difficulty was proposed by Monteiro and Tsuchiya [19], who developed a variant of Vavasis and Ye's algorithm which has the same complexity as theirs and computes only one LLS step per iteration without any explicit knowledge of $\bar{\chi}_A$. The method is a P-C type algorithm like the one described in [13] except that at the predictor stage it takes a step along either the primal-dual AS step or the LLS step. In contrast to the LLS step used in the algorithm of Vavasis and Ye, the partition of variables used in the algorithm of [19] for computing the LLS step is constructed from the information provided by the AS direction and hence does not require any knowledge and/or guess on $\bar{\chi}_A$.

In this paper we present a new iteration-complexity bound for the Mizuno–Todd–Ye predictor-corrector (MTY P-C) primal-dual interior-point algorithm for LP. Using the notion of crossover events and a few other nontrivial ideas, we show that the MTY P-C algorithm started from a well-centered interior-feasible solution with duality gap $n\mu_0$ finds an interior-feasible solution with duality gap less than $n\eta$ in $\mathcal{O}(T(\mu_0/\eta) + n^{3.5} \log(\bar{\chi}_A^* + n))$ iterations, where

$$(1.3) \quad T(t) \equiv \min\{n^2 \log(\log t), \log t\} \quad \forall t > 0,$$

and $\bar{\chi}_A^*$ is a scaling invariant condition number associated with the matrix A . More specifically, $\bar{\chi}_A^*$ is the infimum of all the conditions numbers $\bar{\chi}_{AD}$, where D varies over the set of positive diagonal matrices. Thus, the derived iteration-complexity bound is scaling-invariant.

The iteration-complexity bound $\mathcal{O}(T(\mu_0/\eta) + n^{3.5} \log(\bar{\chi}_A^* + n))$ has the following intuitive geometric interpretation. The term $n^{3.5} \log(\bar{\chi}_A^* + n)$ corresponds to the complexity of tracing “the curved part” of the central trajectory, that is, the part of the trajectory where the step-size along the AS direction can be as small as $\Theta(1/\sqrt{n})$. This complexity depends only on n and A and does not depend on the initial duality gap and the tolerance $\eta > 0$. The other term, $T(\mu_0/\eta) = \min\{\mathcal{O}(n^2 \log(\log \mu_0/\eta)), \mathcal{O}(\log \mu_0/\eta)\}$, which corresponds to the complexity of tracing the “straight part” of the trajectory, is the minimum of two bounds. The first bound is obtained by showing that the duality gap reduces R -quadratically over $\mathcal{O}(n^2)$ disjoint sets of consecutive iterations of the algorithm. The second bound, which is independent of the dimension of the problem, is due to the fact that the duality gap along the straight part of the central trajectory is divided by a factor at least 2 at every iteration. To some extent, the second bound gives a plausible explanation to why the number of iterations performed by many interior-point algorithms grows very slowly with n .

Furthermore, using the above result, we show that the MTY algorithm endowed with a certain scaling-invariant finite termination procedure terminates in at most $\mathcal{O}(T(\mu_0/\eta^*) + n^{3.5} \log(\bar{\chi}_A^* + n))$ iterations, where η^* is a scaling invariant threshold number determined by the data (A, b, c) . In particular, our results imply that the MTY P-C algorithm solves (1.1) and (1.2) in $\mathcal{O}(\min\{n^2 \log L, L\} + n^{3.5} L_A)$ iterations under the Turing machine model. This contrasts well with the classical iteration-complexity bound for the MTY P-C algorithm, namely, $\mathcal{O}(\sqrt{n}L)$, which depends linearly on the input size L of the data (A, b, c) instead of the logarithm of L .

The organization of the paper is as follows. Section 2 consists of four subsections. In subsection 2.1, we review the notion of the primal-dual central path and its associated 2-norm neighborhoods. Subsection 2.2 introduces the condition number $\bar{\chi}_A$ of a matrix A and describes the properties of $\bar{\chi}_A$ that will be useful in our analysis. Subsection 2.3 reviews the MTY P-C algorithm and states the main result of this paper, which establishes a new scaling-invariant iteration-complexity bound for the

MTY P-C algorithm to find a near primal-dual optimal solution of (1.1) and (1.2). Subsection 2.4 describes a scaling-invariant finite termination procedure and gives an alternative scaling-invariant iteration-complexity bound for the MTY P-C algorithm to find a primal-dual optimal solution of (1.1) and (1.2). Section 3, which consists of four subsections, introduces some basic tools which are used in our convergence analysis. Subsection 3.1 discusses the notion of a crossover event. Subsection 3.2 describes the notion of an LLS direction and states a proximity result that gives sufficient conditions under which the AS direction can be well approximated by an LLS direction. Subsection 3.3 reviews, from a different perspective, an important result from Vavasis and Ye [30], which basically provides sufficient conditions for the occurrence of crossover events. Subsection 3.4 describes two ordered partitions of the set of variables which are frequently used in our analysis. Section 4 is dedicated to the proof of the main result stated in subsection 2.3. Section 5 deals with a few implications of our main result under the Turing machine model. Section 6 provides some concluding remarks.

The following notation is used throughout our paper. We denote the vector of all ones by e . Its dimension is always clear from the context. The symbols $\mathbb{R}^n$, $\mathbb{R}_+^n$, and $\mathbb{R}_{++}^n$ denote n -dimensional Euclidean space, the nonnegative orthant of $\mathbb{R}^n$, and the positive orthant of $\mathbb{R}^n$, respectively. The set of all $m \times n$ matrices with real entries is denoted by $\mathbb{R}^{m \times n}$. If J is a finite index set, then $|J|$ denotes its cardinality, that is, the number of elements of J . For $J \subseteq \{1, \dots, n\}$ and $w \in \mathbb{R}^n$, we let w_J denote the subvector $[w_i]_{i \in J}$; moreover, if E is an $m \times n$ matrix, then E_J denotes the $m \times |J|$ submatrix of E corresponding to J . For a vector $w \in \mathbb{R}^n$, we let $\max(w)$ and $\min(w)$ denote the largest and the smallest component of w , respectively; $\text{Diag}(w)$ denotes the diagonal matrix whose i th diagonal element is w_i for $i = 1, \dots, n$; and w^{-1} denotes the vector $[\text{Diag}(w)]^{-1}e$ whenever it is well defined. For two vectors $u, v \in \mathbb{R}^n$, uv denotes their Hadamard product, i.e., the vector in $\mathbb{R}^n$ whose i th component is $u_i v_i$. The Euclidean norm, the 1-norm, and the ∞ -norm are denoted by $\|\cdot\|$, $\|\cdot\|_1$, and $\|\cdot\|_\infty$, respectively. For a matrix E , $\text{Im}(E)$ denotes the subspace generated by the columns of E , and $\text{Ker}(E)$ denotes the subspace orthogonal to the rows of E . The superscript T denotes transpose.

2. Problem and primal-dual P-C interior-point algorithms. In this section we review the MTY P-C algorithm [13] for solving the pair of LP problems (1.1) and (1.2). We also present our main convergence result, which establishes a new polynomial iteration-complexity bound for this algorithm, namely, $\mathcal{O}(T(\mu_0/\eta) + n^{3.5} \log(\bar{\chi}_A^* + n))$, where $T(\mu_0/\eta)$ is defined in (1.3) and $\bar{\chi}_A^*$ is a certain scaling-invariant condition number associated with the matrix A , $n\mu_0$ is the initial duality gap, and $n\eta$ is the required upper bound on the duality gap of the final iterate.

2.1. The problem, the central path, and its 2-norm neighborhood. In this subsection we state our assumptions and describe the primal-dual central path and its corresponding 2-norm neighborhoods.

Given $A \in \mathbb{R}^{m \times n}$, $c \in \mathbb{R}^n$, and $b \in \mathbb{R}^m$, consider the pairs of linear programs (1.1) and (1.2), where $x \in \mathbb{R}^n$ and $(y, s) \in \mathbb{R}^m \times \mathbb{R}^n$ are their respective variables. The set of strictly feasible solutions for these problems are

$$\begin{aligned} \mathcal{P}^{++} &\equiv \{x \in \mathbb{R}^n : Ax = b, x > 0\}, \\ \mathcal{D}^{++} &\equiv \{(y, s) \in \mathbb{R}^{m \times n} : A^T y + s = c, s > 0\}. \end{aligned}$$

Throughout the paper we make the following assumptions on the pair of problems

(1.1) and (1.2).

A.1. $\mathcal{P}^{++}$ and $\mathcal{D}^{++}$ are nonempty.

A.2. The rows of A are linearly independent.

Under the above assumptions, it is well known that for any $\nu > 0$ the system

$$(2.1) \quad xs = \nu e,$$

$$(2.2) \quad Ax = b, \quad x > 0,$$

$$(2.3) \quad A^T y + s = c, \quad s > 0,$$

has a unique solution (x, y, s) , which we denote by $(x(\nu), y(\nu), s(\nu))$. The central path is the set consisting of all these solutions as ν varies in $(0, \infty)$. As ν converges to zero, the path $(x(\nu), y(\nu), s(\nu))$ converges to a primal-dual optimal solution (x^*, y^*, s^*) for problems (1.1) and (1.2). Given a point $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$, its duality gap and its normalized duality gap are defined as $x^T s$ and $\mu = \mu(x, s) \equiv x^T s / n$, respectively, and the point $(x(\mu), y(\mu), s(\mu))$ is said to be the central point associated with w . Note that $(x(\mu), y(\mu), s(\mu))$ also has normalized duality gap μ . We define the proximity measure of a point $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ with respect to the central path by

$$\phi(w) = \|xs/\mu - e\|.$$

Clearly, $\phi(w) = 0$ if and only if $w = (x(\mu), y(\mu), s(\mu))$, or equivalently, w coincides with its associated central point. The 2-norm neighborhood of the central path with opening $\beta > 0$ is defined as

$$\mathcal{N}(\beta) \equiv \{w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++} : \phi(w) \leq \beta\}.$$

Finally, for any point $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$, we define

$$(2.4) \quad \delta(w) \equiv s^{1/2} x^{-1/2} \in \mathbb{R}^n.$$

2.2. Condition number. In this subsection we define the condition number $\bar{\chi}_A$ associated with the constraint matrix A and state the properties of $\bar{\chi}_A$, which will play an important role in our analysis.

Let $\mathcal{D}$ denote the set of all positive definite $n \times n$ diagonal matrices and define

$$(2.5) \quad \begin{aligned} \bar{\chi}_A &\equiv \sup\{\|A^T(A\tilde{D}A^T)^{-1}A\tilde{D}\| : \tilde{D} \in \mathcal{D}\} \\ &= \sup\left\{\frac{\|A^T y\|}{\|c\|} : y = \operatorname{argmin}_{\tilde{y} \in \mathbb{R}^n} \|\tilde{D}^{1/2}(A^T \tilde{y} - c)\| \text{ for some } 0 \neq c \in \mathbb{R}^n \text{ and } \tilde{D} \in \mathcal{D}\right\}. \end{aligned}$$

The parameter $\bar{\chi}_A$ plays a fundamental role in the complexity analysis of algorithms for LP and LLS problems (see [30] and references therein). Its finiteness was first established by Dikin [1]. Other authors have also given alternative derivations of the finiteness of $\bar{\chi}_A$ (see, for example, Stewart [21], Todd [24], and Vanderbei and Lagarias [29]).

We summarize in the next proposition a few important facts about the parameter $\bar{\chi}_A$.

PROPOSITION 2.1. *Let $A \in \mathbb{R}^{m \times n}$ with full row rank be given. Then, the following statements hold:*

- (a) $\bar{\chi}_{HA} = \bar{\chi}_A$ for any nonsingular matrix $H \in \mathbb{R}^{m \times m}$;
- (b) $\bar{\chi}_A = \max\{\|G^{-1}A\| : G \in \mathcal{G}\}$, where $\mathcal{G}$ denotes the set of all $m \times m$ nonsingular submatrices of A ;
- (c) if the entries of A are all integers, then $\bar{\chi}_A$ is bounded by $2^{\mathcal{O}(L_A)}$, where L_A is the input bit length of A ;
- (d) $\bar{\chi}_A = \bar{\chi}_F$ for any $F \in \mathbb{R}^{(n-m) \times n}$ such that $\text{Ker}(A) = \text{Im}(F^T)$.

Proof. Statement (a) readily follows from the definition (2.5). The inequality $\bar{\chi}_A \geq \max\{\|G^{-1}A\| : G \in \mathcal{G}\}$ is established in Lemma 3 of [30] while the proof of the reverse inequality is given in [24] (see also Theorem 1 of [25]). Hence, (b) holds. The proof of (c) can be found in Lemma 24 of [30]. A proof of (d) can be found in [4]. $\square$

We now state a Hoffman-type result for a system of linear equalities whose proof can be found in Lemma 2.3 of [19].

LEMMA 2.2. *Let $A \in \mathbb{R}^{m \times n}$ with full row rank be given and let $(\mathcal{K}, \mathcal{L})$ be an arbitrary bipartition of the index set $\{1, \dots, n\}$. Assume that $\bar{w} \in \mathbb{R}^{|\mathcal{L}|}$ is an arbitrary vector such that the system $A_{\mathcal{K}}u = A_{\mathcal{L}}\bar{w}$ is feasible. Then, this system has a feasible solution $\bar{u}$ such that $\|\bar{u}\| \leq \bar{\chi}_A \|\bar{w}\|$.*

2.3. P-C step and its properties. In this subsection we review the well-known MTY P-C algorithm [13] and its main properties. We also state the main result of this paper, which establishes a new scaling-invariant iteration-complexity bound of the MTY P-C algorithm for finding a near primal-dual optimal solution of (1.1) and (1.2).

Each iteration of the MTY P-C algorithm consists of two steps, namely, the predictor (or AS) step and the corrector (or centrality) step. The search direction used by both steps at a given point in $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ is the unique solution of the following linear system of equations:

$$(2.6) \quad \begin{aligned} S\Delta x + X\Delta s &= \sigma\mu e - xs, \\ A\Delta x &= 0, \\ A^T\Delta y + \Delta s &= 0, \end{aligned}$$

where $\mu = \mu(x, s)$ and $\sigma \in \mathbb{R}$ is a prespecified parameter, commonly referred to as the centrality parameter. When $\sigma = 0$, we denote the solution of (2.6) by $(\Delta x^a, \Delta y^a, \Delta s^a)$ and refer to it as the (primal-dual) AS direction at w ; it is the direction used in the predictor step. When $\sigma = 1$, we denote the solution of (2.6) by $(\Delta x^c, \Delta y^c, \Delta s^c)$ and refer to it as the centrality direction at w ; it is the direction used in the corrector step.

To describe an entire iteration of the MTY P-C algorithm, suppose that a constant $\beta \in (0, 1/4]$ is given. Given a point $w = (x, y, s) \in \mathcal{N}(\beta)$, this algorithm generates the next point $w^+ = (x^+, y^+, s^+) \in \mathcal{N}(\beta)$ as follows. It first moves along the direction $(\Delta x^a, \Delta y^a, \Delta s^a)$ until it hits the boundary of the enlarged neighborhood $\mathcal{N}(2\beta)$. More specifically, it computes the point $w^a = (x^a, y^a, s^a) \equiv (x, y, s) + \alpha_a(\Delta x^a, \Delta y^a, \Delta s^a)$, where

$$(2.7) \quad \alpha_a \equiv \sup\{\alpha \in [0, 1] : (x, y, s) + \alpha'(\Delta x^a, \Delta y^a, \Delta s^a) \in \mathcal{N}(2\beta) \ \forall \alpha' \in [0, \alpha]\}.$$

Next, the point $w^+ = (x^+, y^+, s^+)$ inside the smaller neighborhood $\mathcal{N}(\beta)$ is generated by taking a unit step along the centrality direction $(\Delta x^c, \Delta y^c, \Delta s^c)$ at the point w^a , that is, $(x^+, y^+, s^+) \equiv (x^a, y^a, s^a) + (\Delta x^c, \Delta y^c, \Delta s^c) \in \mathcal{N}(\beta)$. Starting from a point $w^0 \in \mathcal{N}(\beta)$ and successively performing iterations as described above, the MTY P-C

algorithm generates a sequence of points $\{w^k\} \subseteq \mathcal{N}(\beta)$, which converges to the primal-dual optimal face of problems (1.1) and (1.2).

The convergence analysis of the sequence $\{w^k\}$, as well as the finite termination of the MTY P-C algorithm, has been the subject of study of several papers. In what follows, we review the properties of the P-C iteration, which yields the classical polynomial iteration-complexity bound for the MTY P-C algorithm. We also discuss alternative properties of the P-C iteration, which will be used in our analysis to derive a new polynomial iteration-complexity bound for the MTY P-C algorithm. A scaling-invariant finite termination procedure for the MTY P-C algorithm and its relationship with another well-known finite termination procedure will be discussed in subsection 2.4. A detailed proof of the next two propositions can be found, for example, in [34].

PROPOSITION 2.3 (predictor step). *Suppose that $w = (x, y, s) \in \mathcal{N}(\beta)$ for some constant $\beta \in (0, 1/4]$. Let $\Delta w^a = (\Delta x^a, \Delta y^a, \Delta s^a)$ denote the AS direction at w and let α_a be the step-size computed according to (2.7). Then the following statements hold:*

- (a) *the point $w + \alpha \Delta w^a$ has normalized duality gap $\mu(\alpha) = (1 - \alpha)\mu$ for all $\alpha \in \mathbb{R}$;*
- (b) *$\alpha_a \geq \max\{1 - \chi/\beta, \sqrt{\beta/n}\}$, where $\chi \equiv \|\Delta x^a \Delta s^a\|/\mu$.*

Proof. It is well known (see, for example, section 4.5.1 of [34]) that (a) holds, $\chi \leq n/2$, and

$$\alpha_a \geq \frac{2}{1 + \sqrt{1 + 4\chi/\beta}}.$$

Using these two inequalities we see, after a simple verification, that (b) holds. $\square$

PROPOSITION 2.4 (corrector step). *Suppose that $w = (x, y, s) \in \mathcal{N}(2\beta)$ for some constant $\beta \in (0, 1/4]$ and let $(\Delta x^c, \Delta y^c, \Delta s^c)$ denote the corrector step at w . Then $w + \Delta w^c \in \mathcal{N}(\beta)$. Moreover, the (normalized) duality gap of $w + \Delta w^c$ is the same as that of w .*

For a search direction $\Delta w = (\Delta x, \Delta y, \Delta s)$ at a point $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$, the quantity

$$(2.8) \quad (Rx, Rs) \equiv \left(\frac{\delta(x + \Delta x)}{\sqrt{\mu}}, \frac{\delta^{-1}(s + \Delta s)}{\sqrt{\mu}} \right) = \left(\frac{x^{1/2}s^{1/2} + \delta\Delta x}{\sqrt{\mu}}, \frac{x^{1/2}s^{1/2} + \delta^{-1}\Delta s}{\sqrt{\mu}} \right),$$

where $\delta \equiv \delta(w)$, appears quite often in our analysis. We refer to it as the *residual* of Δw . Throughout this paper, we denote the residual of the AS direction $\Delta w^a = (\Delta x^a, \Delta y^a, \Delta s^a)$ at w as $(Rx^a(w), Rs^a(w))$. Note that if $(Rx^a, Rs^a) \equiv (Rx^a(w), Rs^a(w))$, then

$$(2.9) \quad Rx^a = -\frac{1}{\sqrt{\mu}}\delta^{-1}\Delta s^a, \quad Rs^a = -\frac{1}{\sqrt{\mu}}\delta\Delta x^a,$$

and

$$(2.10) \quad Rx^a + Rs^a = \frac{x^{1/2}s^{1/2}}{\sqrt{\mu}},$$

due to the fact that $(\Delta x^a, \Delta y^a, \Delta s^a)$ satisfies the first equation in (2.6) with $\sigma = 0$. The following quantity plays an important role in our analysis:

$$(2.11) \quad \varepsilon_\infty^a(w) \equiv \max_i \{\min\{|Rx_i^a(w)|, |Rs_i^a(w)|\}\}.$$

In terms of this quantity, we have the following bound on the reduction on the duality gap during an iteration of the MTY P-C algorithm.

LEMMA 2.5. *Suppose that $w \in \mathcal{N}(\beta)$ for some constant $\beta \in (0, 1/4]$ and let w^+ be the point obtained after a single iteration of the MTY P-C algorithm with base point w . Then, $w^+ \in \mathcal{N}(\beta)$ and*

$$(2.12) \quad \mu(w^+) \leq \min \left\{ 1 - \sqrt{\frac{\beta}{n}}, \frac{\sqrt{n}\varepsilon_\infty^a(w)}{\beta} \right\} \mu(w).$$

Proof. Using (2.6) and (2.9), we easily see that $(Rx^a)^T Rs^a = 0$. Using this observation together with (2.10), we easily see that $\max\{\|Rx^a\|, \|Rs^a\|\} \leq \sqrt{n}$. The result now immediately follows from this conclusion, relation (2.11), and Propositions 2.3 and 2.4. $\square$

We end this section by stating the main result of this paper. This result establishes a new iteration-complexity bound for the MTY P-C algorithm.

THEOREM 2.6. *Given a termination tolerance $\eta > 0$ for the normalized duality gap and an initial point $w^0 \in \mathcal{N}(\beta)$ with $\beta \in (0, 1/4]$, the MTY P-C algorithm generates an iterate $w^k \in \mathcal{N}(\beta)$ satisfying $\mu(w^k) \leq \eta$ in at most*

$$(2.13) \quad \mathcal{O}(\min\{\sqrt{n}\log(\mu_0/\eta), T(\mu_0/\eta) + n^{3.5}\log(\bar{\chi}_A^* + n)\})$$

iterations, where $\mu_0 \equiv \mu(w^0)$, $T(\cdot)$ is defined in (1.3), and

$$\bar{\chi}_A^* \equiv \inf\{\bar{\chi}_{AD} : D \in \mathcal{D}\}.$$

A few observations are in order at this point. First, the first bound in (2.13) is the classical one derived in [13] (see also [8, 15, 16, 17]), and follows as an immediate consequence of the first bound on the duality gap reduction obtained in Lemma 2.5. The second bound in (2.13) is the one which will be established in this paper. Observe that, in contrast to the classical iteration-complexity bound which is proportional to $\log(\mu_0/\eta)$, the new bound depends linearly on $\log(\log(\mu_0/\eta))$. Second, note that the MTY P-C algorithm is scaling-invariant; i.e., if the change of variables $(x, y, s) = (D\tilde{x}, \tilde{y}, D^{-1}\tilde{s})$ for some $D \in \mathcal{D}$ is performed on the pair of problems (1.1) and (1.2) and the MTY P-C algorithm is applied to the new dual pair of scaled problems, then the sequence of iterates $\{\tilde{w}^k\}$ generated satisfies $(x^k, y^k, s^k) = (D\tilde{x}^k, \tilde{y}^k, D^{-1}\tilde{s}^k)$ for all $k \geq 1$ as long as the initial iterate $\tilde{w}^0 \in \mathcal{N}(\beta)$ in the $\tilde{w}$ -space satisfies $(x^0, y^0, s^0) = (D\tilde{x}^0, \tilde{y}^0, D^{-1}\tilde{s}^0)$. For this reason, the MTY P-C algorithm should have an iteration-complexity bound which does not depend on the scaled space where the sequence of iterates is generated. Note that the iteration-complexity bound (2.13) has this property since μ_0 , $\bar{\chi}_A^*$, and the inequality $\mu(w^k) \leq \eta$ are all scaling-invariant. Third, to establish the iteration-complexity bound stated in Theorem 2.6, it is sufficient to establish that, in the scaled space, the iteration-complexity bound is

$$(2.14) \quad \mathcal{O}(\min\{\sqrt{n}\log(\mu_0/\eta), T(\mu_0/\eta) + n^{3.5}\log(\bar{\chi}_{AD} + n)\}).$$

Indeed, since the number of iterations of the MTY P-C algorithm does not depend on the underlying scaled space, it follows that this number is majorized by the infimum of (2.14) over all $D \in \mathcal{D}$, i.e., by the bound (2.13). Moreover, without loss of generality, we will consider the MTY P-C algorithm applied to (1.1) and (1.2) without any scaling and will establish the iteration-complexity bound (2.14) with $D = I$.

Vavasis and Ye [30, 31] developed a primal-dual path-following interior-point algorithm which solves the LP pair (1.1) and (1.2) in $O(n^{3.5} \log(\bar{\chi}_A + n))$ iterations. In contrast to the other standard primal-dual path-following interior-point methods developed in the literature, including the MTY P-C algorithm presented in this subsection, Vavasis and Ye's algorithm uses from time to time a crucial step, namely, the LLS step, which unfortunately is not scaling-invariant. Hence, the quantity $\bar{\chi}_A$ in its complexity bound cannot be replaced by $\bar{\chi}_A^*$. For this reason, a definitive comparison of the iteration-complexity bound corresponding to Vavasis and Ye's algorithm and the one stated in Theorem 2.6 is not apparent. While the one of Theorem 2.6 contains the extra term $T(\mu_0/\eta)$, which is not present in Vavasis and Ye's iteration-complexity bound, the iteration-complexity bound of Theorem 2.6 depends on $\bar{\chi}_A^*$ instead of the larger quantity $\bar{\chi}_A$. The following simple example shows that the difference between these two quantities can be substantial. It exhibits a family of matrices A such that the ratio $\bar{\chi}_A^*/\bar{\chi}_A$ converges to 0.

Example. For $\varepsilon > 0$, define $A(\varepsilon) := (\varepsilon I | e_1)$, where I is the $m \times m$ identity matrix and e_1 is the first unit vector. It is easy to see that $\bar{\chi}_{A(\varepsilon)} = \Theta(\varepsilon^{-1})$ and that $\chi_{A(\varepsilon)}^* = O(1)$ for every $\varepsilon > 0$.

2.4. A scaling-invariant finite termination procedure. In this section, we describe a scaling-invariant finite termination procedure which, used in conjunction with the MTY P-C algorithm, allows one to find an exact primal-dual optimal solution of (1.1) and (1.2). We also derive an alternative scaling-invariant iteration-complexity bound for the MTY P-C algorithm to find an exact primal-dual optimal solution of (1.1) and (1.2).

The finite termination procedure described in this subsection is similar to the one described in Mehrotra and Ye [12] (see also [34]) except that ours uses a scaling-invariant scheme for guessing the optimal partition (B_*, N_*) associated with the pair of LP problems (1.1) and (1.2). (Recall that, by definition, $B_* \equiv \{i : x_i > 0 \text{ for some } x \in \text{opt}(1.1)\}$ and $N_* \equiv \{i : s_i > 0 \text{ for some } (y, s) \in \text{opt}(1.2)\}$, where $\text{opt}(\cdot)$ denotes the set of optimal solutions of problem $(\cdot)$.) Namely, given a point $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$, we define the AS-bipartition $(B(w), N(w))$ at w as

$$(2.15) \quad B(w) \equiv \{i : |Rs_i^a(w)| \leq |Rx_i^a(w)|\}, \quad N(w) \equiv \{i : |Rs_i^a(w)| > |Rx_i^a(w)|\}.$$

We now state the following finite termination procedure which uses the AS-bipartition for guessing the optimal partition. Note that this partition is a primal-dual symmetric variant of the indicators discussed in [2].

FINITE TERMINATION (FT) PROCEDURE.

Given $\gamma \in (0, 1)$ and $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ such that $xs \geq \gamma\mu(w)$.

- (1) Find the AS-bipartition $(B, N) = (B(w), N(w))$ at w .
- (2) Solve the following projection problems:

$$(2.16) \quad x^* \equiv \operatorname{argmin}_{\tilde{x}} \{\|\delta(x - \tilde{x})\|^2 : A\tilde{x} = b, \tilde{x}_N = 0\},$$

$$(2.17) \quad (y^*, s^*) \equiv \operatorname{argmin}_{(\tilde{y}, \tilde{s})} \{\|\delta^{-1}(s - \tilde{s})\|^2 : A^T \tilde{y} + \tilde{s} = c, \tilde{s}_B = 0\},$$

where $\delta \equiv \delta(w)$.

- (3) If $x_B^* > 0$ and $s_N^* > 0$, then $w^* = (x^*, y^*, s^*)$ is optimal; output w^* and declare success. Otherwise, exit the procedure and declare failure.

Our goal now will be to show that the FT procedure always finds a primal-dual optimal solution of (1.1) and (1.2) whenever $\mu(w)$ is less than a certain positive threshold constant defined in terms of additional condition measures associated with the pair of LP problems (1.1) and (1.2). We start by defining these condition measures. The first two condition measures are given by

$$\begin{aligned}\xi^P(A, b, c) &\equiv \min_{i \in B_*} \max\{\bar{x}_i : \bar{x} \in \text{opt}(1.1)\}, \\ \xi^D(A, b, c) &\equiv \min_{i \in N_*} \max\{\bar{s}_i : (\bar{y}, \bar{s}) \in \text{opt}(1.2)\}.\end{aligned}$$

The third condition measure is defined as

$$(2.18) \quad \zeta(A, (B_*, N_*)) \equiv \max\{\zeta_1, \zeta_2\},$$

where

$$\begin{aligned}\zeta_1 &\equiv \max_{d_{N_*} \neq 0} \left\{ \min \left\{ \frac{\|d_{B_*}\|}{\|d_{N_*}\|} : d = (d_{B_*}, d_{N_*}) \in \text{Ker}(A) \right\} \right\}, \\ \zeta_2 &\equiv \max_{d_{B_*} \neq 0} \left\{ \min \left\{ \frac{\|d_{N_*}\|}{\|d_{B_*}\|} : d = (d_{B_*}, d_{N_*}) \in \text{Im}(A^T) \right\} \right\}.\end{aligned}$$

Using Lemma 2.2 and Proposition 2.1(d), it is easy to see that $\zeta(A, (B_*, N_*)) \leq \bar{\chi}_A$.

The following result states some well-known estimates on the size of the components of a point $w \in \mathcal{P}^{++} \times \mathcal{D}^{++}$.

LEMMA 2.7. *Let $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ be given and define $\mu \equiv \mu(w)$, $\delta \equiv \delta(w)$, $\xi^P \equiv \xi^P(A, b, c)$, and $\xi^D \equiv \xi^D(A, b, c)$. Then*

$$(2.19) \quad \max(x_{N_*}) \leq \frac{n\mu}{\xi^D}, \quad \max(s_{B_*}) \leq \frac{n\mu}{\xi^P}.$$

If we further assume that $xs \geq \gamma\mu e$ for some $\gamma \in (0, 1)$, then

$$(2.20) \quad \min(x_{B_*}) \geq \frac{\gamma\xi^D}{n}, \quad \min(s_{N_*}) \geq \frac{\gamma\xi^P}{n},$$

$$(2.21) \quad \|\delta_{B_*}\|_\infty \|\delta_{N_*}^{-1}\|_\infty \leq \frac{n^2\mu}{\gamma\xi^P\xi^D}.$$

Proof. The inequality (2.19) follows immediately from the definitions of $\xi^P(A, b, c)$ and $\xi^D(A, b, c)$ and the following identities: $x^T s = n\mu$, $\bar{x}^T \bar{s} = 0$, and $(x - \bar{x})^T (s - \bar{s}) = 0$ for all $\bar{w} = (\bar{x}, \bar{y}, \bar{s}) \in \text{opt}(1.1) \times \text{opt}(1.2)$ (see Ye [33] for details). Moreover, using the assumption that $xs \geq \gamma\mu e$ and (2.19), we obtain $s_i \geq \gamma\mu/x_i \geq \gamma\xi^D/n$ for every $i \in N_*$ and $x_i \geq \gamma\mu/s_i \geq \gamma\xi^P/n$ for every $i \in B_*$, from which (2.20) follows. Inequality (2.21) follows immediately from (2.4), (2.19), and (2.20). $\square$

The following result shows not only that the AS-bipartition $(B(w), N(w))$ is a correct guess for the optimal partition (B_*, N_*) but also that the FT procedure yields a primal-dual optimal solution of (1.1) and (1.2) whenever the duality gap $\mu(w)$ is less than a suitable threshold value defined in terms of the condition measures ξ and ζ introduced above, the dimension n , and the degree of centrality γ of w .

LEMMA 2.8. *Suppose that $\gamma \in (0, 1)$ and $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ are such that*

$$(2.22) \quad xs \geq \gamma\mu \quad \text{and} \quad \mu < \frac{\gamma^{1.5}\xi^P\xi^D}{2\zeta n^{2.5}},$$

where $\mu = \mu(w)$, $\zeta \equiv \zeta(A, (B_*, N_*))$, $\xi^P \equiv \xi^P(A, b, c)$, and $\xi^D \equiv \xi^D(A, b, c)$. Then we have that

- (a) $(B(w), N(w)) = (B_*, N_*)$;
- (b) the FT procedure yields a strictly complementary primal-dual optimal solution, that is, a triple $w^* = (x^*, y^*, s^*) \in \text{opt}(1.1) \times \text{opt}(1.2)$ such that $x^* + s^* > 0$.

Proof. Suppose $\gamma \in (0, 1)$ and $w \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ are such that (2.22) holds. We will first prove (a). Let $\Delta w^a = (\Delta x^a, \Delta y^a, \Delta s^a)$ denote the AS direction at w . It is well known that

$$(2.23) \quad \max\{\|\delta \Delta x^a\|, \|\delta^{-1} \Delta s^a\|\} \leq \sqrt{n\mu},$$

where $\delta \equiv \delta(w)$. Moreover, it is easy to see that

$$\Delta x^a = \operatorname{argmin} \left\{ \tilde{s}^T \Delta x + \frac{1}{2} \|\delta \Delta x\|^2 : A \Delta x = 0 \right\}$$

for every $\tilde{s} \in c + \operatorname{Im}(A^T)$. Now fix some $(\bar{y}, \bar{s}) \in \text{opt}(1.2)$ and let $\tilde{s} = \bar{s}$ in the above optimization problem. Splitting the variable Δx^a according to the partition (B_*, N_*) , noting that $\bar{s}_{B_*} = 0$, and fixing the component Δx_{N_*} to be $\Delta x_{N_*}^a$, we conclude that

$$(2.24) \quad \Delta x_{B_*}^a = \operatorname{argmin}_{\Delta x_{B_*}} \left\{ \frac{1}{2} \|\delta_{B_*} \Delta x_{B_*}\|^2 : A_{B_*} \Delta x_{B_*} = -A_{N_*} \Delta x_{N_*}^a \right\}.$$

Now, let $\tilde{\Delta x}_{B_*}$ denote the minimum norm solution of the system $A_{B_*} \Delta x_{B_*} = -A_{N_*} \Delta x_{N_*}^a$. Using (2.21), (2.23), (2.24), and the definition of $\zeta = \zeta(A, (B_*, N_*))$, we obtain

$$\begin{aligned} \|\delta_{B_*} \Delta x_{B_*}^a\| &\leq \|\delta_{B_*} \tilde{\Delta x}_{B_*}\| \leq \|\delta_{B_*}\|_\infty \|\tilde{\Delta x}_{B_*}\| \leq \zeta \|\delta_{B_*}\|_\infty \|\Delta x_{N_*}^a\| \\ &\leq \zeta \|\delta_{B_*}\|_\infty \|\delta_{N_*}^{-1}\|_\infty \|\delta_{N_*} \Delta x_{N_*}^a\| \leq \sqrt{n\mu} \zeta \|\delta_{B_*}\|_\infty \|\delta_{N_*}^{-1}\|_\infty \\ &\leq \frac{n^{2.5} \zeta \mu^{1.5}}{\gamma \xi^P \xi^D} < \frac{\sqrt{\gamma\mu}}{2}, \end{aligned}$$

where the last inequality is due to (2.22). The last inequality, together with (2.22), implies that for every $i \in B_*$, we have

$$\left| \frac{\delta_i^{-1} \Delta s_i^a}{\sqrt{\mu}} \right| = \left| \frac{x_i^{1/2} s_i^{1/2}}{\sqrt{\mu}} + \frac{\delta_i \Delta x_i^a}{\sqrt{\mu}} \right| \geq \frac{x_i^{1/2} s_i^{1/2}}{\sqrt{\mu}} - \left| \frac{\delta_i \Delta x_i^a}{\sqrt{\mu}} \right| > \sqrt{\gamma} - \frac{\sqrt{\gamma}}{2} = \frac{\sqrt{\gamma}}{2} \geq \left| \frac{\delta_i \Delta x_i^a}{\sqrt{\mu}} \right|,$$

or equivalently, $|R x_i^a(w)| > |R s_i^a(w)|$ for every $i \in B_*$. Hence, $B_* \subseteq B(w)$ in view of (2.15). Similarly, we can show that $N_* \subseteq N(w)$. Therefore, $(B(w), N(w)) = (B_*, N_*)$.

We now prove (b). Let (x^*, y^*, s^*) denote the point determined by (2.16) and (2.17). Observe that $x_{B_*} - x_{B_*}^*$ is a feasible solution of the system $A_{B_*} \Delta x_{B_*} = -A_{N_*} x_{N_*}$. Now let $\widehat{\Delta x}_{B_*}$ denote the minimum norm solution of this system. Using (2.21), (2.22), (2.16), and the definition of $\zeta = \zeta(A, (B_*, N_*))$, we obtain

$$\begin{aligned} &\|x_{B_*}^{-1}(x_{B_*} - x_{B_*}^*)\| \\ &\leq \left\| x_{B_*}^{-1/2} s_{B_*}^{-1/2} \right\|_\infty \|\delta_{B_*} (x_{B_*} - x_{B_*}^*)\| \leq \frac{1}{\sqrt{\gamma\mu}} \|\delta_{B_*} (x_{B_*} - x_{B_*}^*)\| \\ &\leq \frac{1}{\sqrt{\gamma\mu}} \|\delta_{B_*} \widehat{\Delta x}_{B_*}\| \leq \frac{1}{\sqrt{\gamma\mu}} \|\delta_{B_*}\|_\infty \|\widehat{\Delta x}_{B_*}\| \leq \frac{\zeta}{\sqrt{\gamma\mu}} \|\delta_{B_*}\|_\infty \|x_{N_*}\| \\ &\leq \frac{\zeta}{\sqrt{\gamma\mu}} \|\delta_{B_*}\|_\infty \|\delta_{N_*}^{-1}\|_\infty \|\delta_{N_*} x_{N_*}\| = \frac{\zeta}{\sqrt{\gamma\mu}} \|\delta_{B_*}\|_\infty \|\delta_{N_*}^{-1}\|_\infty \left\| x_{N_*}^{1/2} s_{N_*}^{1/2} \right\| \\ &\leq \frac{\zeta \sqrt{n}}{\sqrt{\gamma}} \|\delta_{B_*}\|_\infty \|\delta_{N_*}^{-1}\|_\infty \leq \frac{n^{2.5} \zeta \mu}{\gamma^{1.5} \xi^P \xi^D} < \frac{1}{2}. \end{aligned}$$

The above inequality clearly implies that $x_{B_*}^* > 0$. In a similar way, we can prove that $s_{N_*}^* > 0$. Hence, (b) follows. $\square$

The next result gives an iteration-complexity bound for the MTY P-C algorithm, used in conjunction with the FT procedure to find a primal-dual optimal solution of (1.1) and (1.2).

THEOREM 2.9. *Suppose that $w^0 \in \mathcal{N}(\beta)$ with $\beta \in (0, 1/4]$ is given. Then, the version of the MTY P-C algorithm in which the FT procedure is invoked at every iterate w^k started from w^0 finds a primal-dual strictly complementary optimal solution w^* in at most*

$$(2.25) \quad \mathcal{O}(\min\{\sqrt{n} \log(n\mu_0/\eta_*), T(\mu_0/\eta^*) + n^{3.5} \log(\bar{\chi}_A^* + n)\})$$

iterations, where $\mu_0 = \mu(w^0)$, the function $T(\cdot)$ is defined in (1.3), and

$$\eta_* \equiv \sup \left\{ \frac{\xi^P(AD, b, Dc) \xi^D(AD, b, Dc)}{\zeta(AD, (B_*, N_*))} : D \in \mathcal{D} \right\}.$$

Proof. This result follows immediately from Theorem 2.6, Lemma 2.8, and the scaling-invariance of the algorithm under consideration. $\square$

A few implications of Theorem 2.9 under the Turing machine model will be discussed in section 5. For now, we mention that the complexity bound (2.25) remains invariant not only under the transformation that changes the LP data from (A, c, b) to (AD, Dc, b) for some $D \in \mathcal{D}$, but also under the transformation that changes (A, c, b) to $(A, \alpha c, \beta b)$ for some fixed positive scalars α and β .

3. Basic tools. In this section we introduce the basic tools that will be used in the proof of Theorem 2.6. The analysis relies heavily on the notion of crossover events due to Vavasis and Ye [30]. In subsection 3.1, we give a definition of crossover event, which is slightly different than the one introduced in [30], and then discuss some of its properties. In subsection 3.2, we describe the notion of an LLS direction introduced in [30] and then state a proximity result that gives sufficient conditions under which the AS direction can be well approximated by an LLS direction. Subsection 3.3 reviews from a different perspective an important result from [30], namely, Lemma 17, that essentially guarantees the occurrence of crossover events. Since this result is stated in terms of the residual of an LLS step, the use of the proximity result of subsection 3.2 allows us to obtain a similar result stated in terms of the residual of the AS direction. In subsection 3.4, we introduce two ordered partitions of the set of variables, which play an important role in our analysis.

3.1. Crossover events. In this subsection we discuss the important notion of a crossover event developed by Vavasis and Ye [30].

DEFINITION. *For two indices $i, j \in \{1, \dots, n\}$ and a constant $\mathcal{C} \geq 1$, a $\mathcal{C}$ -crossover event for the pair (i, j) is said to occur on the interval $(\nu', \nu]$ if*

$$(3.1) \quad \begin{aligned} &\text{there exists } \nu_0 \in (\nu', \nu] \text{ such that } \frac{s_j(\nu_0)}{s_i(\nu_0)} \leq \mathcal{C}, \\ &\text{and } \frac{s_j(\tilde{\nu})}{s_i(\tilde{\nu})} > \mathcal{C} \ \forall \ \tilde{\nu} \leq \nu'. \end{aligned}$$

Moreover, the interval $(\nu', \nu]$ is said to contain a $\mathcal{C}$ -crossover event if (3.1) holds for some pair (i, j) .

Hence, the notion of a crossover event is independent of any algorithm and is a property of the central path only. Note that in view of (2.1), condition (3.1) can be

reformulated into an equivalent condition involving only the primal variable. For our purposes, we will use only (3.1).

We have the following simple but crucial result about crossover events.

PROPOSITION 3.1. *Let $\mathcal{C} \geq 1$ be a given constant. There can be at most $n(n-1)/2$ disjoint intervals of the form $(\nu', \nu]$ containing $\mathcal{C}$ -crossover events.*

The notion of $\mathcal{C}$ -crossover events can be used to define the notion of $\mathcal{C}$ -crossover events between two iterates of the MTY P-C algorithm as follows. We say that a $\mathcal{C}$ -crossover event occurs between two iterates w^k and w^l , $k < l$, generated by the MTY P-C algorithm if the interval $(\mu(w^l), \mu(w^k)]$ contains a $\mathcal{C}$ -crossover event. Note that in view of Proposition 3.1, there can be at most $n(n-1)/2$ disjoint intervals of this type. We will show in the remainder of this paper that there exists a constant $\mathcal{C} \geq 1$ with the following property: for any index k , there exists an index $l > k$ with the property $l - k = \mathcal{O}(\log(\log(\mu(w^0)/\eta)) + n^{1.5} \log(\bar{\chi}_A + n))$ and, if the MTY P-C algorithm does not terminate before or at the l th iteration, then a $\mathcal{C}$ -crossover event must occur between the iterates w^k and w^l . Proposition 3.1 and a simple argument then show that the MTY P-C algorithm must terminate within $\mathcal{O}(n^2 \log(\log(\mu(w^0)/\eta)) + n^{3.5} \log(\bar{\chi}_A + n))$ iterations, yielding part of Theorem 2.6. The other part of Theorem 2.6 is obtained using slightly different reasoning.

3.2. LLS directions and their relationship with the AS direction. In this subsection we describe another type of direction, which plays an important role on a criterion that guarantees the occurrence of crossover events (see Lemma 3.3), namely, the LLS direction. We also state a proximity result, which describes how the AS direction can be well approximated by suitable LLS directions.

The LLS direction was first introduced by Vavasis and Ye in [30] and is one of two directions used in their algorithm. While the algorithm in this paper does not rely on this direction, its analysis relies heavily on it by means of the implications of Lemma 3.3.

Let $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ and a partition $(J_1, \dots, J_p)$ of the index set $\{1, \dots, n\}$ be given and define $\delta \equiv \delta(w)$. The primal LLS direction $\Delta x^{\text{ll}} = (\Delta x_{J_1}^{\text{ll}}, \dots, \Delta x_{J_p}^{\text{ll}})$ at w with respect to J is defined recursively according to the order $\Delta x_{J_p}^{\text{ll}}, \dots, \Delta x_{J_1}^{\text{ll}}$ as follows. Assume that the components $\Delta x_{J_p}^{\text{ll}}, \dots, \Delta x_{J_{k+1}}^{\text{ll}}$ have been determined. Let $\Pi_{J_k} : \mathbb{R}^n \rightarrow \mathbb{R}^{J_k}$ denote the projection map defined as $\Pi_{J_k}(u) = u_{J_k}$ for all $u \in \mathbb{R}^n$. Then $\Delta x_{J_k}^{\text{ll}} \equiv \Pi_{J_k}(L_k^x)$, where L_k^x is given by

$$(3.2) \quad \begin{aligned} L_k^x &\equiv \text{Argmin}_{p \in \mathbb{R}^n} \{ \|\delta_{J_k}(x_{J_k} + p_{J_k})\|^2 : p \in L_{k+1}^x \} \\ &= \text{Argmin}_{p \in \mathbb{R}^n} \{ \|\delta_{J_k}(x_{J_k} + p_{J_k})\|^2 : p \in \text{Ker}(A), p_{J_i} = \Delta x_{J_i}^{\text{ll}} \\ &\quad \forall i = k+1, \dots, p \}, \end{aligned}$$

with the convention that $L_{p+1}^x \equiv \text{Ker}(A)$. The slack component $\Delta s^{\text{ll}} = (\Delta s_{J_1}^{\text{ll}}, \dots, \Delta s_{J_p}^{\text{ll}})$ of the dual LLS direction $(\Delta y^{\text{ll}}, \Delta s^{\text{ll}})$ at w with respect to J is defined recursively as follows. Assume that the components $\Delta s_{J_1}^{\text{ll}}, \dots, \Delta s_{J_{k-1}}^{\text{ll}}$ have been determined. Then $\Delta s_{J_k}^{\text{ll}} \equiv \Pi_{J_k}(L_k^s)$, where L_k^s is given by

$$(3.3) \quad \begin{aligned} L_k^s &\equiv \text{Argmin}_{q \in \mathbb{R}^n} \{ \|\delta_{J_k}^{-1}(s_{J_k} + q_{J_k})\|^2 : q \in L_{k-1}^s \} \\ &= \text{Argmin}_{q \in \mathbb{R}^n} \{ \|\delta_{J_k}^{-1}(s_{J_k} + q_{J_k})\|^2 : q \in \text{Im}(A^T), q_{J_i} = \Delta s_{J_i}^{\text{ll}} \\ &\quad \forall i = 1, \dots, k-1 \}, \end{aligned}$$

with the convention that $L_0^s \equiv \text{Im}(A^T)$. Finally, once Δs^{ll} has been determined, the component Δy^{ll} is determined from the relation $A^T \Delta y^{\text{ll}} + \Delta s^{\text{ll}} = 0$.

It is easy to verify that the AS direction is a special LLS direction, namely, the one with respect to the only partition in which $p = 1$ (see section 5 of [30]). Clearly, the LLS direction at a given $w \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ depends on the partition $J = (J_1, \dots, J_p)$ used.

A partition $J = (J_1, \dots, J_p)$ is said to be *ordered* at a point $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ if $\max(\delta_{J_i}) \leq \min(\delta_{J_{i+1}})$ for all $i = 1, \dots, p-1$, where $\delta \equiv \delta(w)$. In this case, the gap of J , denoted by $\text{gap}(J)$, is defined as

$$\text{gap}(J) = \min_{1 \leq i \leq p-1} \left\{ \frac{\min(\delta_{J_{i+1}})}{\max(\delta_{J_i})} \right\} = \frac{1}{\max_{1 \leq i \leq p-1} (\|\delta_{J_i}\|_\infty \|\delta_{J_{i+1}}^{-1}\|_\infty)} \geq 1,$$

with the convention that $\text{gap}(J) = \infty$ if $p = 1$.

In the remainder of this subsection, we describe how the AS direction at a given $w \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ can be well approximated by suitable LLS steps, a result that will be important in our convergence analysis. Another result along this direction has also been obtained by Vavasis and Ye [32]. However, our result is more general and better suited for the development of this paper.

The proximity result below can be proved using the projection decomposition techniques developed in [26]. Another proof, using instead the techniques developed in [18], has been given in [19]. The result essentially states that the larger the gap of J is, the closer the AS direction and the LLS direction with respect to J will be to each other.

PROPOSITION 3.2. *Let $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ and an ordered partition $J = (J_1, \dots, J_p)$ at w be given, and let (Rx^a, Rs^a) and (Rx^l, Rs^l) , respectively, denote the residuals of the AS direction at w and of the LLS direction at w with respect to J . If $\text{gap}(J) \geq 4p\bar{\chi}_A$, then*

$$\max\{\|Rx^a - Rx^l\|_\infty, \|Rs^a - Rs^l\|_\infty\} \leq \frac{12\sqrt{n}\bar{\chi}_A}{\text{gap}(J)}.$$

In view of the above result, the AS direction can be well approximated by the LLS directions with respect to ordered partitions J , which have large gaps. The LLS direction with $p = 1$, which is the AS direction, provides the perfect approximation to the AS direction itself. However, this kind of trivial approximation is not useful for us due to the need for keeping under control the “spread” of some layers J_k . For an ordered partition $J = (J_1, \dots, J_p)$ at w , the spread of the layer J_k , denoted by $\text{spr}(J_k)$, is defined as

$$\text{spr}(J_k) \equiv \frac{\max(\delta_{J_k})}{\min(\delta_{J_k})} \quad \forall k = 1, \dots, p.$$

3.3. Relation between crossover events, the AS step, and the LLS directions. In this subsection we develop some variants of Lemma 17 of Vavasis and Ye [30] which are particularly suitable for our analysis. Specifically, we develop two estimates on the number of iterations that the MTY P-C algorithm needs to perform for some crossover event to occur. While the first estimate essentially depends on the size of the residual of the LLS step and the step-size at the initial iterate, the second one, derived with the aid of Proposition 3.2, depends only on the size of the residual of the AS direction at the initial iterate.

We start by stating an immediate consequence of Lemma 17 of [30], whose proof can be found in Lemma 3.4 of Monteiro and Tsuchiya [19].

LEMMA 3.3. Let $w = (x, y, s) \in \mathcal{N}(\beta)$ for some $\beta \in (0, 1)$ and an ordered partition $J = (J_1, \dots, J_p)$ at w be given. Let $\delta \equiv \delta(w)$, $\mu = \mu(w)$, and $(Rx^{\text{ll}}, Rs^{\text{ll}})$ denote the residual of the LLS direction $(\Delta x^{\text{ll}}, \Delta y^{\text{ll}}, \Delta s^{\text{ll}})$ at w with respect to J . Then, for any index $q \in \{1, \dots, p\}$, any constant

$$\mathcal{C}_q \geq (1 + \beta) \text{spr}(J_q) / (1 - \beta)^2,$$

and any $\mu' \in (0, \mu)$ such that

$$\frac{\mu'}{\mu} \leq \frac{\|Rx_{J_q}^{\text{ll}}\|_{\infty} \|Rs_{J_q}^{\text{ll}}\|_{\infty}}{n^3 \mathcal{C}_q^2 \bar{\chi}_A^2},$$

the interval $(\mu', \mu]$ contains a $\mathcal{C}_q$ -crossover event.

Using the above result, we can now derive the main result of this subsection, which provides two estimates on the number of iterations that the MTY P-C algorithm needs to perform until some crossover event occurs. This result is a slight variation of Lemma 3.5 of Monteiro and Tsuchiya [19], which is more suitable for our analysis.

LEMMA 3.4. Suppose $\beta \in (0, 1/4]$ and that $J = (J_1, \dots, J_p)$ is an ordered partition at w . Let $w = (x, y, s) \in \mathcal{N}(2\beta)$ be such that $\mu(w^{k+1}) \leq \mu(w) \leq \mu(w^k)$ for some iteration index k of the MTY P-C algorithm. Let $(Rx^{\text{ll}}, Rs^{\text{ll}})$ denote the residual of the LLS direction $(\Delta x^{\text{ll}}, \Delta y^{\text{ll}}, \Delta s^{\text{ll}})$ at w with respect to J . Then, for every $q \in \{1, \dots, p\}$ and every $\mathcal{C}_q \geq (1 + 2\beta) \text{spr}(J_q) / (1 - 2\beta)^2$, the following statements hold:

(a) There exists an iteration index $l > k$ such that

$$(3.4) \quad l - k = \mathcal{O} \left(\sqrt{n} \left(\log(\bar{\chi}_A + n) + \log \mathcal{C}_q + \log \left(\frac{\mu(w^{k+1}) / \mu(w)}{\|Rx_{J_q}^{\text{ll}}\|_{\infty} \|Rs_{J_q}^{\text{ll}}\|_{\infty}} \right) \right) \right)$$

and with the property that either a $\mathcal{C}_q$ -crossover event occurs between w^k and w^l or the algorithm terminates at or before the l th iteration.

(b) If, in addition,

$$(3.5) \quad \text{gap}(J) \geq \max \left\{ 4n\bar{\chi}_A, \frac{24\sqrt{n}\bar{\chi}_A}{\varepsilon_{J_q}^a} \right\},$$

where $\varepsilon_{J_q}^a \equiv \min\{\|Rx_{J_q}^a\|_{\infty}, \|Rs_{J_q}^a\|_{\infty}\}$, then the iteration index l above satisfies

$$(3.6) \quad l - k = \mathcal{O}(\sqrt{n}(\log(\bar{\chi}_A + n) + \log \mathcal{C}_q + \log[(\varepsilon_{J_q}^a)^{-1}])).$$

Proof. To prove (a), assume that the MTY P-C algorithm does not terminate at or before the l th iteration. Lemma 3.3 guarantees that the interval $(\mu(w^l), \mu(w))$ contains a $\mathcal{C}_q$ -crossover event, and hence that a $\mathcal{C}_q$ -crossover event occurs between w^k and w^l whenever

$$(3.7) \quad \frac{\mu(w^l)}{\mu(w)} = \frac{\mu(w^l)}{\mu(w^{k+1})} \frac{\mu(w^{k+1})}{\mu(w)} \leq \frac{\|Rx_{J_q}^{\text{ll}}\|_{\infty} \|Rs_{J_q}^{\text{ll}}\|_{\infty}}{n^3 \mathcal{C}_q^2 \bar{\chi}_A^2}.$$

Since, by Lemma 2.5, $\mu(w^{j+1}) / \mu(w^j) \leq 1 - \sqrt{\beta/n}$ for all $j \geq 0$, we conclude that (3.7) holds for any l satisfying

$$\log \left(\frac{\mu(w^{k+1})}{\mu(w)} \right) + (l - k - 1) \log \left(1 - \sqrt{\frac{\beta}{n}} \right) \leq \log \left[\frac{\|Rx_{J_q}^{\text{ll}}\|_{\infty} \|Rs_{J_q}^{\text{ll}}\|_{\infty}}{n^3 \mathcal{C}_q^2 \bar{\chi}_A^2} \right].$$

Now, using the fact that $\log(1+x) < x$ for any $x > -1$, it is easy to see that (3.4) holds for the smallest l satisfying the above inequality. Hence, (a) follows.

To prove (b), it is sufficient to show that the bound in (3.4) is bounded above by the one in (3.6) when (3.5) holds. Indeed, by Proposition 3.2 and (3.5), it follows that

$$\max\{\|Rx^a - Rx^l\|_\infty, \|Rs^a - Rs^l\|_\infty\} \leq \frac{12\sqrt{n}\bar{\chi}_A}{\text{gap}(J)} \leq \frac{\varepsilon_{J_q}^a}{2}.$$

Hence, we have

$$\begin{aligned} & \min\{\|Rx_{J_q}^l\|_\infty, \|Rs_{J_q}^l\|_\infty\} \\ & \geq \min\{\|Rx_{J_q}^a\|_\infty - \|Rx^a - Rx^l\|_\infty, \|Rs_{J_q}^a\|_\infty - \|Rs^a - Rs^l\|_\infty\} \\ & \geq \min\{\|Rx_{J_q}^a\|_\infty, \|Rs_{J_q}^a\|_\infty\} - \frac{\varepsilon_{J_q}^a}{2} = \varepsilon_{J_q}^a - \frac{\varepsilon_{J_q}^a}{2} = \frac{\varepsilon_{J_q}^a}{2}. \end{aligned}$$

Using this estimate in (3.4) together with the fact that $\mu(w^{k+1})/\mu(w) \leq 1$, we conclude that the right-hand side of (3.4) is bounded above by the right-hand side of (3.6). $\square$

3.4. Two important ordered partitions. In this subsection we describe two ordered partitions which are crucial in the analysis of this paper.

The first ordered partition is due to Vavasis and Ye [30]. Given a point $w \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ and a parameter $\bar{g} \geq 1$, this partition, which we refer to as the *VY $\bar{g}$ -partition* at w , is defined as follows. Let $(i_1, \dots, i_n)$ be an ordering of $\{1, \dots, n\}$ such that $\delta_{i_1} \leq \dots \leq \delta_{i_n}$, where $\delta = \delta(w)$. For $k = 2, \dots, n$, let $r_k \equiv \delta_{i_k}/\delta_{i_{k-1}}$ and define $r_1 \equiv \infty$. Let $k_1 < \dots < k_p$ be all the indices k such that $r_k > \bar{g}$ for all $j = 1, \dots, p$. The VY $\bar{g}$ -partition J is then defined as $J = (J_1, \dots, J_p)$, where $J_q \equiv \{i_{k_q}, i_{k_q+1}, \dots, i_{k_{q+1}-1}\}$ for all $q = 1, \dots, p$. (Here, by convention, $k_{p+1} \equiv n+1$.) More generally, given a subset $I \subseteq \{1, \dots, n\}$, we can similarly define the *VY $\bar{g}$ -partition* of I at w by taking an ordering $(i_1, \dots, i_m)$ of I satisfying $\delta_{i_1} \leq \dots \leq \delta_{i_m}$ where $m = |I|$, defining the ratios $r_1, \dots, r_m$ as above, and proceeding exactly as in the construction above to obtain an ordered partition $J = (J_1, \dots, J_p)$ of I .

It is easy to see that the following result holds for the partition J described in the previous paragraph (see section 5 of [30]).

PROPOSITION 3.5. *Given a subset $I \subseteq \{1, \dots, n\}$, a point $w \in \mathcal{P}^{++} \times \mathcal{D}^{++}$, and a constant $\bar{g} \geq 1$, the VY $\bar{g}$ -partition $J = (J_1, \dots, J_p)$ of I at w satisfies $\text{gap}(J) > \bar{g}$ and $\text{spr}(J_q) \leq \bar{g}^{|J_q|} \leq \bar{g}^n$ for all $q = 1, \dots, p$.*

The second ordered partition, which is used heavily in our analysis, is obtained as follows. Given a point $w \in \mathcal{P}^{++} \times \mathcal{D}^{++}$, we first use (2.15) to compute the AS-bipartition $(B, N) = (B(w), N(w))$ at w . Next, an order $(i_1, \dots, i_n)$ of the index variables is chosen such that $\delta_{i_1} \leq \dots \leq \delta_{i_n}$. Then, the first block of consecutive indices in the n -tuple $(i_1, \dots, i_n)$ lying in the same set B or N are placed in the first layer $\mathcal{J}_1$, the next block of consecutive indices lying in the other set is placed in $\mathcal{J}_2$, and so on. As an example assume that $(i_1, i_2, i_3, i_4, i_5, i_6, i_7) \in B \times B \times N \times B \times B \times N \times N$. In this case, we have $\mathcal{J}_1 = \{i_1, i_2\}$, $\mathcal{J}_2 = \{i_3\}$, $\mathcal{J}_3 = \{i_4, i_5\}$, and $\mathcal{J}_4 = \{i_6, i_7\}$. A partition obtained according to the above construction is clearly ordered at w . We refer to it as an *ordered AS-partition* and denote it by $\mathcal{J}(w)$.

Note that an ordered AS-partition is not uniquely determined since there can be more than one n -tuple $(i_1, \dots, i_n)$ satisfying $\delta_{i_1} \leq \dots \leq \delta_{i_n}$. This situation occurs exactly when there are two or more indices i with the same value for δ_i . It can be

easily seen that there exists a unique ordered AS-partition at w if and only if there do not exist $i \in B(w)$ and $j \in N(w)$ such that $\delta_i = \delta_j$. Hence, if the AS-bipartition $(B(w), N(w))$ does not have the latter property, there can be multiple ordered AS-partitions at w . In spite of this ambiguity, our analysis in this paper is valid for any chosen ordered AS-partition. So there is no need to have the notion of ordered AS-partition uniquely defined although this can be easily accomplished.

4. Convergence analysis of the MTY P-C algorithm. In this section, we provide the proof of Theorem 2.6.

We first introduce some global constants which will be used in the convergence analysis of this section. Let

$$(4.1) \quad \mathcal{C} \equiv \frac{(1+2\beta)}{(1-2\beta)^2} (2\bar{g})^n \quad \text{and} \quad \bar{g} \equiv \frac{24\bar{\chi}_A n}{\tau},$$

where

$$(4.2) \quad \tau = \tau(\beta) \equiv \frac{\beta(1-\beta)^3(1-2\beta)^2}{4(1+\beta)^2(1+2\beta)}.$$

Note that, for a point $w \in \mathcal{P}^{++} \times \mathcal{D}^{++}$, we have in view of (2.11) and (2.15) that

$$(4.3) \quad \varepsilon_\infty^a(w) = \max\{\|Rx_N^a(w)\|_\infty, \|Rs_B^a(w)\|_\infty\},$$

where $B \equiv B(w)$ and $N \equiv N(w)$. Clearly, $\varepsilon_\infty^a(w)$ is an upper bound on the absolute value of the small components of the residual $(Rx^a(w), Rs^a(w))$. The next result gives a lower bound on the absolute value of the large components of the residual $(Rx^a(w), Rs^a(w))$. For a proof of this result, we refer the reader to Lemma 2.6 of Monteiro and Tsuchiya [19].

LEMMA 4.1. *Suppose that $w = (x, y, s) \in \mathcal{N}(\beta)$ for some $\beta \in (0, 1)$. Then, we have*

$$\max\{|Rx_i^a(w)|, |Rs_i^a(w)|\} \geq \frac{\sqrt{1-\beta}}{2}$$

for all $i = 1, \dots, n$, or equivalently, in view of the definition of $(B, N) \equiv (B(w), N(w))$, we have

$$\min \left\{ \min_{i \in B} |Rx_i^a(w)|, \min_{i \in N} |Rs_i^a(w)| \right\} \geq \frac{\sqrt{1-\beta}}{2}.$$

Lemma 3.4 gives a good idea of part of the effort that will be undertaken in this section, namely, to find conditions under which the bounds (3.4) or (3.6) obtained in this result can be majorized by the quantity $\mathcal{O}(n \log(\bar{\chi}_A + n) + \log(\log(\mu(w^0)/\eta)))$. We will break our analysis into the following three cases:

- (i) $\text{gap}(\mathcal{J}(w)) \leq 2\bar{g}$ (Lemma 4.2);
- (ii) $\text{gap}(\mathcal{J}(w)) \geq \bar{g}$ and $\varepsilon_\infty^a(w) \geq \tau\bar{g}/(\sqrt{n} \text{gap}(\mathcal{J}(w)))$ (Lemma 4.5);
- (iii) neither (i) nor (ii) holds (Lemma 4.11).

We will now give an outline of the approaches used to tackle each of the above three cases. In case (i), it is easy to see that at least one variable in N and one variable in B are in the same layer J_q of a VY $2\bar{g}$ -partition $J = (J_1, \dots, J_p)$ at w . We call a layer of this type a “mixed VY-layer.” The proof of Lemma 4.2 essentially shows that the existence of a mixed layer implies that the quantity $\varepsilon_{J_q}^a$ in Lemma 3.4(b)

is not too small and that a $\mathcal{C}$ -crossover event must occur within $\mathcal{O}(n^{1.5} \log(\bar{\chi}_A + n))$ iterations of the MTY P-C algorithm.

In case (ii), the quantity $\text{gap}(\mathcal{J}(w))$ is sufficiently large to guarantee that the AS-direction can be well approximated by an LLS direction at w with respect to a suitable ordered partition at w obtained from an ordered AS-partition at w by breaking one of its layers into smaller layers. Using the close proximity of these two directions and the fact that a step along this LLS direction followed by at most $\mathcal{O}(n^{1.5} \log(\bar{\chi}_A + n))$ regular MTY P-C steps yields a $\mathcal{C}$ -crossover event, we show that case (ii) also implies that a $\mathcal{C}$ -crossover event must occur within $\mathcal{O}(n^{1.5} \log(\bar{\chi}_A + n))$ iterations of the MTY P-C algorithm.

In case (iii), the proximity between the AS direction and the LLS direction mentioned above is not small enough to guarantee that the reduction of the duality gap along these two directions are of the same order of magnitude. However, in this case we manage to show that within $\mathcal{O}(\log(\log(\mu(w^0)/\eta)))$ iterations of the MTY P-C algorithm we return to a situation in which either a mixed layer arises, or case (ii) above holds, or an index changes status (moves from B to N or vice versa) from one iteration to the next. The same kind of techniques used to handle case (i) can be used to show that the latter possibility also implies that a $\mathcal{C}$ -crossover event must occur within $\mathcal{O}(n^{1.5} \log(\bar{\chi}_A + n))$ iterations of the MTY P-C algorithm. Overall, we conclude that case (iii) also implies that a $\mathcal{C}$ -crossover event must occur within $\mathcal{O}(n^{1.5} \log(\bar{\chi}_A + n) + \log(\log(\mu(w^0)/\eta)))$ iterations of the MTY P-C algorithm.

The first result below is a slight variation of Lemma 4.2 of [19]. It considers case (i) above, which can be handled by applying Lemma 3.4(b) with the ordered partition $J = (J_1, \dots, J_p)$ chosen to be the VY $2\bar{g}$ -partition at w .

LEMMA 4.2. *Suppose $\beta \in (0, 1/4]$ and $w = (x, y, s) \in \mathcal{N}(2\beta)$ is such that $\mu(w^{k+1}) \leq \mu(w) \leq \mu(w^k)$ for some iteration index k of the MTY P-C algorithm. Assume that $\text{gap}(\mathcal{J}) \leq 2\bar{g}$, where $\mathcal{J} = (\mathcal{J}_1, \dots, \mathcal{J}_r)$ is an ordered AS-partition at w . Then there exists an iteration index $l > k$ such that $l - k = \mathcal{O}(n^{1.5} \log(\bar{\chi}_A + n))$ and with the property that either a $\mathcal{C}$ -crossover event occurs between w^k and w^l or the algorithm terminates at or before the l th iteration.*

Proof. The proof that the required iteration index l exists is based on Lemma 3.4(b). Indeed, let $J = (J_1, \dots, J_p)$ be a VY $2\bar{g}$ -partition at w . The assumption $\text{gap}(\mathcal{J}) \leq 2\bar{g}$ implies the existence of two indices i, j lying in some layer J_q of J , with one in $B(w)$ and the other in $N(w)$. Without loss of generality, assume that $i \in B(w)$ and $j \in N(w)$. By Lemma 4.1, we then have $|Rx_i^a(w)| \geq \sqrt{1 - 2\beta}/2$ and $|Rs_j^a(w)| \geq \sqrt{1 - 2\beta}/2$. Since $i, j \in J_q$, this implies that

$$(4.4) \quad \varepsilon_{J_q}^a \equiv \min\{\|Rx_{J_q}^a(w)\|_\infty, \|Rs_{J_q}^a(w)\|_\infty\} \geq \frac{\sqrt{1 - 2\beta}}{2}.$$

Using the above inequality, the fact that $\text{gap}(J) \geq 2\bar{g}$, and relations (4.1) and (4.2), we easily see that (3.5) holds. Since by Proposition 3.5 the spread of every layer of a VY $2\bar{g}$ -partition at w is bounded above by $(2\bar{g})^n$, we conclude that $\text{spr}(J_q) \leq (2\bar{g})^n$. Hence, we may set $\mathcal{C}_q = \mathcal{C} \equiv (1 + 2\beta)(2\bar{g})^n / (1 - 2\beta)^2$ in Lemma 3.4, from which it follows that

$$(4.5) \quad \log(\mathcal{C}_q) = \mathcal{O}(n \log \bar{g}) = \mathcal{O}(n \log(\bar{\chi}_A + n)),$$

where the last equality is due to (4.1). The result now follows from Lemma 3.4(b) by noting that the bound in (3.6) is $\mathcal{O}(n^{1.5} \log(\bar{\chi}_A + n))$ in view of (4.4) and (4.5). $\square$

From now on we consider cases (ii) and (iii), i.e., the situation in which $\text{gap}(\mathcal{J}(w)) \geq \bar{g}$. For the sake of future reference, we note that the condition $\text{gap}(\mathcal{J}(w)) \geq \bar{g}$ implies

$$(4.6) \quad \frac{\tau \bar{g}}{\sqrt{n} \text{gap}(\mathcal{J}(w))} \leq \tau \leq \frac{\sqrt{1-2\beta}}{4},$$

due to (4.2). The following lemma is a slight variation of Lemma 4.3 of [19]. For some suitably chosen ordered partition J at w , it provides an upper bound on the right-hand side of (3.4) in terms of the residual of the LLS step with respect to $\mathcal{J}(w)$. Note that for this result we assume only that $\text{gap}(\mathcal{J}(w)) \geq \bar{g}$.

LEMMA 4.3. *Suppose $\beta \in (0, 1/4]$ and $w = (x, y, s) \in \mathcal{N}(2\beta)$ is such that $\mu(w^{k+1}) \leq \mu(w) \leq \mu(w^k)$ for some iteration index k of the MTY P-C algorithm. Assume that $\text{gap}(\mathcal{J}) \geq \bar{g}$, where $\mathcal{J} = (\mathcal{J}_1, \dots, \mathcal{J}_r)$ is an ordered AS-partition at w . Let $(Rx^1(w), Rs^1(w))$ denote the residual of the LLS direction at w with respect to $\mathcal{J}$ and define*

$$(4.7) \quad \varepsilon_\infty^1(w) \equiv \max\{\|Rx_N^1(w)\|_\infty, \|Rs_B^1(w)\|_\infty\}.$$

Then, there exists an iteration index $l > k$ such that

$$(4.8) \quad l - k = \mathcal{O}\left(n^{1.5} \log(\bar{\chi}_A + n) + \sqrt{n} \log\left(\frac{\mu(w^{k+1})/\mu(w)}{\varepsilon_\infty^1(w)}\right)\right)$$

and with the property that either a C-crossover event occurs between w^k and w^l or the algorithm terminates at or before the l th iteration.

Proof. To simplify notation, let $(Rx^1, Rs^1) \equiv (Rx^1(w), Rs^1(w))$ and $\varepsilon_\infty^1 \equiv \varepsilon_\infty^1(w)$. Assume without loss of generality that $\varepsilon_\infty^1 = \|Rx_N^1\|_\infty$; the case in which $\varepsilon_\infty^1 = \|Rs_B^1\|_\infty$ can be proved similarly. Then, $\varepsilon_\infty^1 = |Rx_i^1|$ for some $i \in N$. Let $\mathcal{J}_t$ be the layer of $\mathcal{J}$ containing the index i and note that

$$(4.9) \quad \varepsilon_\infty^1 = |Rx_i^1| = \|Rx_{\mathcal{J}_t}^1\|_\infty \leq \|Rx_{\mathcal{J}_t}^1\|.$$

Now, let $\mathcal{I} = (\mathcal{I}_1, \dots, \mathcal{I}_p)$ be the VY $\bar{g}$ -partition of $\mathcal{J}_t$ at w and consider the ordered partition $\mathcal{J}'$ defined as

$$\mathcal{J}' \equiv (\mathcal{J}_1, \dots, \mathcal{J}_{t-1}, \mathcal{I}_1, \dots, \mathcal{I}_p, \mathcal{J}_{t+1}, \dots, \mathcal{J}_r).$$

Let (Rx^Π, Rs^Π) denote the residual of the LLS direction at w with respect to $\mathcal{J}'$. Using the definition of the LLS step, it is easy to see that $Rx_{\mathcal{J}_j}^1 = Rx_{\mathcal{J}_j}^\Pi$ for all $j = t+1, \dots, r$. Moreover, we have $\|Rx_{\mathcal{J}_t}^1\| \leq \|Rx_{\mathcal{J}_t}^\Pi\|$ since $\|Rx_{\mathcal{J}_t}^1\|$ is the optimal value of the least squares problem which determines the $\Delta x_{\mathcal{J}_t}^1$ -component of the LLS step with respect to $\mathcal{J}$, whereas $\|Rx_{\mathcal{J}_t}^\Pi\|$ is the objective value at a certain feasible solution for the same least squares problem. Hence, for some $q \in \{1, \dots, p\}$, we have

$$(4.10) \quad \|Rx_{\mathcal{I}_q}^\Pi\|_\infty = \|Rx_{\mathcal{J}_t}^\Pi\|_\infty \geq \frac{1}{\sqrt{|\mathcal{J}_t|}} \|Rx_{\mathcal{J}_t}^\Pi\| \geq \frac{1}{\sqrt{n}} \|Rx_{\mathcal{J}_t}^\Pi\| \geq \frac{1}{\sqrt{n}} \|Rx_{\mathcal{J}_t}^1\|.$$

Combining (4.9) and (4.10), we then obtain

$$(4.11) \quad \|Rx_{\mathcal{I}_q}^\Pi\|_\infty \geq \frac{1}{\sqrt{n}} \varepsilon_\infty^1.$$

Let us now bound the quantity $\|Rs_{\mathcal{I}_q}^{\parallel}\|_{\infty}$ from below. Using the triangle inequality for norms, Lemma 4.1 together with the fact that $\mathcal{I}_q \subseteq N$, and Proposition 3.2 together with the fact that $\text{gap}(\mathcal{J}') \geq \bar{g} = 24\bar{\chi}_A n/\tau \geq 96\bar{\chi}_A \sqrt{n}$, where the last inequality is due to (4.2), we obtain

$$(4.12) \quad \|Rs_{\mathcal{I}_q}^{\parallel}\|_{\infty} \geq \|Rs_{\mathcal{I}_q}^a\|_{\infty} - \|Rs_{\mathcal{I}_q}^{\parallel} - Rs_{\mathcal{I}_q}^a\|_{\infty} \geq \frac{1}{4} - \frac{12\sqrt{n}\bar{\chi}_A}{\text{gap}(\mathcal{J}')} \geq \frac{1}{4} - \frac{1}{8} = \frac{1}{8},$$

where $Rs^a \equiv Rs^a(w)$. Also, note that by (4.1) we have

$$(4.13) \quad \log \mathcal{C} = \mathcal{O}(n \log(\bar{\chi}_A + n)).$$

The result now follows from Lemma 3.4(a) with $J = \mathcal{J}'$ and $\mathcal{C}_q = \mathcal{C}$, the observation that Proposition 3.5 and (4.1) imply that $\mathcal{C} \geq (1 + 2\beta)\bar{g}^n/(1 - 2\beta)^2 \geq (1 + 2\beta)\text{spr}(\mathcal{I}_q)/(1 - 2\beta)^2$, and the fact that the estimates (4.11)–(4.13) imply that the bound in (3.4) is majorized by the one in (4.8). $\square$

Our goal now will be to estimate the second logarithm that appears in the iteration-complexity bound (4.8). The next result gives a condition under which $\varepsilon_{\infty}^a(w) = \mathcal{O}(\varepsilon_{\infty}^1(w))$.

LEMMA 4.4. *Let $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ be given and let $\mathcal{J} = (\mathcal{J}_1, \dots, \mathcal{J}_r)$ denote an ordered AS-partition at w . If*

$$(4.14) \quad \text{gap}(\mathcal{J}) \geq \max \left\{ 4n\bar{\chi}_A, \frac{24\sqrt{n}\bar{\chi}_A}{\varepsilon_{\infty}^a(w)} \right\},$$

then $\varepsilon_{\infty}^1(w) \geq \varepsilon_{\infty}^a(w)/2$, where $\varepsilon_{\infty}^1(w)$ is defined in (4.7).

Proof. Let (Rx^a, Rs^a) and (Rx^1, Rs^1) , respectively, denote the residuals of the AS direction at w and the LLS direction at w with respect to $\mathcal{J}(w)$. By Proposition 3.2 and condition (4.14), we have

$$\max\{\|Rx^a - Rx^1\|_{\infty}, \|Rs^a - Rs^1\|_{\infty}\} \leq \frac{12\sqrt{n}\bar{\chi}_A}{\text{gap}(\mathcal{J})} \leq \frac{\varepsilon_{\infty}^a(w)}{2}.$$

Hence, we have

$$\begin{aligned} \varepsilon_{\infty}^1(w) &\equiv \max\{\|Rx_N^1\|_{\infty}, \|Rs_B^1\|_{\infty}\} \\ &\geq \max\{\|Rx_N^a\|_{\infty} - \|Rx_N^a - Rx_N^1\|_{\infty}, \|Rs_B^a\|_{\infty} - \|Rs_B^a - Rs_B^1\|_{\infty}\} \\ &\geq \max\{\|Rx_N^a\|_{\infty}, \|Rs_B^a\|_{\infty}\} - \frac{\varepsilon_{\infty}^a(w)}{2} = \varepsilon_{\infty}^a(w) - \frac{\varepsilon_{\infty}^a(w)}{2} = \frac{\varepsilon_{\infty}^a(w)}{2}. \quad \square \end{aligned}$$

We are now ready to state and prove the result which takes care of case (ii).

LEMMA 4.5. *Suppose $\beta \in (0, 1/4]$ and $w = (x, y, s) \in \mathcal{N}(2\beta)$ is such that $\mu(w^{k+1}) \leq \mu(w) \leq \mu(w^k)$ for some iteration index k of the MTY P-C algorithm. Assume that $\text{gap}(\mathcal{J}) \geq \bar{g}$ and $\varepsilon_{\infty}^a(w) \geq \tau\bar{g}/(\sqrt{n}\text{gap}(\mathcal{J}))$, where $\mathcal{J} = (\mathcal{J}_1, \dots, \mathcal{J}_r)$ is an ordered AS-partition at w . Then, there exists an iteration index $l > k$ such that $l - k = \mathcal{O}(n^{1.5} \log(\bar{\chi}_A + n) + \sqrt{n} \log(\varepsilon_{\infty}^a(w))^{-1})$ and with the property that either a C-crossover event occurs between w^k and w^l or the algorithm terminates at or before the l th iteration. Furthermore, if $w = w^k$, then $l - k = \mathcal{O}(n^{1.5} \log(\bar{\chi}_A + n))$.*

Proof. The proof is based on Lemma 4.3 and essentially consists of showing that the term inside the second logarithm that appears in the bound (4.8) can always be bounded by $\mathcal{O}((\varepsilon_{\infty}^a(w))^{-1})$ and that, when $w = w^k$, this term can also

be bounded by $\mathcal{O}(\sqrt{n})$. Indeed, first note that the conditions $\text{gap}(\mathcal{J}) \geq \bar{g}$ and $\varepsilon_\infty^a(w) \geq \tau\bar{g}/\sqrt{n}\text{gap}(\mathcal{J})$ and relations (4.1) and (4.2) clearly imply (4.14). Hence, by Lemma 4.4, it follows that $\varepsilon_\infty^1(w) \geq \varepsilon_\infty^a(w)/2$. Thus, we have

$$\frac{\mu(w^{k+1})/\mu(w)}{\varepsilon_\infty^1(w)} \leq \frac{2\mu(w^{k+1})/\mu(w)}{\varepsilon_\infty^a(w)} \leq \frac{2}{\varepsilon_\infty^a(w)} = \mathcal{O}((\varepsilon_\infty^a(w))^{-1}),$$

where in the second inequality we used the fact that $\mu(w) \geq \mu(w^{k+1})$. When $w = w^k$, it follows from the first inequality above and Lemma 2.5 that

$$\frac{\mu(w^{k+1})/\mu(w)}{\varepsilon_\infty^1(w)} \leq \frac{2\mu(w^{k+1})/\mu(w)}{\varepsilon_\infty^a(w)} \leq \frac{2\sqrt{n}}{\beta} = \mathcal{O}(\sqrt{n}).$$

The result now follows from Lemma 4.3 and the bounds obtained above. $\square$

From now on, we will consider the case in which an iterate $w = w^k$ of the MTY P-C algorithm satisfies $\text{gap}(\mathcal{J}(w)) \geq \bar{g}$ and $\varepsilon_\infty^a(w) \leq \tau\bar{g}/(\sqrt{n}\text{gap}(\mathcal{J}(w)))$. Before tackling this case, we first need to establish two technical lemmas, namely, Lemmas 4.6 and 4.9, which give other sufficient conditions for the occurrence of $\mathcal{C}$ -crossover events. Lemma 4.7 is used only in the proof of Lemma 4.8, which in turn is used only in the proof of Lemma 4.9.

LEMMA 4.6. *Suppose that $\beta \in (0, 1/4]$ and that $w = w^k \in \mathcal{N}(\beta)$ is an iterate of the MTY P-C algorithm and let $w^+ \equiv w^{k+1}$. Assume that either one of the following conditions holds:*

- (a) *either $B(w) \cap N(w^+) \neq \emptyset$ or $N(w) \cap B(w^+) \neq \emptyset$;*
- (b) *there exist indices i and j , one lying in $B(w)$ and the other in $N(w)$, such that $\delta_i(w)/\delta_j(w) \geq \bar{g}$ and $\delta_i(w^+)/\delta_j(w^+) \leq \bar{g}$.*

Then, there exists an iteration index $l > k$ such that $l - k = \mathcal{O}(n^{1.5} \log(\bar{\chi}_A + n))$ and with the property that either a $\mathcal{C}$ -crossover event occurs between w^k and w^l or the algorithm terminates at or before the l th iteration.

Proof. Let $\mathcal{J} \equiv \mathcal{J}(w)$. If either one of the conditions $\text{gap}(\mathcal{J}) \geq \bar{g}$ or $\varepsilon_\infty^a(w) \leq \tau\bar{g}/(\sqrt{n}\text{gap}(\mathcal{J}))$ does not hold, then the conclusion of the lemma follows immediately from Lemmas 4.2 and 4.5, regardless of whether one of conditions (a) or (b) holds. Assume then that $\text{gap}(\mathcal{J}) \geq \bar{g}$ and $\varepsilon_\infty^a(w) \leq \tau\bar{g}/(\sqrt{n}\text{gap}(\mathcal{J}))$. This together with (4.6) imply that $\varepsilon_\infty^a(w) \leq \sqrt{1-2\beta}/4$. We will first show that the conclusion of the lemma holds under condition (a). Let $w : [0, 1] \rightarrow \mathcal{N}(2\beta)$ be a continuous path such that $w(0) = w$, $w(1) = w^+$, and $\mu(w^+) \leq \mu(w(t)) \leq \mu(w)$ for all $t \in [0, 1]$; e.g., consider the path that traces the line segment from w to $w + \alpha_a \Delta w^a$ and then the line segment from $w + \alpha_a \Delta w^a$ to w^+ . (It is straightforward to verify that these two segments lie in $\mathcal{N}(2\beta)$.) We will show more generally that if there exists $0 < t \leq 1$ such that either $B(w) \cap N(w(t)) \neq \emptyset$ or $N(w) \cap B(w(t)) \neq \emptyset$ (for $t = 1$ this is condition (a)), then the conclusion of the lemma follows. Indeed, assume that for some $0 < t \leq 1$ there exists an index j in the set $B(w) \cap N(w(t))$. (The proof is similar for the case in which $j \in N(w) \cap B(w(t))$.) Since $j \in B(w)$ and $\varepsilon_\infty^a(w) \leq \sqrt{1-2\beta}/4$, we have $|Rs_j^a(w(0))| = |Rs_j^a(w)| \leq \varepsilon_\infty^a(w) \leq \sqrt{1-2\beta}/4$. Moreover, since $j \in N(w(t))$, we have $|Rs_j^a(w(t))| \geq \sqrt{1-2\beta}/2$ in view of Lemma 4.1. The intermediate value theorem applied to the continuous function $t \rightarrow |Rs_j^a(w(t))|$ implies the existence of some $\bar{t} \in [0, t]$ such that $|Rs_j^a(w(\bar{t}))| = \sqrt{1-2\beta}/4$. Letting $\bar{w} \equiv w(\bar{t})$, we have $\bar{w} \in \mathcal{N}(2\beta)$ and $|Rs_j^a(\bar{w})| = \sqrt{1-2\beta}/4$. By Lemma 4.1 with β replaced by 2β we have that $\max\{|Rx_j^a(\bar{w})|, |Rs_j^a(\bar{w})|\} \geq \sqrt{1-2\beta}/2$. Since $|Rs_j^a(\bar{w})| = \sqrt{1-2\beta}/4 < \sqrt{1-2\beta}/2$, we must have $|Rx_j^a(\bar{w})| \geq \sqrt{1-2\beta}/2$. We thus proved that $\varepsilon_\infty^a(\bar{w}) \geq$

$\min\{|Rx_j^a(\bar{w})|, |Rs_j^a(\bar{w})|\} \geq \sqrt{1-2\beta}/4$. If $\text{gap}(\mathcal{J}(\bar{w})) \geq \bar{g}$, then by (4.6) and the fact that $\varepsilon_\infty^a(\bar{w}) \geq \sqrt{1-2\beta}/4$, the conclusion of the lemma follows from Lemma 4.5 with $w = \bar{w}$. If, on the other hand, $\text{gap}(\mathcal{J}(\bar{w})) < \bar{g}$, then the conclusion of the lemma follows from Lemma 4.2 with $w = \bar{w}$.

We will now show that the conclusion of the lemma holds under condition (b). Without loss of generality, assume that $i \in B(w)$ and $j \in N(w)$ is such that $\delta_i(w)/\delta_j(w) \geq \bar{g}$ and $\delta_i(w^+)/\delta_j(w^+) \leq \bar{g}$. The intermediate value theorem applied to the continuous function $t \rightarrow \delta_i(w(t))/\delta_j(w(t))$ implies the existence of some $\bar{t} \in [0, 1]$ such that $\delta_i(w(\bar{t}))/\delta_j(w(\bar{t})) = \bar{g}$. If either $i \in N(w(\bar{t}))$ or $j \in B(w(\bar{t}))$, then the conclusion of the lemma holds in view of what we have already shown in the previous paragraph. Consider now the case in which $i \in B(w(\bar{t}))$ and $j \in N(w(\bar{t}))$. Since $\delta_i(w(\bar{t}))/\delta_j(w(\bar{t})) = \bar{g}$, this case implies that $\text{gap}(\mathcal{J}(w(\bar{t}))) \leq \bar{g}$. Thus, the conclusion of the lemma follows from Lemma 4.2 with $w = w(\bar{t})$. $\square$

It is worth noting that Lemma 4.6(a) gives a simple and intuitive scaling-invariant sufficient condition for a crossover event to occur within a reasonable number of iterations, namely, that an index changes its status from the iterate w to the next iterate w^+ .

We now make an important observation that will be used in the proof of Lemma 4.11 below. Namely, if $\text{gap} \mathcal{J}(w) \geq \bar{g}$ and neither condition (a) nor (b) of Lemma 4.6 holds, then we must have $(B(w^+), N(w^+)) = (B(w), N(w))$ and $\mathcal{J}(w) = \mathcal{J}(w^+)$; i.e., the AS-bipartition and the ordered AS-partition do not change during the iteration from w to w^+ . Indeed, assume that $i, j \in \{1, \dots, n\}$ are indices such that $\delta_i(w) > \delta_j(w)$ and one of them lies in $B(w)$ while the other lies in $N(w)$. Since $\text{gap} \mathcal{J}(w) \geq \bar{g}$, this implies that $\delta_i(w)/\delta_j(w) \geq \bar{g}$. Moreover, since we are assuming that (b) of Lemma 4.6 does not hold, we must have $\delta_i(w^+)/\delta_j(w^+) \geq \bar{g}$, and hence that $\delta_i(w^+) > \delta_j(w^+)$. We have thus shown that the order of the δ_i 's for indices of different types are preserved while moving from w to w^+ . Moreover, since we are assuming that (a) of Lemma 4.6 does not hold, it follows that indices do not change status while moving from w to w^+ . Using these two conclusions together, we easily see that $(B(w^+), N(w^+)) = (B(w), N(w))$ and $\mathcal{J}(w) = \mathcal{J}(w^+)$.

The next lemma describes how the ratio between any pair of dual slacks varies as we move from one primal-dual feasible point to another one.

LEMMA 4.7. *For some $\beta, \beta' \in (0, 1)$, let $w = (x, s, y) \in \mathcal{N}(\beta)$ and $w' = (x', s', y') \in \mathcal{N}(\beta')$ be points such that $\mu(w) = \mu(w')$. Then, for every $i, j \in \{1, \dots, n\}$, we have*

$$(4.15) \quad \frac{s_i}{s_j} \leq \frac{(1+\beta)(1+\beta')}{(1-\beta)^2(1-\beta')^2} \frac{s'_i}{s'_j}.$$

Proof. Let $\mu \equiv \mu(w) = \mu(w')$. It is well known that

$$\frac{1-\beta}{1+\beta} s \leq s(\mu) \leq \frac{1}{1-\beta} s, \quad \frac{1-\beta'}{1+\beta'} s' \leq s(\mu) \leq \frac{1}{1-\beta'} s'$$

(see, for example, Lemma 2.4(ii) of Gonzaga [3] and Proposition 2.1 of Monteiro and Tsuchiya [19]). For $i, j \in \{1, \dots, n\}$, these relations then imply that

$$\begin{aligned} s_i &\leq \frac{1+\beta}{1-\beta} s_i(\mu) \leq \frac{1+\beta}{(1-\beta)(1-\beta')} s'_i, \\ s_j &\geq (1-\beta) s_j(\mu) \geq \frac{(1-\beta)(1-\beta')}{1+\beta'} s'_j, \end{aligned}$$

from which (4.15) immediately follows. $\square$

The following technical lemma describes how the ratio function $\delta_i(\cdot)/\delta_j(\cdot)$, for a fixed pair $(i, j) \in B(w) \times N(w)$, varies from $w = w^k$ to $w^+ = w^{k+1}$.

LEMMA 4.8. *Suppose $\beta \in (0, 1/4]$ and assume that $w = (x, s, y)$ and $w^+ = (x^+, s^+, y^+)$ are two consecutive iterates of the MTY P-C algorithm. Assume also that $\varepsilon_\infty^a(w) \leq \sqrt{1-\beta}/2$. Then, for every $i \in B(w)$ and $j \in N(w)$, we have*

$$(4.16) \quad \frac{\delta_i(w^+)}{\delta_j(w^+)} \leq \frac{\sqrt{n}}{\tau} \frac{\delta_i(w)}{\delta_j(w)} \varepsilon_\infty^a(w),$$

where $\tau = \tau(\beta)$ is the constant defined in (4.2).

Proof. To simplify notation, let $\varepsilon_\infty^a \equiv \varepsilon_\infty^a(w)$, $\delta^+ \equiv \delta_i(w^+)$, and $\delta \equiv \delta(w)$. Also, let $\Delta w^a \equiv (\Delta x^a, \Delta s^a, \Delta y^a)$ denote the AS direction at w and define $\tilde{w} = (\tilde{x}, \tilde{s}, \tilde{y}) \equiv w + \alpha_a \Delta w^a$. Using (2.4), the fact that $w^+ \in \mathcal{N}(\beta)$, and Lemma 4.7 with $w = w^+$ and $w' = \tilde{w}$, we obtain

$$(4.17) \quad \frac{\delta_i^+}{\delta_j^+} = \frac{(x_j^+ s_j^+)^{1/2} s_i^+}{(x_i^+ s_i^+)^{1/2} s_j^+} \leq \frac{\sqrt{1+\beta} s_i^+}{\sqrt{1-\beta} s_j^+} \leq \frac{(1+\beta)^{3/2}(1+2\beta)}{(1-\beta)^{5/2}(1-2\beta)^2} \frac{\tilde{s}_i}{\tilde{s}_j}.$$

Now, letting $(Rx^a, Rs^a) \equiv (Rx^a(w), Rs^a(w))$ and using Lemma 2.5, (2.11), and the fact that $i \in B(w)$, we obtain

$$(4.18) \quad \begin{aligned} \tilde{s}_i &= s_i + \alpha_a \Delta s_i^a = (1 - \alpha_a) s_i + \alpha_a (s_i + \Delta s_i^a) \\ &= \sqrt{\mu} \delta_i \left[(1 - \alpha_a) \frac{x_i^{1/2} s_i^{1/2}}{\sqrt{\mu}} + \alpha_a R s_i^a \right] \\ &\leq \sqrt{\mu} \delta_i \left[\frac{\sqrt{n} \varepsilon_\infty^a}{\beta} \sqrt{1+\beta} + \varepsilon_\infty^a \right] \leq \sqrt{\mu} \delta_i \varepsilon_\infty^a \frac{2\sqrt{(1+\beta)n}}{\beta}. \end{aligned}$$

Now, using relations (2.4), (2.9), and (4.3), the fact that $j \in N(w)$ and $\alpha_a \leq 1$, and the assumption that $\varepsilon_\infty^a \leq \sqrt{1-\beta}/2$, we obtain

$$(4.19) \quad \begin{aligned} \tilde{s}_j &= s_j + \alpha_a \Delta s_j^a = \sqrt{\mu} \delta_j \left[\frac{x_j^{1/2} s_j^{1/2}}{\sqrt{\mu}} + \alpha_a \frac{\delta_j^{-1} \Delta s_j^a}{\sqrt{\mu}} \right] \\ &= \sqrt{\mu} \delta_j \left[\frac{x_j^{1/2} s_j^{1/2}}{\sqrt{\mu}} - \alpha_a R x_j^a \right] \geq \sqrt{\mu} \delta_j [\sqrt{1-\beta} - \varepsilon_\infty^a] \geq \sqrt{\mu} \delta_j \frac{\sqrt{1-\beta}}{2}. \end{aligned}$$

Merging (4.18) and (4.19) into (4.17) and noting (4.2), we obtain

$$\frac{\delta_i^+}{\delta_j^+} = \frac{(1+\beta)^{3/2}(1+2\beta)}{(1-\beta)^{5/2}(1-2\beta)^2} \frac{\tilde{s}_i}{\tilde{s}_j} \leq \sqrt{n} \frac{4(1+\beta)^2(1+2\beta)}{\beta(1-\beta)^3(1-2\beta)^2} \frac{\delta_i}{\delta_j} \varepsilon_\infty^a = \frac{\sqrt{n}}{\tau} \frac{\delta_i}{\delta_j} \varepsilon_\infty^a. \quad \square$$

The next result gives another alternative sufficient condition for the occurrence of $\mathcal{C}$ -crossover events. By the definition of the quantity $\text{gap}(\mathcal{J}(w))$, there exists two indices i and j , one lying in $B(w)$ and the other in $N(w)$, such that $\text{gap}(\mathcal{J}(w)) = \delta_i(w)/\delta_j(w)$. The lemma below considers the situation where $\text{gap}(\mathcal{J}(w)) = \delta_i(w)/\delta_j(w)$ for some $i \in B(w)$ and $j \in N(w)$.

LEMMA 4.9. *Suppose $\beta \in (0, 1/4]$ and $w = w^k \in \mathcal{N}(\beta)$ is an iterate of the MTY P-C algorithm. Assume that $\text{gap}(\mathcal{J}(w)) = \delta_i(w)/\delta_j(w)$ for some $i \in B(w)$ and $j \in$*

$N(w)$. Then there exists an iteration index $l > k$ such that $l - k = \mathcal{O}(n^{1.5} \log(\bar{\chi}_A + n))$ and with the property that either a $\mathcal{C}$ -crossover event occurs between w^k and w^l or the algorithm terminates at or before the l th iteration.

Proof. Let $\mathcal{J} \equiv \mathcal{J}(w)$. If either condition $\text{gap}(\mathcal{J}) \geq \bar{g}$ or $\varepsilon_\infty^a(w) \leq \tau \bar{g} / (\sqrt{n} \text{gap}(\mathcal{J}))$ does not hold, then the conclusion of the lemma follows immediately from Lemmas 4.2 and 4.5. Assume then that $\text{gap}(\mathcal{J}) \geq \bar{g}$ and $\varepsilon_\infty^a(w) \leq \tau \bar{g} / (\sqrt{n} \text{gap}(\mathcal{J}))$. By assumption, there exist $i \in B(w)$ and $j \in N(w)$ such that $\delta_i(w) / \delta_j(w) = \text{gap}(\mathcal{J}) \geq \bar{g}$. The assumptions on $\text{gap}(\mathcal{J})$ and $\varepsilon_\infty^a(w)$ together with (4.6) imply that $\varepsilon_\infty^a(w) \leq \sqrt{1 - 2\beta}/4 \leq \sqrt{1 - \beta}/2$. Hence, by Lemma 4.8 and the assumption on $\varepsilon_\infty^a(w)$, we have

$$(4.20) \quad \frac{\delta_i(w^+)}{\delta_j(w^+)} \leq \frac{\sqrt{n} \delta_i(w)}{\tau \delta_j(w)} \varepsilon_\infty^a(w) = \frac{\sqrt{n}}{\tau} \text{gap}(\mathcal{J}) \varepsilon_\infty^a(w) \leq \bar{g},$$

where $w^+ \equiv w^{k+1}$. The conclusion of the lemma now follows from Lemma 4.6(b). $\square$

In order to prove our main lemma, namely Lemma 4.11, followed by the proof of Theorem 2.6, we introduce the set $\mathcal{K}$ consisting of those indices k such that the iterates w^k and w^{k+1} of the MTY P-C algorithm satisfy the following set of conditions, where $\mathcal{J}_l \equiv \mathcal{J}(w^l)$ for all $l \geq 0$:

- (C1) $\text{gap}(\mathcal{J}_k) > 2\bar{g}$ and $\varepsilon_\infty^a(w^k) < \tau \bar{g} / (\sqrt{n} \text{gap}(\mathcal{J}_k))$. (Note that if (C1) does not hold, then $w = w^k$ satisfies the hypotheses of either Lemma 4.2 or 4.5.)
- (C2) $\text{gap}(\mathcal{J}_{k+1}) = \delta_i(w^{k+1}) / \delta_j(w^{k+1})$ for some $i \in N(w^{k+1})$ and $j \in B(w^{k+1})$. (Note that if (C2) does not hold, then $w = w^{k+1}$ satisfies the hypotheses of Lemma 4.9.)
- (C3) $(B(w^{k+1}), N(w^{k+1})) = (B(w^k), N(w^k))$ and $\mathcal{J}_{k+1} = \mathcal{J}_k$. (Note that if (C3) does not hold, then the observation made in the second paragraph after Lemma 4.6 implies that $w = w^k$ satisfies the hypotheses of either Lemma 4.2 or 4.6.)

Note that if $k \notin \mathcal{K}$, then a $\mathcal{C}$ -crossover event must occur after $\mathcal{O}(\log(\bar{\chi}_A + n))$ iterations, since $w = w^k$ satisfies the hypotheses of either Lemma 4.2, 4.5, or 4.6, or w^{k+1} satisfies the hypotheses of Lemma 4.9. The following result establishes a number of interesting properties regarding the above set $\mathcal{K}$.

LEMMA 4.10. *For every $k \in \mathcal{K}$, we have*

$$(4.21) \quad \frac{\text{gap}(\mathcal{J}_{k+1})}{\bar{g}} \geq \left(\frac{\text{gap}(\mathcal{J}_k)}{\bar{g}} \right)^2, \quad \frac{\mu(w^k)}{\mu(w^{k+1})} \geq \left(\frac{\text{gap}(\mathcal{J}_k)}{\bar{g}} \right) \geq 2.$$

Moreover, for some given constant $\eta > 0$, if $l, l+1, \dots, \hat{l}-1, \hat{l}$ are consecutive indices in the set $\mathcal{K}(\eta) \equiv \{k \in \mathcal{K} : \mu(w^{k+1}) > \eta\}$, then $\hat{l} - l = \mathcal{O}(\log(\log(\mu(w^0)/\eta)))$.

Proof. Fix some $k \in \mathcal{K}$. Then, conditions (C1), (C2), and (C3) above hold for this k . In particular, by conditions (C2) and (C3), we know that there exist $i \in N(w^k)$ and $j \in B(w^k)$ such that

$$\text{gap}(\mathcal{J}_{k+1}) = \frac{\delta_i(w^{k+1})}{\delta_j(w^{k+1})}.$$

Using (4.6) and condition (C1), we easily see that w^k satisfies the hypothesis of Lemma 4.8. Hence, it follows from the above identity, Lemma 4.8, and condition (C1) that

$$\text{gap}(\mathcal{J}_{k+1}) \geq \frac{\tau}{\sqrt{n} \varepsilon_\infty^a(w^k)} \frac{\delta_i(w^k)}{\delta_j(w^k)} \geq \frac{\tau}{\sqrt{n} \varepsilon_\infty^a(w^k)} \text{gap}(\mathcal{J}_k) \geq \frac{1}{\bar{g}} [\text{gap}(\mathcal{J}_k)]^2,$$

where the second inequality uses the fact that $\mathcal{J}_k = \mathcal{J}_{k+1}$, which holds in view of (C3). The first inequality in (4.21) follows by dividing the above inequality by $\bar{g}$. Using Lemma 2.5, condition (C1), and the fact that $\tau \leq \beta$, we obtain

$$(4.22) \quad \frac{\mu(w^k)}{\mu(w^{k+1})} \geq \frac{\beta}{\sqrt{n}\varepsilon_\infty^a(w^k)} \geq \frac{\beta \operatorname{gap}(\mathcal{J}_k)}{\tau \bar{g}} \geq \frac{\operatorname{gap}(\mathcal{J}_k)}{\bar{g}} > 2 \quad \forall k \in \mathcal{K},$$

showing that the second relation in (4.21) also holds. To show the last statement of the lemma, assume now that $l, l+1, \dots, \hat{l}$ are consecutive indices in $\mathcal{K}(\eta)$. In view of the first relation in (4.21), we have

$$\log \left(\frac{\operatorname{gap}(\mathcal{J}_{k+1})}{\bar{g}} \right) \geq 2 \log \left(\frac{\operatorname{gap}(\mathcal{J}_k)}{\bar{g}} \right) \quad \forall k = l, l+1, \dots, \hat{l},$$

from where it follows that

$$\log \left(\frac{\operatorname{gap}(\mathcal{J}_{\hat{l}})}{\bar{g}} \right) \geq 2^{\hat{l}-l} \log \left(\frac{\operatorname{gap}(\mathcal{J}_l)}{\bar{g}} \right) \geq 2^{\hat{l}-l} \log 2.$$

Using the fact that $\hat{l} \in \mathcal{K}(\eta)$, relation (4.22) with $k = \hat{l}$, and the previous inequality, we obtain

$$\log \frac{\mu(w^0)}{\eta} \geq \log \frac{\mu(w^{\hat{l}})}{\mu(w^{\hat{l}+1})} \geq \log \left(\frac{\operatorname{gap}(\mathcal{J}_{\hat{l}})}{\bar{g}} \right) \geq 2^{\hat{l}-l} \log 2,$$

from which the last statement of the lemma immediately follows. $\square$

We are now in a position to establish the main lemma of this section. Even though the main goal of this lemma is to cover the case where the hypotheses of neither Lemma 4.2 nor Lemma 4.5 hold, this result imposes no condition on the iterate w of the MTY P-C algorithm. Moreover, in contrast to the bounds of Lemmas 4.2 and 4.5, the bound derived in this lemma on the number of iterations needed to guarantee the occurrence of a $\mathcal{C}$ -crossover event depends not only on the quantity $n^{1.5} \log(\bar{\chi}_A + n)$ but also on the term $\log(\log(\mu(w^0)/\eta))$.

LEMMA 4.11. *Suppose $\beta \in (0, 1/4]$ and $w = w^k \in \mathcal{N}(\beta)$ is an iterate of the MTY P-C algorithm. Then, there exists an iteration index $l > k$ such that*

$$(4.23) \quad l - k = \mathcal{O}(n^{1.5} \log(\bar{\chi}_A + n) + \log(\log(\mu(w^0)/\eta)))$$

and with the property that either a $\mathcal{C}$ -crossover event occurs between w^k and w^l or the algorithm terminates at or before the l th iteration. Furthermore, we have

$$(4.24) \quad l - k = \mathcal{O}(n^{1.5} \log(\bar{\chi}_A + n) + \log(\mu(w^k)/\mu(w^l))).$$

Proof. Consider the set $\mathcal{K}(\eta)$ defined in Lemma 4.10 and let $\hat{k}$ be the first index greater than or equal to k such that $k \notin \mathcal{K}(\eta)$. Note that by Lemma 4.10, such an index $\hat{k}$ exists and satisfies $\hat{k} - k = \mathcal{O}(\log(\log(\mu(w^0)/\eta)))$. The condition $\hat{k} \notin \mathcal{K}(\eta)$ means that either $\mu(w^{\hat{k}+1}) \leq \eta$ or $\hat{k} \notin \mathcal{K}$. If the first condition holds, then the conclusion of the lemma obviously holds with $l = \hat{k} + 1$. (The algorithm terminates at the l th iteration in this case.) If the latter condition holds, it follows that $w = w^{\hat{k}}$ satisfies the hypotheses of either Lemma 4.2, 4.5, or 4.6, or that $w^{\hat{k}+1}$ satisfies the hypotheses of Lemma 4.9. In any of these cases, we know that there exists an iteration index $l > \hat{k}$

such that $l - \hat{k} = \mathcal{O}(n^{1.5} \log(\bar{\chi}_A + n))$ and with the property that either a $\mathcal{C}$ -crossover event occurs between $w^{\hat{k}}$ and w^l or the algorithm terminates at or before the l th iteration. Clearly, such an iteration index l satisfies the first conclusion of the lemma. To show that (4.24) also holds, it suffices to show that $\hat{k} - k = \mathcal{O}(\log(\mu(w^k)/\mu(w^l)))$. Indeed, since this conclusion obviously holds when $\hat{k} = k$, we may assume that $\hat{k} > k$. This implies that the indices $k, k+1, \dots, \hat{k}-1$ are in $\mathcal{K}$, and hence, by the second relation in (4.22), we conclude that

$$\log \frac{\mu(w^k)}{\mu(w^l)} \geq \log \frac{\mu(w^k)}{\mu(w^{\hat{k}})} = \sum_{j=k}^{\hat{k}-1} \log \frac{\mu(w^j)}{\mu(w^{j+1})} \geq (\hat{k} - k) \log 2,$$

which clearly implies our claim above. $\square$

We are now ready to give the proof of Theorem 2.6.

Proof of Theorem 2.6. Let k^* be the first index such that $\mu(w^{k^*}) \leq \eta$. Using Lemma 4.11 inductively starting from w^0 , we can partition the set $\{0, 1, \dots, k^*\}$ into disjoint subset of consecutive indices $K_1 := \{k_0, k_0 + 1, \dots, k_1 - 1\}$, $K_2 := \{k_1, k_1 + 1, \dots, k_2 - 1\}$, $\dots$, $K_r := \{k_{r-1}, k_{r-1} + 1, \dots, k_r - 1\}$ such that $k_0 = 0$, $k_r = k^*$, and satisfying the following: (i) a $\mathcal{C}$ -crossover event occurs between $w^{k_{i-1}}$ and w^{k_i-1} for $i = 1, \dots, r-1$; and (ii) the cardinality of K_i satisfies

$$|K_i| = \mathcal{O} \left(\min \left\{ \log \left(\log \frac{\mu^0}{\eta} \right), \log \left(\frac{\mu^{k_{i-1}}}{\mu^{k_i}} \right) \right\} + n^{1.5} \log(\bar{\chi}_A + n) \right) \quad \forall i = 1, \dots, r.$$

Since the number of crossover event is $\mathcal{O}(n^2)$, it follows that $r = \mathcal{O}(n^2)$. These observations, (1.3), and the fact that $k^* = \sum_{i=1}^r |K_i|$ clearly imply that $k^* = \mathcal{O}(T(\mu_0/\eta) + n^{3.5} \log(\bar{\chi}_A + n))$. $\square$

5. Implications of the main result under the Turing machine model.

In section 2, we have presented two iteration-complexity results for the MTY P-C algorithm under the real number computational model. In this section, we discuss a few implications of these results under the Turing machine model. In this model, the entries of the data (A, b, c) are rational numbers of finite bit length and the arithmetic operations on these types of numbers are carried out approximately, i.e., in finite precision. For the sake of simplicity, here we just focus on the number of arithmetic operations and do not deal with the issue of finite precision. However, we observe that the issue of finite precision in the context of the MTY interior-point algorithm can be dealt with using standard and well-known techniques developed in the context of other interior-point methods (e.g., see [20]).

Let L and L_A be the input size of (A, b, c) and A , respectively. It is well known that $\bar{\chi}_A^* \leq \bar{\chi}_A \leq 2^{\mathcal{O}(L_A)}$ and $\xi(A, b, c) \geq 2^{-\mathcal{O}(L)}$. As was mentioned before, we have $\zeta(A, (B_*, N_*)) \leq \bar{\chi}_A$. Therefore, we have the following corollary which immediately follows from Theorem 2.9.

COROLLARY 5.1. *Assume that the data (A, b, c) is integral, and let L and L_A be defined as in the paragraph above. For some $\beta \in (0, 1/4]$, suppose that a point $w^0 \in \mathcal{N}(\beta)$ such that $\mu(w^0) = 2^{\mathcal{O}(L)}$ is given. Then, the version of the MTY P-C algorithm, in which the FT procedure is invoked at every iterate w^k , started from w^0 finds a primal-dual strictly complementary optimal solution w^* in at most*

$$(5.1) \quad \mathcal{O}(\min\{\sqrt{n}L, \min\{L, n^2 \log L\} + n^{3.5}(L_A + \log n)\})$$

iterations.

Corollary 5.1 assumes that the initial iterate of the MTY P-C algorithm is a well-centered strictly feasible point whose duality gap is not too large. For a general dual pair of linear programs, even if such a point exists, computing this point is as hard as solving the pair of LP problems. In such a case, an auxiliary pair of dual LP problems is constructed whose optimal solution yields that of the original pair of LP problems. Using the big M idea, Vavasis and Ye [30] construct an auxiliary pair of LP problems, which we refer to as the *VY auxiliary LP pair*, associated with the pair of problems (1.1) and (1.2) in order to resolve the initialization issue for their LLS step algorithm. The VY auxiliary LP pair has the following properties: (i) the input size of its coefficient matrix is bounded by $\mathcal{O}(L_A)$; (ii) the sizes of its cost and right-hand coefficients are bounded by $\mathcal{O}(L)$; (iii) it admits a readily available well-centered initial point whose duality gap is $n2^{\mathcal{O}(L)}$; and (iv) if (1.1) and (1.2) have a primal-dual optimal solution, then this solution can be easily obtained from an optimal solution of the VY auxiliary LP pair. Therefore, we obtain the following theorem for solving a general pair of LP problems under the Turing machine model.

THEOREM 5.2. *Assume that the data (A, b, c) is integral, and suppose that (1.1) and (1.2) have a primal-dual optimal solution. Then, the MTY P-C algorithm, with the FT procedure invoked at every iteration, applied to the VY auxiliary LP pair, finds a strict complementary primal-dual optimal solution of (1.1) and (1.2) in a number of iterations bounded by (5.1).*

6. Concluding remarks. In this paper, we have developed a new iteration-complexity bound for the MTY P-C algorithm using the notion of crossover events due to Vavasis and Ye [30]. In contrast to the iteration-complexity bound developed in [30], ours is scaling-invariant and has an extra but relatively small term, namely, the term $T(\mu_0/\eta)$, where $T(\cdot)$ is defined in (1.3).

Note that the second bound in (2.12), i.e., the inequality $\mu(w^+)/\mu(w) \leq \varepsilon_\infty^a(w)\sqrt{n}/\beta$, plays an important role in our analysis. This is also the inequality which plays an important role in establishing that the sequence $\{\mu(w^k)\}$ generated by the MTY P-C algorithm is quadratically convergent. Observe also that if the sequence $\{\mu(w^k)\}$ generated by a fictitious algorithm satisfied $\mu(w^{k+1}) \leq C_1\mu(w^k)^2$ for all k , where $C_1 = \mathcal{O}(1)$, then its iteration-complexity bound would be $\mathcal{O}(\log(\log(\mu(w^0)/\eta)))$. The term $n^2 \log(\log(\mu(w^0)/\eta))$ which appears in our iteration-complexity bound is due to the fact that this type of quadratic convergence happens over $\mathcal{O}(n^2)$ disjoint finite sets of consecutive iteration indices. A natural conjecture is that any primal-dual interior-point algorithm which achieves the duality gap reduction given by (2.12) at every iteration has the same iteration-complexity bound as the one obtained in this paper. More generally, we conjecture that all interior-point algorithms whose corresponding sequence $\{\mu(w^k)\}$ converges superlinearly or quadratically are suitable for the same type of analysis performed in this paper.

Note that the iteration-complexity bound obtained in this paper under the real-computation model is with respect to a pair of dual LP problems satisfying Assumptions A.1 and A.2. For a pair of dual LP problems which does not satisfy A.1 or A.2, a natural open question is whether one can develop the same type of scaling-invariant iteration-complexity bound under the real-computation model obtained in this paper. Observe that one of the difficulties in this context is the proper choice of the big M constant in the VY auxiliary LP pair.

Our work was strongly motivated by the work of Vavasis and Ye [30]. Therefore, we wonder whether the MTY P-C algorithm with the FT procedure has the iteration-complexity bound $\mathcal{O}(n^{3.5} \log(\bar{\chi}_A^* + n))$, i.e., the iteration-complexity bound of Vavasis

and Ye's algorithm but with $\bar{\chi}_A$ replaced by $\bar{\chi}_A^*$. The term involving the two log operators in our iteration-complexity bound is due to the possibility that the assumption of Lemma 4.9 is violated. When this assumption is violated, we have shown that the gap reduces quadratically until some suitable conditions are met in $\mathcal{O}(\log(\log(\mu(w^0)/\eta)))$ iterations which guarantee the occurrence of a $\mathcal{C}$ -crossover event. A very challenging open question is whether these conditions are met in $\mathcal{O}(p(n, \log \bar{\chi}_A))$ iterations, where $p(x_1, x_2)$ is a polynomial in $\mathbb{R}^2$. Note that this would provide a scaling-invariant iteration-complexity bound for the MTY P-C algorithm depending on n and $\bar{\chi}_A^*$ only.

Finally, another interesting topic for future research is whether the results obtained in this paper can be extended, possibly under some extra assumptions, to more general classes of problems such as convex quadratic programming or semidefinite programming.

Acknowledgments. The authors would like to thank the two anonymous referees and the associate editor for their constructive comments which have helped to improve the presentation of the paper. In particular, one of the referees found a few minor errors in the original proofs of Lemmas 4.6 and 4.8, which have been corrected in this present version.

This work was done while the first author was visiting Professor Cornelis Roos at the Department of Information Systems and Algorithms of Delft University of Technology during the summer of 2001; while visiting the second author at the Institute of Statistical Mathematics in August of 2002 and October of 2003; and while the second author was visiting the first author at the School of Industrial and Systems Engineering in June and September of 2002. They are grateful to their host institutions for the financial support and congenial scientific atmosphere provided during these visits.

REFERENCES

- [1] I. I. DIKIN, *On the speed of an iterative process*, Upr. Sist., 12 (1974), pp. 54–60 (in Russian).
- [2] A. S. EL-BAKRY, R. A. TAPIA, AND Y. ZHANG, *A study of indicators for identifying zero variables in interior-point methods*, SIAM Rev., 36 (1994), pp. 45–72.
- [3] C. C. GONZAGA, *Path-following methods for linear programming*, SIAM Rev., 34 (1992), pp. 167–224.
- [4] C. C. GONZAGA AND H. J. LARA, *A note on properties of condition numbers*, Linear Algebra Appl., 261 (1997), pp. 269–273.
- [5] O. GÜLER AND Y. YE, *Convergence behavior of interior-point algorithms*, Math. Program., 60 (1993), pp. 215–228.
- [6] N. KARMARKAR, *A new polynomial-time algorithm for linear programming*, Combinatorica, 4 (1984), pp. 373–395.
- [7] M. KOJIMA, S. MIZUNO, AND A. YOSHISE, *A primal-dual interior point algorithm for linear programming*, in Progress in Mathematical Programming, Interior-Point and Related Methods, N. Megiddo, ed., Springer-Verlag, New York, 1989, pp. 29–47.
- [8] M. KOJIMA, S. MIZUNO, AND A. YOSHISE, *A polynomial-time algorithm for a class of linear complementarity problems*, Math. Program., 44 (1989), pp. 1–26.
- [9] L. McLINDEN, *The complementarity problem for maximal monotone multifunction*, in Variational Inequalities and Complementarity Problems, R. W. Cottle, F. Giannessi, and J. L. Lions, eds., John Wiley, New York, 1980, pp. 251–270.
- [10] N. MEGIDDO, *Pathways to the optimal set in linear programming*, in Progress in Mathematical Programming (Pacific Grove, CA, 1987), Springer, New York, Berlin, 1989, pp. 131–158.
- [11] N. MEGIDDO, S. MIZUNO, AND T. TSUCHIYA, *A modified layered-step interior-point algorithm for linear programming*, Math. Program., 82 (1998), pp. 339–355.
- [12] S. MEHROTRA AND Y. YE, *Finding an interior point in the optimal face of linear programs*, Math. Program., 62 (1993), pp. 497–515.
- [13] S. MIZUNO, M. J. TODD, AND Y. YE, *On adaptive-step primal-dual interior-point algorithms for linear programming*, Math. Oper. Res., 18 (1993), pp. 964–981.
- [14] S. MIZUNO, N. MEGIDDO, AND T. TSUCHIYA, *A linear programming instance with many crossover events*, J. Complexity, 12 (1996), pp. 474–479.

- [15] R. D. C. MONTEIRO AND I. ADLER, *Interior path following primal-dual algorithms. Part I: Linear programming*, Math. Program., 44 (1989), pp. 27–41.
- [16] R. D. C. MONTEIRO AND I. ADLER, *Interior path-following primal-dual algorithms. Part II: Convex quadratic programming*, Math. Program., 44 (1989), pp. 43–66.
- [17] R. D. C. MONTEIRO, I. ADLER, AND M. G. C. RESENDE, *A polynomial-time primal-dual affine scaling algorithm for linear and convex quadratic programming and its power series extension*, Math. Oper. Res., 15 (1990), pp. 191–214.
- [18] R. D. C. MONTEIRO AND T. TSUCHIYA, *Global convergence of the affine scaling algorithm for convex quadratic programming*, SIAM J. Optim., 8 (1998), pp. 26–58.
- [19] R. D. C. MONTEIRO AND T. TSUCHIYA, *A variant of the Vavasis–Ye layered-step interior-point algorithm for linear programming*, SIAM J. Optim., 13 (2003), pp. 1054–1079.
- [20] J. RENEGAR, *A polynomial-time algorithm, based on Newton’s method, for linear programming*, Math. Program., 40 (1988), pp. 59–93.
- [21] G. W. STEWART, *On scaled projections and pseudoinverses*, Linear Algebra Appl., 112 (1989), pp. 189–193.
- [22] K. TANABE, *Centered Newton method for mathematical programming*, in System Modeling and Optimization, M. Iri and K. Yajima, eds., Springer-Verlag, Berlin, 1988, pp. 197–206.
- [23] É. TARDOS, *A strongly polynomial algorithm to solve combinatorial linear programs*, Oper. Res., 34 (1986), pp. 250–256.
- [24] M. J. TODD, *A Dantzig-Wolfe-like variant of Karmarkar’s interior-point linear programming algorithm*, Oper. Res., 38 (1990), pp. 1006–1018.
- [25] M. J. TODD, L. TUNÇEL, AND Y. YE, *Characterizations, bounds, and probabilistic analysis of two complexity measures for linear programming problems*, Math. Program., 90 (2001), pp. 59–70.
- [26] T. TSUCHIYA, *Global convergence property of the affine scaling method for primal degenerate linear programming problems*, Math. Oper. Res., 17 (1992), pp. 527–557.
- [27] L. TUNÇEL, *Approximating the complexity measure of Vavasis–Ye algorithm is NP-hard*, Math. Program., 86 (1999), pp. 219–223.
- [28] L. TUNÇEL, *On the condition numbers for polyhedra in Karmarkar’s form*, Oper. Res. Lett., 24 (1999), pp. 149–155.
- [29] R. J. VANDERBEI AND J. C. LAGARIAS, *I. I. Dikin’s convergence result for the affine-scaling algorithm*, in Mathematical Developments Arising from Linear Programming, Contemp. Math. 114, AMS, Providence, RI, 1990, pp. 109–119.
- [30] S. VAVASIS AND Y. YE, *A primal-dual accelerated interior point method whose running time depends only on A*, Math. Program., 74 (1996), pp. 79–120.
- [31] S. VAVASIS AND Y. YE, *A simplification to “a primal-dual interior point method whose running time depends only on the constraint matrix,”* in High Performance Optimization, Appl. Optim. 33, Kluwer, Dordrecht, 2000, pp. 233–243.
- [32] S. VAVASIS AND Y. YE, *On the relationship between layered least squares and affine scaling steps*, in The Mathematics of Numerical Analysis, Lectures in Appl. Math. 32, AMS, Providence, RI, 1996, pp. 857–865.
- [33] Y. YE, *On the finite convergence of interior-point algorithms for linear programming*, Math. Program., 57 (1992), pp. 325–335.
- [34] Y. YE, *Interior-Point Algorithms: Theory and Analysis*, John Wiley and Sons, New York, 1997.

ERROR BOUNDS AND LIMITING BEHAVIOR OF WEIGHTED PATHS ASSOCIATED WITH THE SDP MAP $X^{1/2}SX^{1/2}$ *

ZHAOSONG LU[†] AND RENATO D. C. MONTEIRO[†]

Abstract. This paper studies the limiting behavior of weighted infeasible central paths for semidefinite programming (SDP) obtained from centrality equations of the form $X^{1/2}SX^{1/2} = \nu W$, where W is a fixed positive definite matrix and $\nu > 0$ is a parameter, under the assumption that the problem has a strictly complementary primal-dual optimal solution. It is shown that a weighted central path as a function of $\sqrt{\nu}$ can be extended analytically beyond 0 and hence that the path converges as $\nu \downarrow 0$. Characterization of the limit points of the path and its normalized first-order derivatives are also provided. As a consequence, it is shown that a weighted central path can have two types of behavior: it converges either as $\Theta(\nu)$ or as $\Theta(\sqrt{\nu})$ depending on whether the matrix W on a certain scaled space is block diagonal or not, respectively. We also derive an error bound on the distance between a point lying in a certain neighborhood of the central path and the set of primal-dual optimal solutions. Finally, in light of the results of this paper, we give a characterization of a sufficient condition proposed by Potra and Sheng which guarantees the superlinear convergence of a class of primal-dual interior-point SDP algorithms.

Key words. limiting behavior, weighted central path, error bound, superlinear convergence, semidefinite programming (SDP)

AMS subject classifications. 90C22, 90C25, 90C30, 65K05

DOI. 10.1137/S1052623403430828

1. Introduction. Let $\mathcal{S}^n$ denote the space of $n \times n$ real symmetric matrices. We consider the semidefinite programming (SDP) problem

$$(P) \quad \begin{aligned} & \text{minimize} && C \bullet X \\ & \text{subject to} && \mathcal{A}X = b, \\ & && X \succeq 0, \end{aligned}$$

and its associated dual SDP problem

$$(D) \quad \begin{aligned} & \text{maximize} && b^T y \\ & \text{subject to} && \mathcal{A}^* y + S = C, \\ & && S \succeq 0, \end{aligned}$$

where the data consists of $C \in \mathcal{S}^n$, $b \in \mathbb{R}^m$, and a linear operator $\mathcal{A} : \mathcal{S}^n \rightarrow \mathbb{R}^m$, the primal variable is $X \in \mathcal{S}^n$, and the dual variable consists of $(S, y) \in \mathcal{S}^n \times \mathbb{R}^m$. For a matrix $V \in \mathcal{S}^n$, the notation $V \succeq 0$ means that V is positive semidefinite. Given a fixed positive definite matrix $W \in \mathcal{S}^n$, $\Delta b \in \mathbb{R}^m$, and $\Delta C \in \mathcal{S}^n$, our interest in this paper is to study the set of solutions of the following system of nonlinear equations parametrized by the parameter $\nu > 0$:

$$(3) \quad \mathcal{A}X = b + \nu \Delta b, \quad X \succ 0,$$

*Received by the editors July 3, 2003; accepted for publication (in revised form) March 18, 2004; published electronically December 30, 2004.

<http://www.siam.org/journals/siopt/15-2/43082.html>

[†]School of ISyE, Georgia Institute of Technology, Atlanta, GA 30332 (zhaosong@isye.gatech.edu, monteiro@isye.gatech.edu). The authors were supported in part by NSF grants CCR-9902010, CCR-0203113, and INT-9910084 and ONR grant N00014-03-1-0401.

$$(4) \quad \mathcal{A}^* y + S = C + \nu \Delta C, \quad S \succ 0,$$

$$(5) \quad X^{1/2} S X^{1/2} = \nu W.$$

Under suitable conditions on $(W, \Delta C, \Delta b)$, it has been shown in Monteiro and Zanjácomo [31] that the above system has a unique solution, denoted by $p(\nu) \equiv (X(\nu), S(\nu), y(\nu))$ for every $\nu \in (0, 1]$. We refer to the path $\nu \in (0, 1] \rightarrow p(\nu)$ as the $(W, \Delta C, \Delta b)$ -weighted central path associated with (P) and (D) . The main objective of this paper is to analyze the limiting behavior of this path as $\nu \downarrow 0$.

When $(W, \Delta C, \Delta b) = (I, 0, 0)$, the path $\nu \in (0, 1] \rightarrow p(\nu)$ is a part of the central path associated with (P) and (D) . Properties of the central path have been extensively studied in several papers due to the important role it plays in the development of interior-point algorithms for cone programming, nonlinear programming, and complementarity problems. Early works dealing with the well-definedness, differentiability, and limiting behavior of weighted central paths in the context of the linear programming and monotone complementarity problems include [1, 2, 3, 8, 9, 10, 11, 16, 22, 23, 24, 26, 27, 29, 32, 36, 37, 38, 39, 40].

Using the fact that every real algebraic variety has a triangulation, Kojima et al. [15] showed that the central path associated with a monotone linear complementarity problem converges to a solution. In [19], Kojima, Shindoh, and Hara claim that similar arguments as the ones used in [15] can also be used to show that the central path of a monotone linear semidefinite complementarity problem (which is equivalent to SDP) converges to a solution of the problem. More generally, Drummond and Peterzil [8] established convergence of the central path for analytic convex nonlinear SDP problems. An alternative proof based on a deep result from algebraic geometry (see, for example, Lemma 3.1 of Milnor [25]) of the convergence of the central path for an SDP problem was given by Halická, de Klerk, and Roos [14]. Characterization of the limit point of the central path has been obtained by De Klerk, Roos, and Terlaky [6] and Luo, Sturm, and Zhang [21] for SDP problems possessing strictly complementary primal-dual optimal solutions. Using an approach based on the implicit function theorem described in Stoer and Wechs [37, 38], Halická [12] showed that the central path of an SDP problem possessing a strictly complementary primal-dual optimal solution can be extended analytically as a function of $\nu > 0$ to $\nu = 0$. For more general SDP problems, the above issues regarding the central path still remain open but some advances have been made in a few papers. These include De Klerk, Roos, and Terlaky [5] and Goldfarb and Scheinberg [7] who proved that any cluster point of the central path must be a maximally complementary optimal solution. Also, Halická, de Klerk, and Roos [13] and Sporre and Forsgren [36] provided partial characterizations of the limit point of the central path as being the analytic center of some convex subset of the optimal solution set and the unique solution of a perturbed log barrier problem over the optimal solution set, respectively. Finally, the recent paper by Cruz Neto, Ferreira, and Monteiro [4], which appeared after the release of the first version of the present work, establishes the convergence of the central path for a special class of SDPs which do not satisfy the strict complementarity condition.

Generalization of the notion of weighted central paths from linear programming to SDP problems is a delicate issue. While for linear programming a weighted central path can be characterized as optimal solutions of certain weighted logarithmic barrier problems, this characterization does not seem to be a good source for obtaining a suitable notion of weighted central paths for SDP. Instead, Monteiro and Zanjácomo [31] (see also Monteiro and Pang [28]) work directly with a system consisting of (3), (4), and an equation of the form $\Phi(X, S) = \nu W$ for some suitable map $\Phi : D \subseteq$

$\mathcal{S}^n \times \mathcal{S}^n \rightarrow \mathcal{S}^n$ and show that this system has a unique solution for every $\nu \in (0, 1]$. Special instances of the map Φ for which the above result applies include the maps $(X, S) \rightarrow (XS + SX)/2$ and $(X, S) \rightarrow X^{1/2}SX^{1/2}$.

Independently of the present work, Preiß and Stoeer [35] have proved that the weighted central paths associated with the map $(X, S) \rightarrow (XS + SX)/2$ are analytically extendible as functions of $\nu \in (0, 1]$ to $\nu = 0$ (see also Lu and Monteiro [20] for another proof of this result). In this paper, we will be interested only in the second map and its corresponding weighted central paths, i.e., the path of solutions of systems of the form (3)–(5). More specifically, we will investigate the asymptotic properties of the weighted central paths $\nu \in (0, 1] \rightarrow p(\nu)$ and their derivatives for the special class of SDPs possessing strictly complementary primal-dual optimal solutions. Using a suitable change of variables together with the technique described in [37, 38] based on the implicit function theorem, we prove in section 3 that the path $t \in (0, 1] \rightarrow p(t^2)$ can be extended analytically to $t = 0$, and we also characterize the limit point of $p(\nu)$ as $\nu \downarrow 0$. In section 4, we characterize the limit of the normalized derivative $p'(\nu)/\|p(\nu)\|$ as $\nu \downarrow 0$. As a consequence, we show that a weighted central path can have two types of behavior: it converges either as $\Theta(\nu)$ or as $\Theta(\sqrt{\nu})$, depending on whether the matrix W on a certain scaled space is block diagonal or not, respectively. Using these results, we derive in section 5 an error bound on the distance between a point lying in a certain neighborhood of the central path and the set of primal-dual optimal solutions. Finally, we consider in section 6 a sufficient condition proposed by Potra and Sheng [33], which guarantees the superlinear convergence of a large class of primal-dual interior-point SDP algorithms, and we obtain a characterization of it in terms of the results obtained in this paper.

The organization of this paper is as follows. Section 2 introduces the assumptions made throughout the paper and discusses some preliminary known results about weighted central paths. Sections 3–6 establish the results mentioned in the previous paragraph. Finally, we end the paper by providing some concluding remarks in section 7.

1.1. Notation. The space of symmetric $n \times n$ matrices will be denoted by $\mathcal{S}^n$. Given matrices X and Y in $\mathbb{R}^{p \times q}$, the standard inner product is defined by $X \bullet Y \equiv \text{tr}(X^T Y)$, where $\text{tr}(\cdot)$ denotes the trace of a matrix. The Euclidean norm and its associated operator norm, i.e., the spectral norm, are both denoted by $\|\cdot\|$. The Frobenius norm of a $p \times q$ matrix X is defined as $\|X\|_F \equiv \sqrt{X \bullet X}$. Given a point f and a set F in a finite dimensional normed vector space, the distance from f to F is defined as $\text{dist}(f, F) \equiv \inf_{\tilde{f} \in F} \|f - \tilde{f}\|$. If $X \in \mathcal{S}^n$ is positive semidefinite (resp., definite), we write $X \succeq 0$ (resp., $X \succ 0$). The cone of positive semidefinite (resp., definite) matrices is denoted by $\mathcal{S}_+^n$ (resp., $\mathcal{S}_{++}^n$). Either the identity matrix or operator will be denoted by I . The image (or range) space of a linear operator $\mathcal{A}$ will be denoted by $\text{Im}(\mathcal{A})$; the dimension of the subspace $\text{Im}(\mathcal{A})$, referred to as the rank of $\mathcal{A}$, will be denoted by $\text{rank}(\mathcal{A})$. Given a linear operator $\mathcal{F} : E \rightarrow F$ between two finite dimensional inner product spaces $(E, \langle \cdot, \cdot \rangle_E)$ and $(F, \langle \cdot, \cdot \rangle_F)$, its *adjoint* is the unique operator $\mathcal{F}^* : F \rightarrow E$ satisfying $\langle \mathcal{F}(u), v \rangle_F = \langle u, \mathcal{F}^*(v) \rangle_E$ for all $u \in E$ and $v \in F$.

Given functions $f : \Omega \rightarrow E$ and $g : \Omega \rightarrow \mathbb{R}_{++}$, where Ω is an arbitrary set and E is a normed vector space, and a subset $\tilde{\Omega} \subset \Omega$, we write $f(w) = \mathcal{O}(g(w))$ for all $w \in \tilde{\Omega}$ to mean that there exists $M \geq 0$ such that $\|f(w)\| \leq M g(w)$ for all $w \in \tilde{\Omega}$; moreover, for a function $U : \Omega \rightarrow \mathcal{S}_{++}^n$, we write $U(w) = \Theta(g(w))$ for all $w \in \tilde{\Omega}$ if $U(w) = \mathcal{O}(g(w))$ and $U(w)^{-1} = \mathcal{O}(1/g(w))$ for all $w \in \tilde{\Omega}$. The latter condition is

equivalent to the existence of a constant $M > 0$ such that

$$\frac{1}{M}I \preceq \frac{1}{g(w)}U(w) \preceq MI \quad \forall w \in \Omega.$$

2. Preliminaries. In this section, we describe the assumptions that will be used in our presentation. We also describe the weighted central path that will be the subject of our study in this paper. Some preliminary results about this path are also stated including conditions for its well-definedness.

Throughout this paper we will be dealing with the pair of dual SDPs (P) and (D) (see (1) and (2), respectively). Denote the feasible sets of (P) and (D) by $\mathcal{F}_P$ and $\mathcal{F}_D$, respectively. Throughout our presentation we make the following assumptions on the pair of problems (P) and (D) .

A.1. $\mathcal{A} : \mathcal{S}^n \rightarrow \mathcal{R}^m$ is an onto linear operator.

A.2. There exists a pair of strictly complementary primal-dual optimal solutions for (P) and (D) , that is, a triple $(X^*, S^*, y^*) \in \mathcal{F}_P \times \mathcal{F}_D$ satisfying $X^*S^* = 0$ and $X^* + S^* \succ 0$.

We will assume that assumptions A.1 and A.2 are in force throughout our presentation. Hence, we will state our results without explicitly mentioning them.

Assumption A.1 is not really crucial for our analysis, but it is convenient to ensure that the variables S and y are in one-to-one correspondence. We will see that the dual weighted central path can always be defined in the S -space. The goal of assumption A.1 is just to ensure that this path can also be extended to the y -space.

Assumption A.2 is the one that is commonly used in the analysis of superlinear convergence of interior-point algorithms, and it plays an important role in our analysis. In fact, it is a very challenging problem to generalize the analysis of this paper to the case where assumption A.2 is dropped or simply relaxed.

By assumption A.2, since $X^*S^* = S^*X^* = 0$, we can diagonalize X^* and S^* simultaneously, i.e., find an orthonormal $P \in \mathcal{R}^{n \times n}$ such that $P^T X^* P$ and $P^T S^* P$ are both diagonal. Performing the change of variables $\hat{X} = P^T X P$ and $(\hat{S}, \hat{y}) = (P^T S P, y)$ on problems (P) and (D) yields another pair of primal and dual SDPs which has a primal-dual optimal solution $(\hat{X}^*, \hat{S}^*, \hat{y}^*)$ such that $\hat{X}^*$ and $\hat{S}^*$ are both diagonal. To simplify our notation, we will assume without loss of generality that the original (P) and (D) already have a primal-dual optimal solution (X^*, S^*, y^*) such that

$$(6) \quad X^* = \begin{bmatrix} \Lambda_B & 0 \\ 0 & 0 \end{bmatrix}, \quad S^* = \begin{bmatrix} 0 & 0 \\ 0 & \Lambda_N \end{bmatrix},$$

where $\Lambda_B \equiv \text{diag}(\lambda_1, \dots, \lambda_K)$, $\Lambda_N \equiv \text{diag}(\lambda_{K+1}, \dots, \lambda_n)$ for some integer $0 \leq K \leq n$ and some scalars $\lambda_i > 0$, $i = 1, 2, \dots, n$. Here the subscripts B and N signify the “basic” and “nonbasic” subspaces (following the terminology of linear programming). Throughout this paper, the decomposition of any $n \times n$ matrix V is always made with respect to the above partition B and N , namely,

$$V = \begin{bmatrix} V_B & V_{BN} \\ V_{NB} & V_N \end{bmatrix},$$

so that V_{BN} and V_{NB} denote the off-diagonal block of V . If $V_{BN} = 0$ and $V_{NB} = 0$, V is called block diagonal; otherwise, it is called non-block diagonal.

Notice that $X \in \mathcal{F}_P$ is an optimal solution of (P) if and only if $XS^* = 0$. Hence, by assumption A.2, the primal optimal solution set $\mathcal{F}_P^*$ is given by

$$\mathcal{F}_P^* \equiv \{X \in \mathcal{F}_P : X_{BN} = 0, X_{NB} = 0, \text{ and } X_N = 0\}.$$

Analogously, the dual optimal solution set $\mathcal{F}_D^*$ is given by

$$\mathcal{F}_D^* \equiv \{(S, y) \in \mathcal{F}_D : S_{BN} = 0, S_{NB} = 0, \text{ and } S_B = 0\}.$$

Define the linear map $\mathcal{G} : \mathcal{S}^n \times \mathcal{S}^n \times \mathbb{R}^m \rightarrow \mathcal{S}^n \times \mathbb{R}^m$ by

$$(7) \quad \mathcal{G}(X, S, y) \equiv (\mathcal{A}^* y + S - C, \mathcal{A}X - b)$$

and the sets $\mathcal{G}_{++}$ and $\mathcal{W}$ by

$$(8) \quad \mathcal{G}_{++} \equiv \mathcal{G}(\mathcal{S}_{++}^n \times \mathcal{S}_{++}^n \times \mathbb{R}^m),$$

$$(9) \quad \mathcal{W} \equiv \left\{ W \in \mathcal{S}_{++}^n : \|W - \nu I\| < \nu/\sqrt{2} \text{ for some } \nu > 0 \right\}.$$

Given $(W, \Delta C, \Delta b) \in \mathcal{S}^n \times \mathcal{S}^n \times \mathbb{R}^m$, in this paper we are interested in the solutions of the system of nonlinear equations (3)–(5) parametrized by the parameter $\nu > 0$. The following result gives a condition on $(W, \Delta C, \Delta b)$ for system (3)–(5) to have a unique solution for each $\nu \in (0, 1]$.

PROPOSITION 2.1. *Assume that $(W, \Delta C, \Delta b) \in \mathcal{W} \times \mathcal{G}_{++}$. Then, for any $\nu \in (0, 1]$, system (3)–(5) has a unique solution, denoted by $(X(\nu), S(\nu), y(\nu))$. Moreover, the path $\nu \in (0, 1] \rightarrow (X(\nu), S(\nu), y(\nu))$ is analytic.*

Proof. By A.2 and the assumption that $(W, \Delta C, \Delta b) \in \mathcal{W} \times \mathcal{G}_{++}$, we easily see that $\nu(W, \Delta C, \Delta b) \in \mathcal{W} \times \mathcal{G}_{++}$ for all $\nu \in (0, 1]$. The first conclusion of the proposition now follows from Theorem 1(b) of Monteiro and Zanjácomo [31] by letting F , Φ , and $\mathcal{V}$ in that theorem be defined as $F = \mathcal{G}$, $\Phi(X, S) = X^{1/2} S X^{1/2}$ for all $(X, S) \in \mathcal{S}_+^n \times \mathcal{S}_+^n$, and $\mathcal{V} = \mathcal{W}$. The second conclusion follows by applying the analytic version of the implicit function theorem to system (3)–(5) viewed as a function of (X, S, y, ν) and using the fact that the assumption $(W, \Delta C, \Delta b) \in \mathcal{W} \times \mathcal{G}_{++}$ implies that the Jacobian of this system with respect to (X, S, y) is nonsingular at $(X(\nu), S(\nu), y(\nu), \nu)$ for every $\nu \in (0, 1]$. (See Theorem 2.4 of [30] and the paragraph following it.) $\square$

For a given $(W, \Delta C, \Delta b) \in \mathcal{W} \times \mathcal{G}_{++}$, the path $\nu \in (0, 1] \rightarrow (X(\nu), S(\nu), y(\nu))$ will be referred to as the $(W, \Delta C, \Delta b)$ -weighted central path. In view of the above proposition, we will assume throughout sections 2–4 that the following condition is true, without explicitly mentioning it in the statements of the results.

A.3. $(W, \Delta C, \Delta b) \in \mathcal{W} \times \mathcal{G}_{++}$.

The next result gives some estimates on the size of the blocks of $X(\nu)$ and $S(\nu)$.

LEMMA 2.2. *For all $\nu > 0$ sufficiently small, we have*

$$(10) \quad X_B(\nu) = \mathcal{O}(1), \quad S_N(\nu) = \mathcal{O}(1),$$

$$(11) \quad X_N(\nu) = \mathcal{O}(\nu), \quad S_B(\nu) = \mathcal{O}(\nu),$$

$$(12) \quad X_{BN}(\nu) = \mathcal{O}(\sqrt{\nu}), \quad S_{BN}(\nu) = \mathcal{O}(\sqrt{\nu}).$$

Proof. Assume that $\nu > 0$ is sufficiently small and, for notational convenience, let $X \equiv X(\nu)$ and $S \equiv S(\nu)$. Also, let (X^*, S^*, y^*) be as in condition A.2. Since $(\Delta C, \Delta b) \in \mathcal{G}_{++}$, we have $(\Delta C, \Delta b) = \mathcal{G}(X^0, S^0, y^0)$ for some $(X^0, S^0, y^0) \in \mathcal{S}_{++}^n \times \mathcal{S}_{++}^n \times \mathbb{R}^m$. It is easy to see that $A(X - \nu X^0 - (1 - \nu)X^*) = 0$ and $S - \nu S^0 - (1 - \nu)S^* \in \text{Im}(\mathcal{A}^*)$ and hence that

$$(13) \quad (X - \nu X^0 - (1 - \nu)X^*) \bullet (S - \nu S^0 - (1 - \nu)S^*) = 0.$$

This equality together with (6) and the facts that $X^* \bullet S^* = 0$, $X \bullet S = \nu \text{tr}(W)$, and all the quantities X, X^0, X^*, S, S^0, S^* are in $\mathcal{S}_+^n$ implies that

$$(14) \quad X \bullet S^0 + X^0 \bullet S \leq \text{tr}(W) + \xi(\nu)$$

and

$$(15) \quad X_N \bullet S_N^* + X_B^* \bullet S_B = X \bullet S^* + X^* \bullet S \leq \frac{\nu(\text{tr}(W) + \xi(\nu))}{1 - \nu},$$

where $\xi(\nu) \equiv \nu(X^0 \bullet S^0) + (1 - \nu)(X^0 \bullet S^* + X^* \bullet S^0)$. The above two inequalities together with the fact that the matrices $X^0, S^0, X_B^*, S_N^*, X_N, S_B$ are all positive definite clearly imply that (10) and (11) hold. Using the fact that $X(\nu) \succ 0$ and $S(\nu) \succ 0$, we obtain that $X_{ij}^2(\nu) \leq X_{ii}(\nu)X_{jj}(\nu)$ and $S_{ij}^2(\nu) \leq S_{ii}(\nu)S_{jj}(\nu)$ for all i, j . These inequalities together with (10) and (11) imply (12). $\square$

The next result gives estimates on the size of the blocks of the matrix $X^{1/2}(\nu) \equiv [X(\nu)]^{1/2}$.

LEMMA 2.3. *Let $U(\nu) \equiv X^{1/2}(\nu)$ for all $\nu \in (0, 1]$. Then, for all $\nu > 0$ sufficiently small, we have*

$$U(\nu) = \begin{pmatrix} U_B(\nu) & U_{BN}(\nu) \\ U_{NB}(\nu) & U_N(\nu) \end{pmatrix} = \begin{pmatrix} \mathcal{O}(1) & \mathcal{O}(\sqrt{\nu}) \\ \mathcal{O}(\sqrt{\nu}) & \mathcal{O}(\sqrt{\nu}) \end{pmatrix}.$$

Proof. For notational convenience, let $U = U(\nu)$. Since $X = UU$, we have $X_B = U_B U_B + U_{BN} U_{BN}^T$ and $X_N = U_N U_N + U_{NB} U_{NB}^T$. Hence,

$$\begin{aligned} n\|X_B\| &\geq \text{tr} X_B = \text{tr} (U_B U_B + U_{BN} U_{BN}^T) \\ &= \|U_B\|_F^2 + \|U_{BN}\|_F^2 \geq \max\{\|U_B\|^2, \|U_{BN}\|^2\}, \\ n\|X_N\| &\geq \text{tr} X_N = \text{tr} (U_N U_N + U_{NB} U_{NB}^T) \\ &= \|U_N\|_F^2 + \|U_{NB}\|_F^2 \geq \max\{\|U_N\|^2, \|U_{NB}\|^2\}, \end{aligned}$$

from which the result follows in view of (10) and (11). $\square$

We end this section by stating a convergence result of the $(W, \Delta C, \Delta b)$ -weighted central path to a primal-dual optimal solution of (1) and (2). We do not provide a proof for it since it is similar to the one given in the appendix of Halická, de Klerk, and Roos [14].

PROPOSITION 2.4. *There exist some $\epsilon > 0$ and an analytic function $\nu : [0, \epsilon) \rightarrow (0, 1)$ such that $\nu(0) = 0$ and the path $t \in (0, \epsilon) \rightarrow (X(\nu(t)), S(\nu(t)), y(\nu(t)))$ is analytic at $t = 0$. In particular, $(X(\nu(t)), S(\nu(t)), y(\nu(t)))$ converges to some primal-dual optimal solution (X^*, S^*, y^*) as $t \downarrow 0$.*

We observe that Proposition 2.4 holds even without requiring assumption A.2. As a consequence, its main advantage is that it holds for any SDP problem. Its main drawbacks are that it neither gives a characterization of the limit point (X^*, S^*, y^*) nor describes how fast $\nu(t)$ converges to 0. These issues and others will be addressed in the remaining sections of this paper in the context of SDPs satisfying assumption A.2.

3. Analyticity of the weighted central path. In the parametrization introduced in the previous section, the weighted central path in general cannot be extended analytically to an interval of the form $(-\epsilon, \infty)$ for some $\epsilon > 0$ (see Corollary 4.3). However, in this section we will show that the reparametrized weighted central path $t \rightarrow p(t^2)$ can be extended analytically to an interval as above.

For the sake of brevity, it is convenient to introduce the following definition.

DEFINITION 3.1. *Let $w : (0, \delta) \rightarrow E$ be a given function where $\delta > 0$ and E is a finite dimensional normed vector space. The function w is said to be analytic at 0 if*

there exist $\epsilon > 0$ and an analytic function $\psi : (-\epsilon, \epsilon) \rightarrow E$ such that $w(t) = \psi(t)$ for all $t \in (0, \epsilon)$.

The basic result that we use to establish that a function $w : (0, \delta) \rightarrow E$ is analytic at 0 is the following corollary of the analytic version of the implicit function theorem.

PROPOSITION 3.2. *Let $w : (0, \delta) \rightarrow E$ be a given function where $\delta > 0$ and E is a finite dimensional normed vector space. Assume that there exists an analytic function $H : \mathcal{O} \times (-\epsilon, \epsilon) \rightarrow E$, where $\epsilon > 0$ and $\mathcal{O}$ is an open subset of E , such that $w = w(t)$ is the unique solution of $H(w, t) = 0$ in $\mathcal{O}$ for every $t \in (0, \epsilon)$. Assume also there exists $\bar{w} \in \mathcal{O}$ such that $H(\bar{w}, 0) = 0$ and $H'_w(\bar{w}, 0)$ is nonsingular. Then,*

- (i) $w = \bar{w}$ is the unique solution of the system $H(w, 0) = 0$;
- (ii) w is analytic at 0 and, as a consequence, $\lim_{t \downarrow 0} w(t) = \bar{w}$, and the limits of all the derivatives of $w(t)$ as $t \downarrow 0$ exist.

The following theorem is one of the main results of this section. Its proof will be given at the end of this section.

THEOREM 3.3. *The reparametrized $(W, \Delta C, \Delta b)$ -weighted central path $t \in (0, 1] \rightarrow (X(t^2), S(t^2), y(t^2))$ is analytic and also analytic at $t = 0$. As a consequence, the $(W, \Delta C, \Delta b)$ -weighted central path $\nu \in (0, 1] \rightarrow (X(\nu), S(\nu), y(\nu))$ converges.*

A key step toward showing the above result is a reformulation of the weighted central path system (3)–(5) as we now discuss. First, observe that (3), (4), and the equations

$$(16) \quad USU = t^2 W,$$

$$(17) \quad UU = X$$

have $(U, X, S, y) = (U(t^2), X(t^2), S(t^2), y(t^2))$ as their unique solution in $\mathcal{S}_{++}^n \times \mathcal{S}_{++}^n \times \mathcal{S}_{++}^n \times \mathbb{R}^m$, where $U(t^2) \equiv [X(t^2)]^{1/2}$. Letting

$$(18) \quad D_B(t) \equiv \begin{bmatrix} I/t & 0 \\ 0 & I \end{bmatrix}, \quad D_N(t) \equiv \begin{bmatrix} I & 0 \\ 0 & I/t \end{bmatrix},$$

and noting that $D_B(t)D_N(t) = I/t$ for every $t \in (0, 1]$, we easily see that $U, X, S \in \mathcal{S}_{++}^n$ satisfies (16) and (17) if and only if $U, \tilde{X} \equiv D_N(t)X D_N(t)$, and $\tilde{S} \equiv D_B(t)S D_B(t)$ satisfy

$$(19) \quad [U D_N(t)] \tilde{S} [D_N(t) U] = W,$$

$$(20) \quad [D_N(t) U] [U D_N(t)] = \tilde{X}.$$

Now, let

$$\mathcal{U}^n = \left\{ U \in \mathbb{R}^{n \times n} : U_B \in \mathcal{S}^{|B|}, U_N \in \mathcal{S}^{|N|}, U_{NB} = 0 \right\},$$

$$\mathcal{U}_{++}^n = \{ U \in \mathcal{U}^n : U_B \succ 0, U_N \succ 0 \},$$

and define $\mathcal{L} : \mathcal{U}^n \rightarrow \mathbb{R}^{n \times n}$ as

$$\mathcal{L}(U) = \begin{bmatrix} 0 & 0 \\ U_{BN}^T & 0 \end{bmatrix} \quad \forall U \in \mathcal{U}^n.$$

It then follows that $(U, \tilde{X}, \tilde{S})$ satisfies (19) and (20) if and only if $(\tilde{U}, \tilde{X}, \tilde{S})$ with

$$\tilde{U} \equiv \begin{bmatrix} U_B & U_{BN}/t \\ 0 & U_N/t \end{bmatrix}$$

satisfies the equations

$$(21) \quad \left[\tilde{U} + t \mathcal{L}(\tilde{U}) \right] \tilde{S} \left[\tilde{U} + t \mathcal{L}(\tilde{U}) \right]^T = W,$$

$$(22) \quad \left[\tilde{U} + t \mathcal{L}(\tilde{U}) \right]^T \left[\tilde{U} + t \mathcal{L}(\tilde{U}) \right] = \tilde{X}.$$

Indeed, the above claim follows from the identity

$$UD_N(t) = \begin{bmatrix} U_B & U_{BN}/t \\ U_{NB} & U_N/t \end{bmatrix} = \tilde{U} + t \mathcal{L}(\tilde{U}).$$

The above arguments establish the following key result.

PROPOSITION 3.4. *Let $(X^*, S^*, y^*) \in \mathcal{F}_P^* \times \mathcal{F}_D^*$ be given. Then, for every $t \in (0, 1]$, the system defined by (21), (22), and the linear equations*

$$(23) \quad \mathcal{A} \left(D_N(t)^{-1} \tilde{X} D_N(t)^{-1} - X^* \right) = t^2 \Delta b,$$

$$(24) \quad D_B(t)^{-1} \tilde{S} D_B(t)^{-1} - S^* \in t^2 \Delta C + \text{Im}(\mathcal{A}^*)$$

has a unique solution, denoted by $(\tilde{U}(t), \tilde{X}(t), \tilde{S}(t))$, in $\mathcal{U}_{++}^n \times \mathcal{S}_{++}^n \times \mathcal{S}_{++}^n$. Moreover, the path $t \in (0, 1] \rightarrow (\tilde{U}(t), \tilde{X}(t), \tilde{S}(t))$ is analytic and, for every $t \in (0, 1]$,

$$(25) \quad \tilde{X}(t) = D_N(t) X(t^2) D_N(t), \quad \tilde{S}(t) = D_B(t) S(t^2) D_B(t),$$

$$(26) \quad \tilde{U}(t) = \begin{bmatrix} U_B(t^2) & U_{BN}(t^2)/t \\ 0 & U_N(t^2)/t \end{bmatrix}.$$

The next result states some basic properties about the accumulation points of the path $t \in (0, 1] \rightarrow (\tilde{U}(t), \tilde{X}(t), \tilde{S}(t))$ as t approaches 0.

PROPOSITION 3.5. *The path $t \in (0, 1] \rightarrow (\tilde{U}(t), \tilde{X}(t), \tilde{S}(t))$ remains bounded as t approaches 0, and any accumulation point $(\tilde{U}^*, \tilde{X}^*, \tilde{S}^*)$ of this path as t approaches 0 is in $\mathcal{U}_{++}^n \times \mathcal{S}_{++}^n \times \mathcal{S}_{++}^n$ and satisfies the equations*

$$(27) \quad \tilde{U} \tilde{S} \tilde{U}^T = W,$$

$$(28) \quad \tilde{U}^T \tilde{U} = \tilde{X}.$$

Proof. By (18) and (25), we have

$$(29) \quad \tilde{X}(t) = \begin{bmatrix} X_B(t^2) & X_{BN}(t^2)/t \\ X_{NB}(t^2)/t & X_N(t^2)/t^2 \end{bmatrix}, \quad \tilde{S}(t) = \begin{bmatrix} S_B(t^2)/t^2 & S_{BN}(t^2)/t \\ S_{NB}(t^2)/t & S_N(t^2) \end{bmatrix},$$

which, together with Lemma 2.2, imply that $(\tilde{X}(t), \tilde{S}(t))$ remains bounded as t approaches 0. Relation (26) and Lemma 2.3 imply that $\tilde{U}(t)$ also remains bounded as t approaches 0.

Consider now an accumulation point $(\tilde{U}^*, \tilde{X}^*, \tilde{S}^*)$ of the path $t \in (0, 1] \rightarrow (\tilde{U}(t), \tilde{X}(t), \tilde{S}(t))$ as t approaches 0. By (25) and (26), we see that $(\tilde{U}(t), \tilde{X}(t), \tilde{S}(t)) \in \mathcal{U}_{++}^n \times \mathcal{S}_{++}^n \times \mathcal{S}_{++}^n$ for all $t \in (0, 1]$, and hence we must have $\tilde{U}^* \in \mathcal{U}^n$, $\tilde{X}^* \succeq 0$, $\tilde{S}^* \succeq 0$, $\tilde{U}_B^* \succeq 0$, and $\tilde{U}_N^* \succeq 0$. Thus, to conclude that $(\tilde{U}^*, \tilde{X}^*, \tilde{S}^*) \in \mathcal{U}_{++}^n \times \mathcal{S}_{++}^n \times \mathcal{S}_{++}^n$, it suffices to show that $\tilde{U}^*$, $\tilde{X}^*$, and $\tilde{S}^*$ are all invertible. Indeed, since $(\tilde{U}(t), \tilde{X}(t), \tilde{S}(t))$ satisfies (21) and (22), we conclude upon letting $t \downarrow 0$ that $(\tilde{U}^*, \tilde{X}^*, \tilde{S}^*)$ satisfies (27) and (28). This conclusion together with the fact that $W \succ 0$ implies that $\tilde{U}^*$, $\tilde{X}^*$, and $\tilde{S}^*$ are all invertible. $\square$

Our next goal is to show that the path $t \in (0, 1] \rightarrow (\tilde{U}(t), \tilde{X}(t), \tilde{S}(t))$ is analytic at $t = 0$. The basic tool we use to establish this fact is the implicit function theorem applied to a specific system of equations parametrized by the parameter $t \in \mathfrak{R}$. A first natural candidate for such a system seems to be the one given by (21), (22), (23), and (24). However, the main drawback of this system is that its Jacobian with respect to $(\tilde{U}, \tilde{X}, \tilde{S})$ is generally singular for $t = 0$ (even though for $t \in (0, 1)$ it is always nonsingular). The main cause for this phenomenon is that the “rank” of the linear equations (23) and (24) changes when t becomes 0.

We will now show how the linear equations (23) and (24) can be reformulated into equivalent linear equations for every $t \in (0, 1]$. Moreover, the new linear equations have the property that their rank remains constant for every $t \in \mathfrak{R}$. First note that the linear operator $\mathcal{A} : \mathcal{S}^n \rightarrow \mathfrak{R}^m$ can be expressed as

$$(30) \quad \mathcal{A}(X) = \mathcal{A}_B(X_B) + \mathcal{A}_{BN}(X_{BN}) + \mathcal{A}_N(X_N) \equiv (\mathcal{A}_B \ \mathcal{A}_{BN} \ \mathcal{A}_N) \begin{pmatrix} X_B \\ X_{BN} \\ X_N \end{pmatrix}$$

for some linear operators $\mathcal{A}_B : \mathcal{S}^{|B|} \rightarrow \mathfrak{R}^m$, $\mathcal{A}_{BN} : \mathfrak{R}^{|B| \times |N|} \rightarrow \mathfrak{R}^m$, and $\mathcal{A}_N : \mathcal{S}^{|N|} \rightarrow \mathfrak{R}^m$.

A well-known result from linear algebra says that any matrix can be put into row-echelon form after a sequence of elementary row operations. A similar type of argument allows one to establish the following result.

LEMMA 3.6. *Let $\mathcal{A} : \mathcal{S}^n \rightarrow \mathfrak{R}^m$ be an onto linear operator. Assume that*

$$i_1 = \text{rank}(\mathcal{A}_B), \quad i_2 = \text{rank}(\mathcal{A}_B \ \mathcal{A}_{BN}) - i_1, \quad i_3 = \text{rank}(\mathcal{A}) - (i_1 + i_2) = m - (i_1 + i_2).$$

Then there exists an isomorphism $T : \mathfrak{R}^m \rightarrow \mathfrak{R}^m$ such that

$$\begin{aligned} (T \circ \mathcal{A})(X) &= \begin{pmatrix} \mathcal{A}_{11}(X_B) & + \mathcal{A}_{12}(X_{BN}) & + \mathcal{A}_{13}(X_N) \\ & \mathcal{A}_{22}(X_{BN}) & + \mathcal{A}_{23}(X_N) \\ & & \mathcal{A}_{33}(X_N) \end{pmatrix} \\ &\equiv \begin{pmatrix} \mathcal{A}_{11} & \mathcal{A}_{12} & \mathcal{A}_{13} \\ 0 & \mathcal{A}_{22} & \mathcal{A}_{23} \\ 0 & 0 & \mathcal{A}_{33} \end{pmatrix} \begin{pmatrix} X_B \\ X_{BN} \\ X_N \end{pmatrix} \end{aligned}$$

for some linear operators

$$\begin{aligned} \mathcal{A}_{11} : \mathcal{S}^{|B|} &\rightarrow \mathfrak{R}^{i_1}, & \mathcal{A}_{12} : \mathfrak{R}^{|B| \times |N|} &\rightarrow \mathfrak{R}^{i_1}, \\ \mathcal{A}_{13} : \mathcal{S}^{|N|} &\rightarrow \mathfrak{R}^{i_1}, & \mathcal{A}_{22} : \mathfrak{R}^{|B| \times |N|} &\rightarrow \mathfrak{R}^{i_2}, \\ \mathcal{A}_{23} : \mathcal{S}^{|N|} &\rightarrow \mathfrak{R}^{i_2}, & \mathcal{A}_{33} : \mathcal{S}^{|N|} &\rightarrow \mathfrak{R}^{i_3} \end{aligned}$$

such that $\text{rank}(\mathcal{A}_{11}) = i_1$, $\text{rank}(\mathcal{A}_{22}) = i_2$, $\text{rank}(\mathcal{A}_{33}) = i_3$.

We can now reformulate the linear system (23) with the use of Lemma 3.6 as follows. Using the fact that

$$D_N^{-1} \tilde{X} D_N^{-1} - X^* = \begin{bmatrix} \tilde{X}_B - X_B^* & t \tilde{X}_{BN} \\ t \tilde{X}_{NB} & t^2 \tilde{X}_N \end{bmatrix}$$

and Lemma 3.6, we easily see that (23) is equivalent to the linear system

$$\begin{pmatrix} \mathcal{A}_{11} & t \mathcal{A}_{12} & t^2 \mathcal{A}_{13} \\ 0 & t \mathcal{A}_{22} & t^2 \mathcal{A}_{23} \\ 0 & 0 & t^2 \mathcal{A}_{33} \end{pmatrix} \begin{pmatrix} \tilde{X}_B - X_B^* \\ \tilde{X}_{BN} \\ \tilde{X}_N \end{pmatrix} = t^2 \begin{pmatrix} \widetilde{\Delta b_1} \\ \widetilde{\Delta b_2} \\ \widetilde{\Delta b_3} \end{pmatrix},$$

where $\widetilde{\Delta b} \equiv T(\Delta b)$. Dividing the second and third blocks of rows in the above system by t and t^2 , respectively, we obtain the system

$$(31) \quad \begin{pmatrix} \mathcal{A}_{11} & t\mathcal{A}_{12} & t^2\mathcal{A}_{13} \\ 0 & \mathcal{A}_{22} & t\mathcal{A}_{23} \\ 0 & 0 & \mathcal{A}_{33} \end{pmatrix} \begin{pmatrix} \tilde{X}_B - X_B^* \\ \tilde{X}_{BN} \\ \tilde{X}_N \end{pmatrix} = \begin{pmatrix} t^2\widetilde{\Delta b_1} \\ t\widetilde{\Delta b_2} \\ \widetilde{\Delta b_3} \end{pmatrix}.$$

Note that the linear system (31) is equivalent to (23) for every $t \in (0, 1]$. Hence, $\tilde{X}(t)$ satisfies (31) for every $t \in (0, 1]$. A nice feature of (31) is that the operator on its left-hand side does not lose full rankness as t becomes 0. We state this fact in the following proposition.

PROPOSITION 3.7. *Let $\mathcal{A}_t : \mathcal{S}^n \rightarrow \mathbb{R}^m$ be the operator such that $\mathcal{A}_t(\tilde{X})$ is defined by the left-hand side of (31). Then, $t \in \mathbb{R} \rightarrow \mathcal{A}_t$ is a continuous map such that $\text{rank}(\mathcal{A}_t) = m$ for every $t \in \mathbb{R}$.*

The linear system (24) can also be reformulated with the aid of Lemma 3.6 as follows. First note that by Lemma 3.6 we have

$$\begin{aligned} \text{Im}(\mathcal{A}^*) &= \text{Im}[(T \circ \mathcal{A})^*] = \text{Im} \left[\begin{pmatrix} \mathcal{A}_{11}^* & 0 & 0 \\ \mathcal{A}_{12}^* & \mathcal{A}_{22}^* & 0 \\ \mathcal{A}_{13}^* & \mathcal{A}_{23}^* & \mathcal{A}_{33}^* \end{pmatrix} \right] \\ &= \text{Im} \left[\begin{pmatrix} t^2\mathcal{A}_{11}^* & 0 & 0 \\ t^2\mathcal{A}_{12}^* & t\mathcal{A}_{22}^* & 0 \\ t^2\mathcal{A}_{13}^* & t\mathcal{A}_{23}^* & \mathcal{A}_{33}^* \end{pmatrix} \right] \end{aligned}$$

for every $t \in (0, 1]$. Hence, for every $t \in (0, 1]$, (24) is equivalent to

$$\begin{pmatrix} t^2\tilde{S}_B \\ t\tilde{S}_{BN} \\ \tilde{S}_N - S_N^* \end{pmatrix} \in t^2 \begin{pmatrix} \Delta C_B \\ \Delta C_{BN} \\ \Delta C_N \end{pmatrix} + \text{Im} \left[\begin{pmatrix} t^2\mathcal{A}_{11}^* & 0 & 0 \\ t^2\mathcal{A}_{12}^* & t\mathcal{A}_{22}^* & 0 \\ t^2\mathcal{A}_{13}^* & t\mathcal{A}_{23}^* & \mathcal{A}_{33}^* \end{pmatrix} \right].$$

Dividing the first and second block of rows in the above system by t^2 and t , respectively, we obtain the system

$$(32) \quad \begin{pmatrix} \tilde{S}_B \\ \tilde{S}_{BN} \\ \tilde{S}_N - S_N^* \end{pmatrix} \in \begin{pmatrix} \Delta C_B \\ t\Delta C_{BN} \\ t^2\Delta C_N \end{pmatrix} + \text{Im} \left[\begin{pmatrix} \mathcal{A}_{11}^* & 0 & 0 \\ t\mathcal{A}_{12}^* & \mathcal{A}_{22}^* & 0 \\ t^2\mathcal{A}_{13}^* & t\mathcal{A}_{23}^* & \mathcal{A}_{33}^* \end{pmatrix} \right],$$

which is equivalent to (24) and hence is satisfied by $\tilde{S}(t)$, for all $t \in (0, 1]$.

Using the definition of $\mathcal{A}_t$ and the fact that $\tilde{X}(t)$ and $\tilde{S}(t)$ satisfy (31) and (32), respectively, for every $t \in (0, 1]$, we conclude that there exists a function $\tilde{y} : (0, 1] \rightarrow \mathbb{R}^m$ such that $(\tilde{X}(t), \tilde{S}(t), \tilde{y}(t))$ satisfies

$$(33) \quad \mathcal{A}_t(\tilde{X} - X^*) = \begin{pmatrix} t^2\widetilde{\Delta b_1} \\ t\widetilde{\Delta b_2} \\ \widetilde{\Delta b_3} \end{pmatrix}, \quad \mathcal{A}_t^* \tilde{y} + (\tilde{S} - S^*) = \begin{pmatrix} \Delta C_B \\ t\Delta C_{BN} \\ t^2\Delta C_N \end{pmatrix}$$

for every $t \in (0, 1]$. Moreover, using Proposition 3.7 and the fact that $\{\tilde{S}(t) : t \in (0, 1]\}$ is bounded, we easily see that $\{\tilde{y}(t) : t \in (0, 1]\}$ is also bounded. We have thus established the following result.

PROPOSITION 3.8. *There exists a curve $\tilde{y} : \mathbb{R}_{++} \rightarrow \mathbb{R}^m$ such that $(\tilde{U}(t), \tilde{X}(t), \tilde{S}(t), \tilde{y}(t))$ is the unique solution of (21), (22), and (33) in $\mathcal{U}_{++}^n \times \mathcal{S}_{++}^n \times \mathcal{S}_{++}^n \times \mathbb{R}^m$ for every $t \in (0, 1]$. Moreover, the path $t \in (0, 1] \rightarrow (\tilde{U}(t), \tilde{X}(t), \tilde{S}(t), \tilde{y}(t))$ remains bounded as t approaches 0 and any of its accumulation points are in $\mathcal{U}_{++}^n \times \mathcal{S}_{++}^n \times \mathcal{S}_{++}^n \times \mathbb{R}^m$.*

The system formed by (21), (22), and (33) is the one which we will use to establish that the path $t \in (0, 1] \rightarrow (\tilde{U}(t), \tilde{X}(t), \tilde{S}(t), \tilde{y}(t))$ is analytic at $t = 0$. This will follow by Proposition 3.2 if we can establish that the Jacobian of this system with $t = 0$ with respect to $(\tilde{U}, \tilde{X}, \tilde{S}, \tilde{y})$ is nonsingular as long as $(\tilde{U}, \tilde{X}, \tilde{S})$ is well centered in the sense that $\|\tilde{U}\tilde{S}\tilde{U} - \nu I\| < \nu/\sqrt{2}$ for some $\nu \in (0, 1]$. The nonsingularity of this Jacobian can be easily seen to be equivalent to showing that $(\widetilde{\Delta U}, \widetilde{\Delta X}, \widetilde{\Delta S}, \widetilde{\Delta y}) = (0, 0, 0, 0) \in \mathcal{U}^n \times \mathcal{S}^n \times \mathcal{S}^n \times \mathbb{R}^m$ is the only solution of the following linear system:

$$\begin{aligned} \widetilde{\Delta U} \tilde{S} \tilde{U}^T + \tilde{U} \widetilde{\Delta S} \tilde{U}^T + \tilde{U} \tilde{S} \widetilde{\Delta U}^T &= 0, \\ \widetilde{\Delta U}^T \tilde{U} + \tilde{U}^T \widetilde{\Delta U} &= \widetilde{\Delta X}, \\ \mathcal{A}_0 \widetilde{\Delta X} &= 0, \\ \mathcal{A}_0^* \widetilde{\Delta y} + \widetilde{\Delta S} &= 0. \end{aligned} \quad (34)$$

Before establishing the above fact, we state and prove two technical results.

LEMMA 3.9. *For any $U \in \mathcal{U}_{++}^n$ and $H \in \mathcal{S}^n$, there exists a unique matrix $V \in \mathcal{U}^n$ such that*

$$H = U^T V + V^T U. \quad (35)$$

Moreover,

$$\|VU^{-1}\|_F \leq \frac{\|U^{-T} H U^{-1}\|_F}{\sqrt{2}}. \quad (36)$$

Proof. The first part of the lemma follows from the fact that the linear map $\Psi_U : \mathcal{U}^n \rightarrow \mathcal{S}^n$ defined by $\Psi_U(V) = U^T V + V^T U$ for all $V \in \mathcal{U}^n$ is an isomorphism. Indeed, since its domain and codomain have the same dimension, it suffices to show that Ψ_U is one-to-one or, equivalently, that $U^T V + V^T U = 0$ implies $V = 0$. In turn, this last implication follows from the fact that any solution V of (35) satisfies (36) (simply set $H = 0$ in (36) to conclude that $V = 0$). To show the last claim, we multiply both sides of (35) on the left by U^{-T} and on the right by U^{-1} to obtain

$$U^{-T} H U^{-1} = V U^{-1} + (V U^{-1})^T. \quad (37)$$

Letting $\tilde{U} \equiv V U^{-1}$ and squaring both sides of the above equation, we obtain

$$\|U^{-T} H U^{-1}\|_F^2 = \|\tilde{U} + \tilde{U}^T\|_F^2 = 2\|\tilde{U}\|_F^2 + 2\text{tr}(\tilde{U}^2). \quad (38)$$

Since

$$\begin{aligned} \tilde{U} = V U^{-1} &= \begin{bmatrix} V_B & V_{BN} \\ 0 & V_N \end{bmatrix} \begin{bmatrix} U_B^{-1} & -U_B^{-1} U_{BN} U_N^{-1} \\ 0 & U_N^{-1} \end{bmatrix} \\ &= \begin{bmatrix} V_B U_B^{-1} & -V_B U_B^{-1} U_{BN} U_N^{-1} + V_{BN} U_N^{-1} \\ 0 & V_N U_N^{-1} \end{bmatrix}, \end{aligned}$$

we have

$$\begin{aligned} \text{tr}(\tilde{U}^2) &= \text{tr}((V_B U_B^{-1})^2) + \text{tr}((V_N U_N^{-1})^2), \\ &= \|U_B^{-1/2} V_B U_B^{-1/2}\|_F^2 + \|U_N^{-1/2} V_N U_N^{-1/2}\|_F^2 \geq 0. \end{aligned} \quad (39)$$

Hence, by (38) and (39), we see that (36) holds. $\square$

LEMMA 3.10. *Suppose that $\gamma \in [0, 1/\sqrt{2})$ and that $(U, S) \in \mathcal{U}_{++}^n \times \mathcal{S}^n$ is such that $\|USU^T - \nu I\| \leq \gamma\nu$ for some $\nu > 0$. For some $H \in \mathcal{S}^n$, if $(\Delta U, \Delta X, \Delta S)$ satisfies*

$$(40) \quad \Delta USU^T + U\Delta SU^T + US\Delta U^T = H,$$

$$(41) \quad \Delta U^T U + U^T \Delta U = \Delta X,$$

$$(42) \quad \Delta X \bullet \Delta S = 0,$$

then

$$(43) \quad \max\{\nu\|U^{-T}\Delta XU^{-1}\|_F, \|U\Delta SU^T\|_F\} \leq \frac{\|H\|_F}{(1 - \sqrt{2}\gamma)}.$$

Proof. Multiply both sides of (41) on the left by U^{-T} and on the right by U^{-1} to obtain

$$U^{-T}\Delta U^T + \Delta U U^{-1} = U^{-T}\Delta XU^{-1}.$$

By this equality and (40), we have

$$(44) \quad \nu U^{-T}\Delta XU^{-1} + U\Delta SU^T = H - \Delta U U^{-1}(USU^T - \nu I) - (USU^T - \nu I)U^{-T}\Delta U^T.$$

Taking the Frobenius norm on both sides of this equality and using (36) and (42), we obtain

$$\begin{aligned} & \max\{\nu\|U^{-T}\Delta XU^{-1}\|_F, \|U\Delta SU^T\|_F\} \\ & \leq (\nu^2\|U^{-T}\Delta XU^{-1}\|_F^2 + \|U\Delta SU^T\|_F^2)^{1/2} \\ & = \|H - \Delta U U^{-1}(USU^T - \nu I) - (USU^T - \nu I)U^{-T}\Delta U^T\|_F \\ & \leq \|H\|_F + 2\|\Delta U U^{-1}\|_F\|USU^T - \nu I\| \\ (45) \quad & \leq \|H\|_F + \sqrt{2}\gamma\nu\|U^{-T}\Delta XU^{-1}\|_F, \end{aligned}$$

which clearly implies that

$$(46) \quad \nu\|U^{-T}\Delta XU^{-1}\|_F \leq \frac{\|H\|_F}{(1 - \sqrt{2}\gamma)}.$$

Using this last inequality to bound the right-hand side of (45), we obtain (43). $\square$

As an immediate consequence of the above lemma, we obtain the following corollary.

COROLLARY 3.11. *Assume that $(\tilde{U}, \tilde{S}) \in \mathcal{U}_{++}^n \times \mathcal{S}^n$ is such that $\|USU^T - \nu I\| < \nu/\sqrt{2}$ for some $\nu > 0$. Then, system (34) has $(\widetilde{\Delta U}, \widetilde{\Delta X}, \widetilde{\Delta S}, \widetilde{\Delta y}) = (0, 0, 0, 0)$ as its unique solution.*

Proof. The last two equations of system (34) imply that $\widetilde{\Delta X} \bullet \widetilde{\Delta S} = 0$. Using this identity and the first two equations of (34), by Lemma 3.10 we easily obtain that $\widetilde{\Delta X} = 0$ and $\widetilde{\Delta S} = 0$, which together with the second equation of (34) and Lemma 3.9 implies that $\widetilde{\Delta U} = 0$. Also, $\widetilde{\Delta y} = 0$ follows from the fact that $\mathcal{A}^*$ is one-to-one and from the last equation of (34). $\square$

We are now ready to establish the analyticity of the path $t \in (0, 1] \rightarrow (\tilde{U}(t), \tilde{X}(t), \tilde{y}(t), \tilde{S}(t))$.

THEOREM 3.12. *Let $(X^*, S^*, y^*) \in \mathcal{F}_P^* \times \mathcal{F}_D^*$ be given. The following hold:*

- (i) the path $t \in (0, 1] \rightarrow \tilde{p}(t) \equiv (\tilde{U}(t), \tilde{X}(t), \tilde{S}(t), \tilde{y}(t))$, where $(\tilde{U}(t), \tilde{X}(t), \tilde{S}(t), \tilde{y}(t))$ is the unique solution of (21), (22), and (33) in $\mathcal{U}_{++}^n \times \mathcal{S}_{++}^n \times \mathcal{S}_{++}^n \times \mathbb{R}^m$, is analytic and also analytic at 0; consequently, $\tilde{p}(t)$ and all its k th-order derivatives, $k \geq 1$, converge as $t \downarrow 0$;
- (ii) $(\tilde{U}^*, \tilde{X}^*, \tilde{S}^*, \tilde{y}^*) \equiv \lim_{t \downarrow 0} (\tilde{U}(t), \tilde{X}(t), \tilde{S}(t), \tilde{y}(t))$ is the unique solution of the system defined by (27), (28), and

$$(47) \quad \mathcal{A}_0(\tilde{X} - X^*) = \begin{pmatrix} 0 \\ 0 \\ \Delta b_3 \end{pmatrix}, \quad \mathcal{A}_0^* \tilde{y} + (\tilde{S} - S^*) = \begin{pmatrix} \Delta C_B \\ 0 \\ 0 \end{pmatrix};$$

- (iii) $(\widetilde{\delta U}^*, \widetilde{\delta X}^*, \widetilde{\delta S}^*, \widetilde{\delta y}^*) \equiv \lim_{t \downarrow 0} (\dot{\tilde{U}}(t), \dot{\tilde{X}}(t), \dot{\tilde{S}}(t), \dot{\tilde{y}}(t))$ is the unique solution of the linear system defined by

$$(48) \quad \begin{aligned} & \widetilde{\delta U}^* \tilde{S}^* (\tilde{U}^*)^T + \tilde{U}^* \tilde{S}^* \widetilde{\delta U}^T + \tilde{U}^* \widetilde{\delta S}^* (\tilde{U}^*)^T \\ & = - \left[\mathcal{L}(\tilde{U}^*) \tilde{S}^* (\tilde{U}^*)^T + \tilde{U}^* \tilde{S}^* \mathcal{L}(\tilde{U}^*)^T \right], \end{aligned}$$

$$(49) \quad \widetilde{\delta U}^T \tilde{U}^* + (\tilde{U}^*)^T \widetilde{\delta U} - \widetilde{\delta X} = - \left[\mathcal{L}(\tilde{U}^*)^T \tilde{U}^* + (\tilde{U}^*)^T \mathcal{L}(\tilde{U}^*) \right],$$

$$(50) \quad \begin{aligned} \mathcal{A}_0 \widetilde{\delta X} &= -\mathcal{B}_0 \tilde{X}^* + \begin{pmatrix} 0 \\ \widetilde{\Delta b}_2 \\ 0 \end{pmatrix}, \\ \mathcal{A}_0^* \widetilde{\delta y} + \widetilde{\delta S} &= -\mathcal{B}_0^* \tilde{y}^* + \begin{pmatrix} 0 \\ \Delta C_{BN} \\ 0 \end{pmatrix}, \end{aligned}$$

where

$$\mathcal{B}_0 \equiv \begin{pmatrix} 0 & \mathcal{A}_{12} & 0 \\ 0 & 0 & \mathcal{A}_{23} \\ 0 & 0 & 0 \end{pmatrix}.$$

Proof. The proof of the theorem is based on Proposition 3.2. Indeed, let $E = \mathcal{U}^n \times \mathcal{S}^n \times \mathcal{S}^n \times \mathbb{R}^m$, $\mathcal{O} = \mathcal{U}_{++}^n \times \mathcal{S}_{++}^n \times \mathcal{S}_{++}^n \times \mathbb{R}^m$, $\delta = \epsilon = 1$, $w : (0, 1) \rightarrow E$ denote the path $t \in (0, 1) \rightarrow (\tilde{U}(t), \tilde{X}(t), \tilde{S}(t), \tilde{y}(t))$, and let $H(w, t) = H(\tilde{U}, \tilde{X}, \tilde{S}, \tilde{y}, t)$ be the map determined by system (21), (22), (33). By Proposition 3.8, the path $\tilde{p}(t) = (\tilde{U}(t), \tilde{X}(t), \tilde{S}(t), \tilde{y}(t))$ has an accumulation point $w^* = (\tilde{U}^*, \tilde{X}^*, \tilde{S}^*, \tilde{y}^*)$ in $\mathcal{O}$ and, by Corollary 3.11, it follows that $H'_w(w^*, 0)$ is nonsingular since (27) with $(\tilde{U}, \tilde{X}, \tilde{S}) = (\tilde{U}^*, \tilde{X}^*, \tilde{S}^*)$ implies that $\|\tilde{U}^* \tilde{S}^* (\tilde{U}^*)^T - \nu I\| = \|W - \nu I\| < \nu/\sqrt{2}$. Hence, (i) and (ii) follow directly from Proposition 3.2. Differentiating the identity $H(\tilde{p}(t), t) = 0$ with respect to t and letting $t \downarrow 0$, we conclude that $\delta w = \delta w^* \equiv (\widetilde{\delta U}^*, \widetilde{\delta X}^*, \widetilde{\delta S}^*, \widetilde{\delta y}^*)$ satisfies

$$H'_w(w^*, 0) \delta w = -H'_t(w^*, 0).$$

Statement (iii) now follows from the fact that $H'_w(w^*, 0)$ is nonsingular and the latter system is equivalent to (48)–(50). $\square$

The proof of Theorem 3.3 is now obvious. Indeed, the analyticity of the map $t \rightarrow (X(t^2), S(t^2))$ follows from (29) and the analyticity of $t \rightarrow (\tilde{X}(t), \tilde{S}(t))$. The analyticity of $t \rightarrow y(t^2)$ follows from the analyticity of $t \rightarrow S(t^2)$ and assumption A.1. The last statement of the theorem is obvious.

In the remainder of this paper, we will let $(\tilde{U}^*, \tilde{X}^*, \tilde{S}^*, \tilde{y}^*)$ and $(\widetilde{\delta U}^*, \widetilde{\delta X}^*, \widetilde{\delta S}^*, \widetilde{\delta y}^*)$ denote the limits of $(\tilde{U}(t), \tilde{X}(t), \tilde{S}(t), \tilde{y}(t))$ and $(\dot{\tilde{U}}(t), \dot{\tilde{X}}(t), \dot{\tilde{S}}(t), \dot{\tilde{y}}(t))$, respectively, as $t \downarrow 0$ (as in Theorem 3.12 above). Observe that Theorem 3.12 provides a characterization of $(\tilde{U}^*, \tilde{X}^*, \tilde{S}^*, \tilde{y}^*)$ as being the unique solution of a certain system of equations which arises by first performing some transformations to the original weighted central path system and then setting $t = 0$ in the resulting system. Hence, it is reasonable to expect that the linear equations (47) can be entirely described in terms of the original data $(W, \mathcal{A}, C, \Delta C, b, \Delta b)$. Indeed, the following result gives this alternative description of (47).

THEOREM 3.13. *$(\tilde{U}^*, \tilde{X}^*, \tilde{S}^*)$ is the unique solution of the system given by (27), (28) and the linear equations*

(51)

$$\mathcal{A}_B(\tilde{X}_B) = b, \quad \mathcal{A}_{BN}(\tilde{X}_{BN}) \in \text{Im}(\mathcal{A}_B), \quad \mathcal{A}_N(\tilde{X}_N) \in \Delta b + \text{Im}(\mathcal{A}_B \mathcal{A}_{BN}),$$

(52)

$$\begin{aligned} \tilde{S}_B &\in \Delta C_B + \text{Im}(\mathcal{A}_B^*), \quad \begin{pmatrix} 0 \\ \tilde{S}_{BN} \end{pmatrix} \in \text{Im} \left[\begin{pmatrix} \mathcal{A}_B^* \\ \mathcal{A}_{BN}^* \end{pmatrix} \right], \\ \begin{pmatrix} 0 \\ \tilde{S}_N \end{pmatrix} &\in C + \text{Im} \left[\begin{pmatrix} \mathcal{A}_B^* \\ \mathcal{A}_{BN}^* \\ \mathcal{A}_N^* \end{pmatrix} \right]. \end{aligned}$$

Proof. From Theorem 3.12(ii), it suffices to show that (47) is equivalent to (51) and (52). Since the first equation of (47) is the same as (31) with $t = 0$, we have that the first equation of (47) holds if and only if

$$(53) \quad \mathcal{A}_{11}(\tilde{X}_B) = \mathcal{A}_{11}(X_B^*), \quad \mathcal{A}_{22}(\tilde{X}_{BN}) = 0, \quad \mathcal{A}_{33}(\tilde{X}_N) = \widetilde{\Delta b}_3.$$

By Lemma 3.6, the first identity in (53) can be written as

$$(T \circ \mathcal{A}) \begin{pmatrix} \tilde{X}_B \\ 0 \\ 0 \end{pmatrix} = (T \circ \mathcal{A}) \begin{pmatrix} X_B^* \\ 0 \\ 0 \end{pmatrix},$$

and hence it is equivalent to $\mathcal{A}_B(\tilde{X}_B) = \mathcal{A}_B(X_B^*) = b$, in view of relation (30) and the fact that T is an isomorphism. By Lemma 3.6 and the fact that $\mathcal{A}_{11}$ is onto, the second identity in (53) holds if and only if

$$(T \circ \mathcal{A}) \begin{pmatrix} \hat{X}_B \\ \tilde{X}_{BN} \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

for some $\hat{X}_B \in \mathcal{S}^{|B|}$, and hence it is equivalent to $\mathcal{A}_{BN}(\tilde{X}_{BN}) \in \text{Im}(\mathcal{A}_B)$, in view of (30) and the fact that T is an isomorphism. Using Lemma 3.6 again and the fact that $\mathcal{A}_{11}$ and $\mathcal{A}_{22}$ are onto, we easily see that the last identity in (53) holds if and only if

$$(T \circ \mathcal{A}) \begin{pmatrix} \hat{X}_B \\ \tilde{X}_{BN} \\ \tilde{X}_N \end{pmatrix} = \begin{pmatrix} \widetilde{\Delta b}_1 \\ \widetilde{\Delta b}_2 \\ \widetilde{\Delta b}_3 \end{pmatrix} = T(\Delta b)$$

for some $(\check{X}_B, \check{X}_{BN}) \in \mathcal{S}^{|B|} \times \mathfrak{R}^{|B| \times |N|}$, and hence it is equivalent to $\mathcal{A}_N(\check{X}_N) \in \Delta b + \text{Im}(\mathcal{A}_B \mathcal{A}_{BN})$, in view of (30) and the fact that T is an isomorphism. We have thus shown that the first equation of (47) is equivalent to (51).

The fact that the second equation of (47) holds if and only if (52) holds can be proved in a similar way as above. $\square$

The following result gives an alternative characterization of $(\widetilde{\delta U}^*, \widetilde{\delta X}^*, \widetilde{\delta S}^*)$ involving the original data $(W, \mathcal{A}, C, \Delta C, b, \Delta b)$.

THEOREM 3.14. *$(\widetilde{\delta U}^*, \widetilde{\delta X}^*, \widetilde{\delta S}^*)$ is the unique solution of the linear system of equations (48), (49) and*

$$(54) \quad \begin{bmatrix} \mathcal{A}_B & \mathcal{A}_{BN} \end{bmatrix} \begin{bmatrix} \widetilde{\delta X}_B \\ \widetilde{X}_{BN}^* \end{bmatrix} = 0, \quad \begin{bmatrix} \mathcal{A}_{BN} & \mathcal{A}_N \end{bmatrix} \begin{bmatrix} \widetilde{\delta X}_{BN} \\ \widetilde{X}_N^* \end{bmatrix} \in \Delta b + \text{Im}(\mathcal{A}_B),$$

$$(55) \quad \begin{aligned} & \mathcal{A}_N(\widetilde{\delta X}_N) \in \text{Im}(\mathcal{A}_B \mathcal{A}_{BN}), \\ & \widetilde{\delta S}_B \in \text{Im}(\mathcal{A}_B^*), \quad \begin{pmatrix} \widetilde{S}_B^* \\ \widetilde{\delta S}_{BN} \end{pmatrix} \in \begin{pmatrix} \Delta C_B \\ \Delta C_{BN} \end{pmatrix} + \text{Im} \left[\begin{pmatrix} \mathcal{A}_B^* \\ \mathcal{A}_{BN}^* \end{pmatrix} \right], \\ & \begin{pmatrix} 0 \\ \widetilde{S}_{BN}^* \\ \widetilde{\delta S}_N \end{pmatrix} \in \text{Im} \left[\begin{pmatrix} \mathcal{A}_B^* \\ \mathcal{A}_{BN}^* \\ \mathcal{A}_N^* \end{pmatrix} \right]. \end{aligned}$$

Proof. From Theorem 3.12(iii), it suffices to show that (50) is equivalent to (54) and (55). Observe that the first equation of (50) can be written as

$$(56) \quad \begin{aligned} \mathcal{A}_{11}(\widetilde{\delta X}_B) + \mathcal{A}_{12}(\widetilde{X}_{BN}^*) &= 0, \\ \mathcal{A}_{22}(\widetilde{\delta X}_{BN}) + \mathcal{A}_{23}(\widetilde{X}_N^*) &= \widetilde{\Delta b}_2, \\ \mathcal{A}_{33}(\widetilde{\delta X}_N) &= 0. \end{aligned}$$

Using Lemma 3.6, the fact that $\mathcal{A}_{11}$ and $\mathcal{A}_{22}$ are onto, and the identities $\mathcal{A}_{22}\widetilde{X}_{BN}^* = 0$ and $\mathcal{A}_{33}\widetilde{X}_N^* = \widetilde{\Delta b}_3$ which hold in view of Theorem 3.12(ii), we easily see that the above three equations are, respectively, equivalent to

$$\begin{aligned} (T \circ \mathcal{A}) \begin{pmatrix} \widetilde{\delta X}_B \\ \widetilde{X}_{BN}^* \\ 0 \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad (T \circ \mathcal{A}) \begin{pmatrix} \hat{X}_B \\ \widetilde{\delta X}_{BN} \\ \widetilde{X}_N^* \end{pmatrix} = \begin{pmatrix} \widetilde{\Delta b}_1 \\ \widetilde{\Delta b}_2 \\ \widetilde{\Delta b}_3 \end{pmatrix}, \\ (T \circ \mathcal{A}) \begin{pmatrix} \check{X}_B \\ \check{X}_{BN} \\ \widetilde{\delta X}_N \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \end{aligned}$$

for some $\hat{X}_B, \check{X}_B \in \mathcal{S}^{|B|}$ and $\check{X}_{BN} \in \mathfrak{R}^{|B| \times |N|}$. The latter conditions in turn are equivalent to (54) in view of (30) and the facts that $\widetilde{\Delta b} = T(\Delta b)$ and T is an isomorphism.

Using similar arguments as to the ones used above, it can be shown that the second equation of (50) holds if and only if (55) holds. $\square$

4. Limiting behavior of the derivative of the weighted central path.

In this section, we will first characterize the limit of the normalized derivatives of a weighted central path as ν approaches 0. We then show that a weighted central path can have two types of behavior: it converges either as $\Theta(\nu)$ or as $\Theta(\sqrt{\nu})$ depending on whether the matrix W on a certain scaled space is block diagonal or not, respectively.

THEOREM 4.1. $\lim_{\nu \downarrow 0} \sqrt{\nu} (\dot{X}(\nu), \dot{S}(\nu), \dot{y}(\nu))$ exists and satisfies

$$(57) \quad \lim_{\nu \downarrow 0} \sqrt{\nu} \dot{X}(\nu) = \begin{bmatrix} \widetilde{\delta X}_B^*/2 & \tilde{X}_{BN}^*/2 \\ \tilde{X}_{NB}^*/2 & 0 \end{bmatrix}, \quad \lim_{\nu \downarrow 0} \sqrt{\nu} \dot{S}(\nu) = \begin{bmatrix} 0 & \tilde{S}_{BN}^*/2 \\ \tilde{S}_{NB}^*/2 & \widetilde{\delta S}_N^*/2 \end{bmatrix}.$$

Proof. By (29), we have

$$(58) \quad X(t^2) = \begin{bmatrix} \tilde{X}_B(t) & t\tilde{X}_{BN}(t) \\ t\tilde{X}_{NB}(t) & t^2\tilde{X}_N(t) \end{bmatrix}, \quad S(t^2) = \begin{bmatrix} t^2\tilde{S}_B(t) & t\tilde{S}_{BN}(t) \\ t\tilde{S}_{NB}(t) & \tilde{S}_N(t) \end{bmatrix}.$$

Differentiating both sides with respect to t , letting $t \downarrow 0$, and using Theorem 3.12, we obtain (57) upon letting $\nu = t^2$. $\square$

We establish one technical lemma as follows, which gives a characterization of block diagonal weighted matrix W . This lemma will play a crucial role in further analyzing the limiting behavior of derivatives of the weighted central path.

LEMMA 4.2. *The following statements hold:*

- (i) $\tilde{X}_{BN}^* \bullet \tilde{S}_{BN}^* = 0$;
- (ii) $\tilde{X}_{BN}^* = \tilde{S}_{BN}^* = 0$ if and only if $W_{BN} = 0$.

Proof. Statement (i) follows from the fact that $\tilde{X}_{BN}^*$ and $\tilde{S}_{BN}^*$ satisfy the second equations in (51) and (52), respectively, which can be easily seen to determine two orthogonal complementary subspaces in $\mathbb{R}^{|B| \times |N|}$.

We now show (ii). Using the fact that $(\tilde{U}^*, \tilde{X}^*, \tilde{S}^*)$ satisfies (27) and (28), it is easy to see that

$$(59) \quad W_{BN} = \tilde{U}_B^* \tilde{S}_{BN}^* \tilde{U}_N^* + \tilde{U}_{BN}^* \tilde{S}_N^* \tilde{U}_N^*, \quad \tilde{X}_B^* = (\tilde{U}_B^*)^2, \quad \tilde{X}_{BN}^* = \tilde{U}_B^* \tilde{U}_{BN}^*.$$

By Proposition 3.5, we know that $(\tilde{X}^*, \tilde{S}^*, \tilde{U}^*) \in \mathcal{S}_{++}^n \times \mathcal{S}_{++}^n \times \mathcal{U}_{++}^n$, and hence $\tilde{X}_B^* \succ 0$, $\tilde{S}_N^* \succ 0$, $\tilde{U}_B^* \succ 0$, $\tilde{U}_N^* \succ 0$. Thus, the last relation in (59) implies that

$$(60) \quad \tilde{X}_{BN}^* = 0 \iff \tilde{U}_{BN}^* = 0.$$

Assume first that $\tilde{X}_{BN}^* = \tilde{S}_{BN}^* = 0$. Then, (60) and the first equation in (59) immediately imply that $W_{BN} = 0$. Assume now that $W_{BN} = 0$. Using (59), we obtain

$$\begin{aligned} \tilde{X}_B^* \tilde{S}_{BN}^* + \tilde{X}_{BN}^* \tilde{S}_N^* &= (\tilde{U}_B^*)^2 \tilde{S}_{BN}^* + \tilde{U}_B^* \tilde{U}_{BN}^* \tilde{S}_N^* \\ &= \tilde{U}_B^* \left(\tilde{U}_B^* \tilde{S}_{BN}^* \tilde{U}_N^* + \tilde{U}_{BN}^* \tilde{S}_N^* \tilde{U}_N^* \right) (\tilde{U}_N^*)^{-1} \\ &= \tilde{U}_B^* W_{BN} (\tilde{U}_N^*)^{-1} = 0. \end{aligned}$$

Multiplying the above equation on the left by $(\tilde{X}_B^*)^{-1/2}$ and on the right by $(\tilde{S}_N^*)^{-1/2}$, squaring both sides of the resulting expression, and using (i), we conclude that

$$\|(\tilde{X}_B^*)^{1/2} \tilde{S}_{BN}^* (\tilde{S}_N^*)^{-1/2}\|_F^2 + \|(\tilde{X}_B^*)^{-1/2} \tilde{X}_{BN}^* (\tilde{S}_N^*)^{1/2}\|_F^2 = 0,$$

from which it follows that $\tilde{X}_{BN}^* = \tilde{S}_{BN}^* = 0$. $\square$

From Lemma 4.2 and Theorem 4.1, the following corollary follows.

COROLLARY 4.3. *If $W_{BN} \neq 0$, then at least one of the limits in (57) is nonzero and*

$$\|(X(\nu), S(\nu), y(\nu)) - (X^*, S^*, y^*)\| = \Theta(\sqrt{\nu}).$$

Proof. Assume that $W_{BN} \neq 0$. By Lemma 4.2(ii), we have that either $\tilde{X}_{BN}^* \neq 0$ or $\tilde{S}_{BN}^* \neq 0$, which together with (57) implies the first claim of the corollary. The second claim follows directly from (57) and the equality

$$\lim_{\nu \downarrow 0} \frac{(X(\nu), S(\nu), y(\nu)) - (X^*, S^*, y^*)}{\sqrt{\nu}} = \lim_{\nu \downarrow 0} 2\sqrt{\nu} \begin{pmatrix} \dot{X}(\nu) \\ \dot{S}(\nu) \\ \dot{y}(\nu) \end{pmatrix},$$

which holds due to Theorem 3.3. $\square$

From Corollary 4.3, we immediately see that the weighted central path as a function of ν in general cannot be extended analytically to an interval of the form $(-\epsilon, \infty)$ for some $\epsilon > 0$. Theorem 4.1 and Corollary 4.3 give a precise characterization of how the primal-dual weighted central path approaches its limit (X^*, S^*, y^*) for the case when W is non-block diagonal, that is, $W_{BN} \neq 0$. However, it is still possible for one of the limits in (57) to be equal to zero in this situation. The following result claims that in this case the corresponding primal or dual weighted central path converges toward (X^*, S^*, y^*) at a $\Theta(\nu)$ rate of convergence.

THEOREM 4.4. *The following statements hold:*

(i) *If $\lim_{\nu \downarrow 0} \sqrt{\nu} \dot{X}(\nu) = 0$, then $X(\nu) - X^* = \Theta(\nu)$ and*

$$(61) \quad \lim_{\nu \downarrow 0} \dot{X}(\nu) = \begin{bmatrix} \widetilde{\delta^{(2)}X_B^*}/2 & \widetilde{\delta X_{BN}^*} \\ (\widetilde{\delta X_{BN}^*})^T & \tilde{X}_N^* \end{bmatrix},$$

where $\widetilde{\delta^{(2)}X_B^*} \equiv \lim_{t \downarrow 0} \ddot{X}_B(t)$;

(ii) *If $\lim_{\nu \downarrow 0} \sqrt{\nu} \dot{S}(\nu) = 0$, then $\|(S(\nu), y(\nu)) - (S^*, y^*)\| = \Theta(\nu)$ and*

$$(62) \quad \lim_{\nu \downarrow 0} \dot{S}(\nu) = \begin{bmatrix} \widetilde{\delta^{(2)}S_B^*}/2 & \widetilde{\delta S_{BN}^*} \\ (\widetilde{\delta S_{BN}^*})^T & \tilde{S}_N^* \end{bmatrix},$$

where $\widetilde{\delta^{(2)}S_B^*} \equiv \lim_{t \downarrow 0} \ddot{S}_B(t)$.

Proof. To prove (i), assume that $\lim_{\nu \downarrow 0} \sqrt{\nu} \dot{X}(\nu) = 0$. By Theorem 4.1, we must have $\widetilde{\delta X_B^*} = 0$ and $\tilde{X}_{BN}^* = 0$. Differentiating both sides of the first identity in (58) with respect to t and then dividing the resulting identity by $2t$, we obtain that

$$\dot{X}(t^2) = \begin{bmatrix} \dot{X}_B(t)/(2t) & \tilde{X}_{BN}(t)/(2t) + \dot{X}_{BN}(t)/2 \\ \tilde{X}_{BN}(t)^T/(2t) + \dot{\tilde{X}}_{BN}(t)^T/2 & \tilde{X}_N(t) + t\dot{\tilde{X}}_N(t)/2 \end{bmatrix}.$$

Using the fact that $\widetilde{\delta X_B^*} = 0$ and $\tilde{X}_{BN}^* = 0$ and using Theorem 3.12, we obtain (61) upon letting $\nu = t^2 \downarrow 0$. The conclusion that $X(\nu) - X^* = \Theta(\nu)$ follows immediately from (61) and the fact that $\tilde{X}_N^* \succ 0$. Using the same arguments as above and assumption A.1, we can similarly show (ii). $\square$

The remainder of this section considers the case when W is block diagonal, that is, $W_{BN} = 0$. We will show in this case that two limits in (57) are equal to zero and hence that $\lim_{\nu \downarrow 0} (\dot{X}(\nu), \dot{S}(\nu), \dot{y}(\nu))$ exists and $\|(X(\nu), S(\nu), y(\nu)) - (X^*, S^*, y^*)\| = \Theta(\nu)$ due to Theorem 4.4.

Note that to establish the above claim, it suffices to show that $\widetilde{\delta X_B^*} = 0$ and $\widetilde{\delta S_N^*} = 0$ in view of Lemma 4.2(ii). Before showing this fact, we state two technical results from Monteiro and Tsuchiya [30].

LEMMA 4.5 (Lemma 2.1 of [30]). *For every $A \in \mathcal{S}_{++}^n$ and $H \in \mathcal{S}^n$, the equation $AU + UA = H$ has a unique solution $U \in \mathcal{S}^n$. Moreover, this solution satisfies $\|AU\|_F \leq \|H\|_F/\sqrt{2}$.*

LEMMA 4.6 (Lemma 2.3 of [30]). *Suppose that $\gamma \in [0, 1/\sqrt{2})$ and that $(U, S) \in \mathcal{S}_{++}^n \times \mathcal{S}_{++}^n$ is such that $\|USU - \nu I\| \leq \gamma\nu$ for some $\nu > 0$. If $(\Delta X, \Delta S, \Delta U) \in \mathcal{S}^n \times \mathcal{S}^n \times \mathcal{S}^n$ is a solution of the system*

$$\begin{aligned}\Delta USU + US\Delta U + U\Delta SU &= H, \\ \Delta UU + U\Delta U &= \Delta X, \\ \Delta X \bullet \Delta S &= 0\end{aligned}$$

and $H \in \mathcal{S}^n$, then

$$\max \{ \nu \|U^{-1}\Delta XU^{-1}\|_F, \|U\Delta SU\|_F \} \leq \frac{\|H\|_F}{(1 - \sqrt{2}\gamma)}.$$

We are now ready to show that $\widetilde{\delta X}_B^* = 0$ and $\widetilde{\delta S}_N^* = 0$.

LEMMA 4.7. *If $W_{BN} = 0$, then*

$$\begin{aligned}\widetilde{\delta X}_B^* &= \widetilde{\delta S}_B^* = \widetilde{\delta U}_B^* = 0, \\ \widetilde{\delta X}_N^* &= \widetilde{\delta S}_N^* = \widetilde{\delta U}_N^* = 0.\end{aligned}$$

Proof. From Lemma 4.2(ii), we know that $\tilde{X}_{BN}^* = \tilde{S}_{BN}^* = 0$. Using this identity and the fact that $(\widetilde{\delta X}_B^*, \widetilde{\delta S}_B^*)$ and $(\widetilde{\delta X}_N^*, \widetilde{\delta S}_N^*)$ satisfy the first and third equations of (54) and (55), respectively, we obtain that

$$(63) \quad \widetilde{\delta X}_B^* \bullet \widetilde{\delta S}_B^* = 0, \quad \widetilde{\delta X}_N^* \bullet \widetilde{\delta S}_N^* = 0.$$

By (60), $\tilde{X}_{BN}^* = 0$ implies $\tilde{U}_{BN}^* = 0$, and thus $\mathcal{L}(\tilde{U}^*) = 0$. Hence, by (48) and (49), we have

$$\begin{aligned}\widetilde{\delta U}^* \tilde{S}^* (\tilde{U}^*)^T + \tilde{U}^* \tilde{S}^* (\widetilde{\delta U}^*)^T + \tilde{U}^* \widetilde{\delta S}^* (\tilde{U}^*)^T &= 0, \\ (\widetilde{\delta U}^*)^T \tilde{U}^* + (\tilde{U}^*)^T \widetilde{\delta U}^* &= \widetilde{\delta X}^*.\end{aligned}$$

These equations together with the fact that $\tilde{S}_{BN}^* = 0$ and $\tilde{U}_{BN}^* = 0$ can be easily seen to imply that

$$(64) \quad \widetilde{\delta U}_B^* \tilde{S}_B^* \tilde{U}_B^* + \tilde{U}_B^* \tilde{S}_B^* \widetilde{\delta U}_B^* + \tilde{U}_B^* \widetilde{\delta S}_B^* \tilde{U}_B^* = 0,$$

$$(65) \quad \widetilde{\delta U}_B^* \tilde{U}_B^* + \tilde{U}_B^* \widetilde{\delta U}_B^* = \widetilde{\delta X}_B^*.$$

Moreover, by (27), $(\tilde{U}, \tilde{X}, \tilde{S}) = (\tilde{U}^*, \tilde{X}^*, \tilde{S}^*)$, and the fact that $\tilde{S}_{BN}^* = \tilde{U}_{BN}^* = 0$, we have

$$\tilde{U}_B^* \tilde{S}_B^* \tilde{U}_B^* = W_B,$$

which together with the assumption that $\|W - \nu I\| < \nu/\sqrt{2}$ for some $\nu > 0$ and the fact that $\tilde{U}^* \in \mathcal{U}_{++}^n$ implies

$$\|\tilde{U}_B^* \tilde{S}_B^* \tilde{U}_B^* - \nu I\| = \|W_B - \nu I\| < \nu/\sqrt{2}$$

and $(\tilde{U}_B^*, \tilde{S}_B^*) \in \mathcal{S}_{++}^{|B|} \times \mathcal{S}_{++}^{|B|}$. Using the conclusions above, relations (64) and (65), the first identity in (63), together with Lemmas 4.5 and 4.6 with $H = 0$, we conclude that $\delta\tilde{X}_B^* = \delta\tilde{S}_B^* = \delta\tilde{U}_B^* = 0$. Using similar arguments, we can also show that $\delta\tilde{X}_N^* = \delta\tilde{S}_N^* = \delta\tilde{U}_N^* = 0$. $\square$

As a consequence of the results obtained above, we have the following theorem when $W_{BN} = 0$.

THEOREM 4.8. *If $W_{BN} = 0$, then the primal-dual weighted central path $(X(\nu), y(\nu), S(\nu))$ satisfies*

- (i) $\lim_{\nu \downarrow 0} \sqrt{\nu} (\dot{X}(\nu), \dot{S}(\nu)) = 0$;
- (ii) $\|(X(\nu), S(\nu), y(\nu)) - (X^*, S^*, y^*)\| = \Theta(\nu)$;
- (iii) $\lim_{\nu \downarrow 0} (\dot{X}(\nu), \dot{S}(\nu), \dot{y}(\nu))$ exists and (61) and (62) hold.

Proof. Using Lemma 4.2(ii), Lemma 4.7, and the condition $W_{BN} = 0$, we obtain that $\tilde{X}_{BN}^* = \tilde{S}_{BN}^* = 0$, $\delta\tilde{X}_B^* = 0$, and $\delta\tilde{S}_N^* = 0$. Consequently, by Theorem 4.1, (i) immediately follows. Statements (ii) and (iii) follow directly from (i) and Theorem 4.4. $\square$

5. Error bound analysis. By strengthening some of the results of the previous sections, in this section we derive a new error bound on the distance of a point lying in a certain neighborhood of the central path to the primal-dual optimal set.

For any given nonempty compact set $\mathcal{K} \subset \mathcal{G}_{++}$ and constants $\gamma, \tau > 0$, define

$$\mathcal{N}(\gamma, \tau, \mathcal{K}) \equiv \{ (X, S, y) \in \mathcal{S}_{++}^n \times \mathcal{S}_{++}^n \times \mathbb{R}^m : \mathcal{G}(X, S, y) \in \tau\mathcal{K}, \\ \|X^{1/2}SX^{1/2} - \tau I\| \leq \gamma\tau \},$$

where the map $\mathcal{G}$ and the set $\mathcal{G}_{++}$ are defined in (7) and (8), respectively.

Observe that the set $\cup_{\tau>0} \mathcal{N}(\gamma, \tau, \mathcal{K})$ forms a neighborhood of the primal-dual central path. This neighborhood and related ones have been frequently used in the development of primal-dual interior-point algorithms for SDP.

The following result gives a new error bound on the distance of a point lying in $\mathcal{N}(\gamma, \tau, \mathcal{K})$ to the primal-dual optimal set $\mathcal{F}_P^* \times \mathcal{F}_D^*$. Its proof will be given at the end of this section after we have derived stronger versions of the results of the previous sections.

THEOREM 5.1. *Let $\gamma \in (0, 1/\sqrt{2})$ and any nonempty compact set $\mathcal{K} \subset \mathcal{G}_{++}$ be given. Then, there exists a constant $M = M(\gamma, \mathcal{K}) > 0$ such that*

$$(66) \quad \text{dist}((X, S, y), \mathcal{F}_P^* \times \mathcal{F}_D^*) \leq M(\tau + \sqrt{\tau}\|W_{BN}\|)$$

for every $\tau \in (0, 1]$ and $(X, S, y) \in \mathcal{N}(\gamma, \tau, \mathcal{K})$, where $W = W(X, S, \tau) \equiv X^{1/2}SX^{1/2}/\tau$.

In view of Proposition 2.1, for each $(\nu, W, \Delta C, \Delta b) \in (0, 1] \times \mathcal{W} \times \mathcal{G}_{++}$, the system of nonlinear equations (3)–(5) has a unique solution, which in this section we denote by $(X(\nu, W, \Delta C, \Delta b), S(\nu, W, \Delta C, \Delta b), y(\nu, W, \Delta C, \Delta b))$ in order to emphasize and study its dependence on $(W, \Delta C, \Delta b)$. Moreover, in view of Theorem 3.3, the limit

$$\lim_{\nu \downarrow 0} (X(\nu, W, \Delta C, \Delta b), S(\nu, W, \Delta C, \Delta b), y(\nu, W, \Delta C, \Delta b)),$$

denoted by $(X(0, W, \Delta C, \Delta b), S(0, W, \Delta C, \Delta b), y(0, W, \Delta C, \Delta b))$, exists for every $(W, \Delta C, \Delta b) \in \mathcal{W} \times \mathcal{G}_{++}$. Hence, the functions $X(\cdot, \cdot, \cdot, \cdot)$, $S(\cdot, \cdot, \cdot, \cdot)$, and $y(\cdot, \cdot, \cdot, \cdot)$ are well defined in the set of points $[0, 1] \times \mathcal{W} \times \mathcal{G}_{++}$. In an obvious way, we can also define the functions $\tilde{X}(t, W, \Delta C, \Delta b)$, $\tilde{S}(t, W, \Delta C, \Delta b)$, and $\tilde{y}(t, W, \Delta C, \Delta b)$ over the set of points $(t, W, \Delta C, \Delta b) \in [0, 1] \times \mathcal{W} \times \mathcal{G}_{++}$.

It turns out that reparametrizations of the above functions are analytic according to the following definition. We say that a function $f : \Omega \subseteq E \rightarrow F$, where E, F are two finite dimensional normed vector spaces, is analytic if there exist an open set $\mathcal{O} \subseteq E$ containing Ω and an analytic function $\tilde{f} : \mathcal{O} \rightarrow F$ such that $\tilde{f}$ restricted to Ω is equal to f .

THEOREM 5.2. *The following hold:*

- (i) *the map $(t, W, \Delta C, \Delta b) \in [0, 1] \times \mathcal{W} \times \mathcal{G}_{++} \rightarrow (\tilde{X}(t, W, \Delta C, \Delta b), \tilde{S}(t, W, \Delta C, \Delta b), \tilde{y}(t, W, \Delta C, \Delta b))$ is analytic;*
- (ii) *the map $(t, W, \Delta C, \Delta b) \in [0, 1] \times \mathcal{W} \times \mathcal{G}_{++} \rightarrow (X(t^2, W, \Delta C, \Delta b), S(t^2, W, \Delta C, \Delta b), y(t^2, W, \Delta C, \Delta b))$ is analytic.*

Proof. The proof of the theorem is identical to the proof of Theorems 3.12 and 3.3, except that when invoking the implicit function theorem, we should view $(t, W, \Delta C, \Delta b)$ as the parameter vector. $\square$

Let

$$\mathcal{W}^b \equiv \{W \in \mathcal{W} : W_{BN} = 0\},$$

where $\mathcal{W}$ is as defined in (9). One important result that we will establish next is that the function $(t, W, \Delta C, \Delta b) \in [0, 1] \times \mathcal{W}^b \times \mathcal{G}_{++} \rightarrow (X'(t^2, W, \Delta C, \Delta b), S'(t^2, W, \Delta C, \Delta b))$ is analytic. We emphasize that this result holds only over the smaller set $[0, 1] \times \mathcal{W}^b \times \mathcal{G}_{++}$. Note also that this result does not follow immediately from Theorem 5.2(ii) since the derivative of the function in Theorem 5.2(ii) is not equal to the above function.

We now state a simple but crucial technical result needed to establish the result stated in the previous paragraph.

PROPOSITION 5.3. *Let $f : I \times E \rightarrow F$ be a given analytic function, where $I \subset \mathbb{R}$ is an interval and E, F are two finite dimensional normed vector spaces. Then, for any $t^* \in I$, the function $g : I \times E \rightarrow F$ defined as*

$$g(t, u) = \begin{cases} \frac{f(t, u) - f(t^*, u)}{t - t^*} & \text{if } t \neq t^*, \\ \frac{\partial f}{\partial t}(t^*, u) & \text{if } t = t^*, \end{cases}$$

is analytic.

We are now ready to establish the result alluded to just before Proposition 5.3.

LEMMA 5.4. *The following hold:*

- (i) *for any $(W, \Delta C, \Delta b) \in \mathcal{W}^b \times \mathcal{G}_{++}$, we have*

$$\lim_{t \downarrow 0} \left(t \frac{\partial X}{\partial \nu}(t^2, W, \Delta C, \Delta b), t \frac{\partial S}{\partial \nu}(t^2, W, \Delta C, \Delta b) \right) = 0;$$

- (ii) *$(t, W, \Delta C, \Delta b) \in [0, 1] \times \mathcal{W}^b \times \mathcal{G}_{++} \rightarrow (X'(t^2, W, \Delta C, \Delta b), S'(t^2, W, \Delta C, \Delta b))$ is analytic.*

Proof. In view of Theorem 4.8(i), we easily see that (i) holds. Since partial derivatives of an analytic function are also analytic, it follows from Theorem 5.2(ii) that the functions

$$\begin{aligned} (t, W, \Delta C, \Delta b) \in [0, 1] \times \mathcal{W}^b \times \mathcal{G}_{++} &\rightarrow (t \frac{\partial X}{\partial \nu}(t^2, W, \Delta C, \Delta b), t \frac{\partial S}{\partial \nu}(t^2, W, \Delta C, \Delta b)), \\ (t, W, \Delta C, \Delta b) \in [0, 1] \times \mathcal{W}^b \times \mathcal{G}_{++} &\rightarrow (\frac{\partial X}{\partial W}(t^2, W, \Delta C, \Delta b), \frac{\partial S}{\partial W}(t^2, W, \Delta C, \Delta b)) \end{aligned}$$

are both analytic. Using (i) and Proposition 5.3, we conclude that the first function above divided by t is also analytic. We have thus shown that (ii) holds. $\square$

For $\gamma > 0$, let

$$\mathcal{W}(\gamma) \equiv \{W \in \mathcal{S}_{++}^n : \|W - I\| \leq \gamma\}, \quad \mathcal{W}^b(\gamma) \equiv \{W \in \mathcal{W}(\gamma) : W_{BN} = 0\}.$$

We can easily see that if $\gamma < 1/\sqrt{2}$, then $\mathcal{W}(\gamma)$ and $\mathcal{W}^b(\gamma)$ are convex compact subsets of $\mathcal{W}$ and $\mathcal{W}^b$, respectively. For the remainder of this section, we let $\mathcal{K} \subset \mathcal{G}_{++}$ be any given nonempty compact set.

The next two results provide estimates on the sizes of the blocks of the matrices $X(\nu, W, \Delta C, \Delta b)$ and $S(\nu, W, \Delta C, \Delta b)$ first when $(W, \Delta C, \Delta b) \in \mathcal{W}^b(\gamma) \times \mathcal{K}$ and then for a general $(W, \Delta C, \Delta b) \in \mathcal{W}(\gamma) \times \mathcal{K}$.

LEMMA 5.5. *Let $\gamma \in (0, 1/\sqrt{2})$ be given. Then, for all $(\nu, W, \Delta C, \Delta b) \in [0, 1] \times \mathcal{W}^b(\gamma) \times \mathcal{K}$, there holds*

$$\|(X(\nu, W, \Delta C, \Delta b), S(\nu, W, \Delta C, \Delta b)) - (X(0, W, \Delta C, \Delta b), S(0, W, \Delta C, \Delta b))\| = \mathcal{O}(\nu).$$

Proof. By the mean value theorem, we have

$$\begin{aligned} & \|(X(\nu, W, \Delta C, \Delta b), S(\nu, W, \Delta C, \Delta b)) - (X(0, W, \Delta C, \Delta b), S(0, W, \Delta C, \Delta b))\| \\ & \leq \sup_{\theta \in [0, 1]} \|(X'(\theta\nu, W, \Delta C, \Delta b), S'(\theta\nu, W, \Delta C, \Delta b))\| \nu. \end{aligned}$$

By Theorem 5.4(ii) and the fact that $\mathcal{W}^b(\gamma) \times \mathcal{K}$ is compact, there exists a constant $M = M(\gamma, \mathcal{K}) > 0$ such that $\|(X'(\theta\nu, W, \Delta C, \Delta b), S'(\theta\nu, W, \Delta C, \Delta b))\| \leq M$ for all $(\theta, \nu, W, \Delta C, \Delta b) \in [0, 1] \times [0, 1] \times \mathcal{W}^b(\gamma) \times \mathcal{K}$. Hence, the lemma follows. $\square$

LEMMA 5.6. *For all $(\nu, W, \Delta C, \Delta b) \in [0, 1] \times \mathcal{W}(\gamma) \times \mathcal{K}$, there hold*

$$\begin{aligned} (67) \quad & X(\nu, W, \Delta C, \Delta b) - X(0, W^b, \Delta C, \Delta b) \\ & = \begin{pmatrix} \mathcal{O}(\|W_{BN}\|) & \mathcal{O}(\sqrt{\nu}\|W_{BN}\|) \\ \mathcal{O}(\sqrt{\nu}\|W_{BN}\|) & \mathcal{O}(\nu\|W_{BN}\|) \end{pmatrix} + \mathcal{O}(\nu), \end{aligned}$$

$$\begin{aligned} (68) \quad & S(\nu, W, \Delta C, \Delta b) - S(0, W^b, \Delta C, \Delta b) \\ & = \begin{pmatrix} \mathcal{O}(\nu\|W_{BN}\|) & \mathcal{O}(\sqrt{\nu}\|W_{BN}\|) \\ \mathcal{O}(\sqrt{\nu}\|W_{BN}\|) & \mathcal{O}(\|W_{BN}\|) \end{pmatrix} + \mathcal{O}(\nu), \end{aligned}$$

where

$$(69) \quad W^b \equiv \begin{pmatrix} W_B & 0 \\ 0 & W_N \end{pmatrix}.$$

Proof. By Theorem 5.2(i), we know that $(\tilde{X}(t, W, \Delta C, \Delta b), \tilde{S}(t, W, \Delta C, \Delta b))$ is analytic over $[0, 1] \times \mathcal{W} \times \mathcal{K}$. Hence, its derivative function $(\tilde{X}'(t, W, \Delta C, \Delta b), \tilde{S}'(t, W, \Delta C, \Delta b))$ is analytic, and hence continuous, over $[0, 1] \times \mathcal{W} \times \mathcal{K}$. Since $[0, 1] \times \mathcal{W}(\gamma) \times \mathcal{K}$ is compact, there exists a constant $M = M(\gamma, \mathcal{K}) > 0$ such that $\|(\tilde{X}'(\nu, W, \Delta C, \Delta b), \tilde{S}'(\nu, W, \Delta C, \Delta b))\| \leq M$ for all $(t, W, \Delta C, \Delta b) \in [0, 1] \times \mathcal{W}(\gamma) \times \mathcal{K}$. This together with (69) and the mean value theorem implies

$$\begin{aligned} & \|\tilde{X}(t, W, \Delta C, \Delta b) - \tilde{X}(t, W^b, \Delta C, \Delta b)\| \\ & = \sup_{\theta \in [0, 1]} \|\tilde{X}'(t, \theta W + (1 - \theta)W^b, \Delta C, \Delta b)\| \|W - W^b\| \\ & \leq M\|W_{BN}\| = \mathcal{O}(\|W_{BN}\|) \end{aligned}$$

for all $(t, W, \Delta C, \Delta b) \in [0, 1] \times \mathcal{W}(\gamma) \times \mathcal{K}$. This estimate and the fact that $\tilde{X}(t, W, \Delta C, \Delta b) = D_N(t)X(t^2, W, \Delta C, \Delta b)D_N(t)$ for all $t \in (0, 1]$ and $(W, \Delta C, \Delta b) \in \mathcal{W} \times \mathcal{K}$ imply

$$\begin{aligned} & X(t^2, W, \Delta C, \Delta b) - X(t^2, W^b, \Delta C, \Delta b) \\ &= \begin{pmatrix} 1 & 0 \\ 0 & t \end{pmatrix} \left(\tilde{X}(t, W, \Delta C, \Delta b) - \tilde{X}(t, W^b, \Delta C, \Delta b) \right) \begin{pmatrix} 1 & 0 \\ 0 & t \end{pmatrix} \\ &= \begin{pmatrix} \mathcal{O}(\|W_{BN}\|) & \mathcal{O}(t\|W_{BN}\|) \\ \mathcal{O}(t\|W_{BN}\|) & \mathcal{O}(t^2\|W_{BN}\|) \end{pmatrix}. \end{aligned}$$

Noting that

$$\begin{aligned} & X(t^2, W, \Delta C, \Delta b) - X(0, W^b, \Delta C, \Delta b) \\ &= (X(t^2, W, \Delta C, \Delta b) - X(t^2, W^b, \Delta C, \Delta b)) \\ &\quad + (X(t^2, W^b, \Delta C, \Delta b) - X(0, W^b, \Delta C, \Delta b)) \end{aligned}$$

and using the above estimate together with Lemma 5.5, we immediately obtain (67) upon letting $\nu = t^2$. The estimate (68) can be proved in a similar way. $\square$

We are now ready to state and prove the main result of this section.

THEOREM 5.7. *For all $(\nu, W, \Delta C, \Delta b) \in [0, 1] \times \mathcal{W}(\gamma) \times \mathcal{K}$, there hold*

$$\begin{aligned} (70) \quad & X(\nu, W, \Delta C, \Delta b) - X(0, W, \Delta C, \Delta b) \\ &= \begin{pmatrix} \mathcal{O}(\sqrt{\nu}\|W_{BN}\|) & \mathcal{O}(\sqrt{\nu}\|W_{BN}\|) \\ \mathcal{O}(\sqrt{\nu}\|W_{BN}\|) & \mathcal{O}(\nu\|W_{BN}\|) \end{pmatrix} + \mathcal{O}(\nu), \end{aligned}$$

$$\begin{aligned} (71) \quad & S(\nu, W, \Delta C, \Delta b) - S(0, W, \Delta C, \Delta b) \\ &= \begin{pmatrix} \mathcal{O}(\nu\|W_{BN}\|) & \mathcal{O}(\sqrt{\nu}\|W_{BN}\|) \\ \mathcal{O}(\sqrt{\nu}\|W_{BN}\|) & \mathcal{O}(\sqrt{\nu}\|W_{BN}\|) \end{pmatrix} + \mathcal{O}(\nu). \end{aligned}$$

Proof. We will prove (70) only since the proof of (71) is similar. Since both $X(0, W, \Delta C, \Delta b)$ and $X(0, W^b, \Delta C, \Delta b)$ are in $\mathcal{F}_P^*$, we have $X_{BN}(0, W, \Delta C, \Delta b) = X_{BN}(0, W^b, \Delta C, \Delta b) = 0$ and $X_N(0, W, \Delta C, \Delta b) = X_N(0, W^b, \Delta C, \Delta b) = 0$ for any $(W, \Delta C, \Delta b) \in \mathcal{W} \times \mathcal{K}$. Hence, in view of Lemma 5.6, it suffices to show that for all $(\nu, W, \Delta C, \Delta b) \in [0, 1] \times \mathcal{W}(\gamma) \times \mathcal{K}$, we have

$$(72) \quad X_B(\nu, W, \Delta C, \Delta b) - X_B(0, W, \Delta C, \Delta b) = \mathcal{O}(\sqrt{\nu}\|W_{BN}\| + \nu).$$

Indeed, using the fact that $\tilde{X}_B(t, W, \Delta C, \Delta b)$ is analytic over the compact set $[0, 1] \times \mathcal{W}(\gamma) \times \mathcal{K}$ due to Theorem 5.2(i), we conclude that

$$\begin{aligned} & X_B(t^2, W, \Delta C, \Delta b) - X_B(0, W, \Delta C, \Delta b) = \tilde{X}_B(t, W, \Delta C, \Delta b) - \tilde{X}_B(0, W, \Delta C, \Delta b) \\ (73) \quad &= \frac{\partial}{\partial t} \tilde{X}_B(0, W, \Delta C, \Delta b)t + \mathcal{O}(t^2) \end{aligned}$$

for every $(t, W, \Delta C, \Delta b) \in [0, 1] \times \mathcal{W}(\gamma) \times \mathcal{K}$. Moreover, by Lemma 4.7, we have

$$\frac{\partial \tilde{X}_B}{\partial t}(0, W^b, \Delta C, \Delta b) = 0$$

for any $(W, \Delta C, \Delta b) \in \mathcal{W} \times \mathcal{K}$, where W^b is as defined in (69). Hence, for every $(W, \Delta C, \Delta b) \in \mathcal{W}(\gamma) \times \mathcal{K}$, we have

$$(74) \quad \frac{\partial}{\partial t} \tilde{X}_B(0, W, \Delta C, \Delta b) = \frac{\partial}{\partial t} \tilde{X}_B(0, W^b, \Delta C, \Delta b) + \mathcal{O}(\|W - W^b\|) = \mathcal{O}(\|W_{BN}\|).$$

Combining (73) and (74), we obtain (72) upon letting $\nu = t^2$. $\square$

The proof of Theorem 5.1 now follows from assumption A.1 and Theorem 5.7 with $\nu = \tau$, $W = X^{1/2}SX^{1/2}/\tau$, $(X, S) = (X(\nu, W, \Delta C, \Delta b), S(\nu, W, \Delta C, \Delta b))$, and the fact that $(X(0, W, \Delta C, \Delta b), S(0, W, \Delta C, \Delta b), y(0, W, \Delta C, \Delta b)) \in \mathcal{F}_P^* \times \mathcal{F}_D^*$.

6. Superlinear convergence criteria. In this section, we consider a sufficient condition introduced by Potra and Sheng [33], which guarantees the superlinear convergence of a class of primal-dual interior-point algorithms for SDP, and we show that it is equivalent to a natural condition about the matrix $W(X, S, \tau)$ of Theorem 5.1.

For the sake of concreteness, we will focus our attention on the algorithm and results obtained in Potra and Sheng (see Algorithm 3.1 in [34]), but we remark that our discussion also applies to a broader class of algorithms. Potra and Sheng [34] have developed a primal-dual infeasible-interior-point algorithm which, for some $\alpha \in (0, 1/2]$, generates a sequence of iterates $\{(X^k, S^k, y^k)\} \subseteq \mathcal{S}_{++}^n \times \mathcal{S}_{++}^n \times \mathfrak{R}^m$ satisfying

$$(75) \quad \|W^k - I\|_F \leq \alpha, \quad r_p^k = \frac{\tau_k}{\tau_0} r_p^0, \quad r_d^k = \frac{\tau_k}{\tau_0} r_d^0$$

for some sequence $\{\tau_k\} \subset \mathfrak{R}_{++}$ converging to 0 at least Q -linearly, where

$$\begin{aligned} r_p^k &\equiv \mathcal{A}X^k - b, \\ r_d^k &\equiv \mathcal{A}^*y^k + S^k - C, \\ W^k &\equiv \frac{(X^k)^{1/2}S^k(X^k)^{1/2}}{\tau_k} \end{aligned}$$

for all $k \geq 0$. The derived linear rate of convergence of the sequence $\{\tau_k\}$ is sufficient to guarantee polynomial convergence of their method under some suitable conditions on the initial point (X^0, S^0, y^0) . Observe that the first condition in (75) implies that $\tau_k = \Theta(X^k \bullet S^k/n)$, and hence asymptotic convergence of $\{X^k \bullet S^k/n\}$ can be derived from the one obtained for $\{\tau_k\}$.

Some sufficient conditions have been developed in the literature which guarantee the Q -superlinear convergence of $\{\tau_k\}$ to zero. One such condition is the tangential condition proposed by Kojima, Shida, and Shindoh [17], namely,

$$(76) \quad \lim_{k \rightarrow \infty} W^k = I.$$

Another such condition, and the one which will be the main subject of this section, is the one that has been proposed by Potra and Sheng [33], namely,

$$(77) \quad \lim_{k \rightarrow \infty} X^k S^k / \sqrt{\tau_k} = 0.$$

We remark that Potra and Sheng [33] have shown that the tangential condition (76) implies their condition (77). Moreover, they have also established the following superlinear convergence result.

PROPOSITION 6.1 (Theorem 6.1 of [33]). *If (77) holds, then the sequence $\{\tau_k\}$ generated by Algorithm 3.1 of [34] converges to zero Q -superlinearly. Moreover, if $X^k S^k = \mathcal{O}(\tau_k^{0.5+\sigma})$ for some constant $\sigma > 0$, then $\{\tau_k\}$ converges to zero with Q -order at least $1 + \min\{\sigma, 0.5\}$.*

A natural relaxation of the tangential condition (76) is the condition that

$$(78) \quad \lim_{k \rightarrow \infty} W_{BN}^k = 0.$$

Surprisingly, the following result shows that it is equivalent to Potra and Sheng's condition (77).

PROPOSITION 6.2. *Let $\theta_k \equiv \|X^k S^k\|/\sqrt{\tau_k}$. Then, $\|W_{BN}^k\| + \sqrt{\tau_k} = \Theta(\theta_k + \sqrt{\tau_k})$.*

Proof. We first show that $\|W_{BN}^k\| + \sqrt{\tau_k} = \mathcal{O}(\theta_k + \sqrt{\tau_k})$. By Lemma 4.2 of [33], we have

$$X^k = \begin{pmatrix} \Theta(1) & \mathcal{O}(\sqrt{\tau_k}) \\ \mathcal{O}(\sqrt{\tau_k}) & \mathcal{O}(\tau_k) \end{pmatrix}, \quad S^k = \begin{pmatrix} \mathcal{O}(\tau_k) & \mathcal{O}(\sqrt{\tau_k}) \\ \mathcal{O}(\sqrt{\tau_k}) & \Theta(1) \end{pmatrix},$$

and hence

$$\begin{aligned} \frac{X^k S^k}{\sqrt{\tau_k}} &= \frac{1}{\sqrt{\tau_k}} \begin{pmatrix} \Theta(1) & \mathcal{O}(\sqrt{\tau_k}) \\ \mathcal{O}(\sqrt{\tau_k}) & \mathcal{O}(\tau_k) \end{pmatrix} \begin{pmatrix} \mathcal{O}(\tau_k) & \mathcal{O}(\sqrt{\tau_k}) \\ \mathcal{O}(\sqrt{\tau_k}) & \Theta(1) \end{pmatrix} \\ &= \begin{pmatrix} \mathcal{O}(\sqrt{\tau_k}) & \mathcal{O}(1) \\ \mathcal{O}(\tau_k) & \mathcal{O}(\sqrt{\tau_k}) \end{pmatrix}. \end{aligned}$$

According to the definition of θ_k , we then conclude that

$$(79) \quad \frac{X^k S^k}{\sqrt{\tau_k}} = \begin{pmatrix} \mathcal{O}(\sqrt{\tau_k}) & \mathcal{O}(\theta_k) \\ \mathcal{O}(\tau_k) & \mathcal{O}(\sqrt{\tau_k}) \end{pmatrix}.$$

By Lemma 4.5 of [33], we have

$$(80) \quad (X^k)^{1/2} = \begin{pmatrix} \mathcal{O}(1) & \mathcal{O}(\sqrt{\tau_k}) \\ \mathcal{O}(\sqrt{\tau_k}) & \mathcal{O}(\sqrt{\tau_k}) \end{pmatrix}, \quad (X^k)^{-1/2} = \begin{pmatrix} \mathcal{O}(1) & \mathcal{O}(1) \\ \mathcal{O}(1) & \mathcal{O}(1/\sqrt{\tau_k}) \end{pmatrix},$$

which together with (79) imply

$$\begin{aligned} W^k &= \frac{(X^k)^{1/2} S^k (X^k)^{1/2}}{\tau_k} = \frac{1}{\sqrt{\tau_k}} (X^k)^{-1/2} \left(\frac{X^k S^k}{\sqrt{\tau_k}} \right) (X^k)^{1/2} \\ &= \frac{1}{\sqrt{\tau_k}} \begin{pmatrix} \mathcal{O}(1) & \mathcal{O}(1) \\ \mathcal{O}(1) & \mathcal{O}(1/\sqrt{\tau_k}) \end{pmatrix} \begin{pmatrix} \mathcal{O}(\sqrt{\tau_k}) & \mathcal{O}(\theta_k) \\ \mathcal{O}(\tau_k) & \mathcal{O}(\sqrt{\tau_k}) \end{pmatrix} \begin{pmatrix} \mathcal{O}(1) & \mathcal{O}(\sqrt{\tau_k}) \\ \mathcal{O}(\sqrt{\tau_k}) & \mathcal{O}(\sqrt{\tau_k}) \end{pmatrix}. \end{aligned}$$

Since the (B, N) -block of the matrix in the right-hand side of the above identity is $\mathcal{O}(\theta_k + \sqrt{\tau_k})$, we conclude that $\|W_{BN}^k\| + \sqrt{\tau_k} = \mathcal{O}(\theta_k + \sqrt{\tau_k})$.

Next we show that $\theta_k + \sqrt{\tau_k} = \mathcal{O}(\|W_{BN}^k\| + \sqrt{\tau_k})$. By the first condition in (75), we have

$$W^k = \begin{pmatrix} \mathcal{O}(1) & \mathcal{O}(\|W_{BN}^k\|) \\ \mathcal{O}(\|W_{BN}^k\|) & \mathcal{O}(1) \end{pmatrix},$$

which together with (80) implies that

$$\begin{aligned} \frac{X^k S^k}{\sqrt{\tau_k}} &= \sqrt{\tau_k} (X^k)^{1/2} \left(\frac{(X^k)^{1/2} S^k (X^k)^{1/2}}{\tau_k} \right) (X^k)^{-1/2} = \sqrt{\tau_k} (X^k)^{1/2} W^k (X^k)^{-1/2} \\ &= \sqrt{\tau_k} \begin{pmatrix} \mathcal{O}(1) & \mathcal{O}(\sqrt{\tau_k}) \\ \mathcal{O}(\sqrt{\tau_k}) & \mathcal{O}(\sqrt{\tau_k}) \end{pmatrix} \begin{pmatrix} \mathcal{O}(1) & \mathcal{O}(\|W_{BN}^k\|) \\ \mathcal{O}(\|W_{BN}^k\|) & \mathcal{O}(1) \end{pmatrix} \begin{pmatrix} \mathcal{O}(1) & \mathcal{O}(1) \\ \mathcal{O}(1) & \mathcal{O}(1/\sqrt{\tau_k}) \end{pmatrix} \\ &= \begin{pmatrix} \mathcal{O}(\sqrt{\tau_k} + \sqrt{\tau_k} \|W_{BN}^k\|) & \mathcal{O}(\sqrt{\tau_k} + \|W_{BN}^k\|) \\ \mathcal{O}(\tau_k + \tau_k \|W_{BN}^k\|) & \mathcal{O}(\sqrt{\tau_k} + \sqrt{\tau_k} \|W_{BN}^k\|) \end{pmatrix}. \end{aligned}$$

This together with the definition of θ_k implies that $\theta_k + \sqrt{\tau_k} = \mathcal{O}(W_{BN}^k + \sqrt{\tau_k})$. $\square$

In view of the equivalence between (77) and (78), it follows that a sufficient condition for the superlinear convergence of the path-following algorithm outlined in this section is that the sequence of matrices $\{W^k\}$ approaches the set of block diagonal matrices. Clearly, this is a much weaker condition than (76), which requires this sequence to approach the identity matrix.

7. Concluding remarks. In this section we provide some final remarks related to the results derived in this paper.

Under the assumptions of this paper, we have shown that the reparametrized $(W, \Delta C, \Delta b)$ -weighted central path $(X(t^2), S(t^2), y(t^2))$ is analytic at $t = 0$ and that the condition $W_{BN} = 0$ implies that $\lim_{\nu \downarrow 0} (\dot{X}(\nu), \dot{S}(\nu), \dot{y}(\nu))$ exists. Based on the latter conclusion, it is natural to wonder whether the path $(X(\nu), S(\nu), y(\nu))$ is analytic at $\nu = 0$ when W is block diagonal. Note that the answer to this question is affirmative when $(W, \Delta C, \Delta b) = (I, 0, 0)$; i.e., the weighted central path is exactly the central path (see Halická [12]).

In this paper, we have proved that the rate of convergence of the $(W, \Delta C, \Delta b)$ -weighted central path $(X(\nu), S(\nu), y(\nu))$ toward the optimal solution set is $\mathcal{O}(\sqrt{\nu})$ (and $\mathcal{O}(\nu)$ when $W_{BN} = 0$). In contrast, Preiß and Stoer [35] have shown that the rate of convergence of the weighted central paths associated with the map $(XS + SX)/2$ is always $\mathcal{O}(\nu)$ (see also Lu and Monteiro [20]). An error bound of this type has also been shown by Kojima, Shida, and Shindoh [18], where it is shown that an interior-point algorithm based on a centering condition associated with the $(XS + SX)/2$ map does not need to approach the central path tangentially in order to converge superlinearly. On the other hand, the iterates of all superlinearly convergent interior-point algorithms based on centering conditions associated with the map $X^{1/2}SX^{1/2}$ that have been proposed in the literature are required to approach the central path tangentially. The latter requirement is natural in view of the fact that it forces $(X^k)^{1/2}S^k(X^k)^{1/2}$ to approach a block diagonal matrix (the identity matrix), and hence it reduces the bound on the distance of (X^k, S^k, y^k) to the optimal solution set from the usual $\mathcal{O}(\sqrt{\mu_k})$ to $o(\sqrt{\mu_k})$ (see Theorem 5.1).

Acknowledgments. The authors would like to thank the two anonymous referees and the associate editor Yin Zhang for their constructive comments which have helped to substantially improve the presentation of the paper.

REFERENCES

- [1] I. ADLER AND R. D. C. MONTEIRO, *Limiting behavior of the affine scaling continuous trajectories for linear programming problems*, Math. Programming, 50 (1991), pp. 29–51.
- [2] M. D. ASIC, V. V. KOVACEVIC-VUJICIC, AND M. D. RADOSAVLJEVIC-NIKOLIC, *A note on limiting behavior of the projective and the affine rescaling algorithms*, in Mathematical Developments Arising from Linear Programming: Proceedings of a Joint Summer Research Conference, Bowdoin College, Brunswick, ME, 1988, Contemp. Math., AMS, 114, J. C. Lagarias and M. J. Todd, eds., Providence, RI, 1990, pp. 151–157.
- [3] D. A. BAYER AND J. C. LAGARIAS, *The nonlinear geometry of linear programming, I: Affine and projective scaling trajectories*, Trans. Amer. Math. Soc., 314 (1989), pp. 499–526.
- [4] J. DA CRUZ NETO, O. FERREIRA, AND R. D. C. MONTEIRO, *Asymptotic Behavior of the Central Path for a Special Class of Degenerate SDP Problems*, manuscript, School of ISyE, Georgia Institute of Technology, Atlanta, GA, 2003.
- [5] E. DE KLERK, C. ROOS, AND T. TERLAKY, *Initialization in semidefinite programming via a self-dual, skew-symmetric embedding*, Oper. Res. Lett., 20 (1997), pp. 213–221.

- [6] E. DE KLERK, C. ROOS, AND T. TERLAKY, *Infeasible-start semidefinite programming algorithms via self-dual embeddings*, in Topics in Semidefinite and Interior-Point Methods, Fields Inst. Commun. 18, P. M. Pardalos and H. Wolkowicz, eds., AMS, Providence, RI, 1998, pp. 215–236.
- [7] D. GOLDFARB AND K. SCHEINBERG, *Interior point trajectories in semidefinite programming*, SIAM J. Optim., 8 (1998), pp. 871–886.
- [8] L. M. G. DRUMMOND AND H. Y. PETERZIL, *The central path in smooth convex semidefinite programs*, Optimization, 51 (2002), pp. 207–233.
- [9] O. GÜLER, *Limiting behavior of the weighted central paths in linear programming*, Math. Programming, 65 (1994), pp. 347–363.
- [10] M. HALICKÁ, *Analytical properties of the central path at the boundary point in linear programming*, Math. Program., 84 (1999), pp. 335–355.
- [11] M. HALICKÁ, *Two simple proofs of analyticity of the central path in linear programming*, Oper. Res. Lett., 28 (2001), pp. 9–19.
- [12] M. HALICKÁ, *Analyticity of the central path at the boundary point in semidefinite programming*, European J. Oper. Res., 143 (2002), pp. 311–324.
- [13] M. HALICKÁ, E. DE KLERK, AND C. ROOS, *Limiting Behavior of the Central Path in Semidefinite Optimization*, preprint, Faculty of Technical Mathematics and Informatics, TU Delft, Delft, The Netherlands, 2002.
- [14] M. HALICKÁ, E. DE KLERK, AND C. ROOS, *On the convergence of the central path in semidefinite optimization*, SIAM J. Optim., 12 (2002), pp. 1090–1099.
- [15] M. KOJIMA, N. MEGIDDO, T. NOMA, AND A. YOSHISE, *A Unified Approach to Interior Point Algorithms for Linear Complementarity Problems*, Lecture Notes in Comput. Sci. 538, Springer-Verlag, Berlin, 1991.
- [16] M. KOJIMA, S. MIZUNO, AND T. NOMA, *Limiting behavior of trajectories by a continuation method for monotone complementarity problems*, Math. Oper. Res., 15 (1990), pp. 662–675.
- [17] M. KOJIMA, M. SHIDA, AND S. SHINDOH, *Local convergence of predictor-corrector infeasible-interior-point algorithms for SDPs and SDLCPs*, Math. Programming, 80 (1998), pp. 129–160.
- [18] M. KOJIMA, M. SHIDA, AND S. SHINDOH, *A predictor-corrector interior-point algorithm for the semidefinite linear complementarity problem using the Alizadeh-Haeberly-Overton search direction*, SIAM J. Optim., 9 (1999), pp. 444–465.
- [19] M. KOJIMA, S. SHINDOH, AND S. HARA, *Interior-point methods for the monotone semidefinite linear complementarity problem in symmetric matrices*, SIAM J. Optim., 7 (1997), pp. 86–125.
- [20] Z. LU AND R. D. C. MONTEIRO, *Limiting Behavior of the Alizadeh-Haeberly-Overton Weighted Paths in Semidefinite Programming*, manuscript, School of ISyE, Georgia Institute of Technology, Atlanta, GA, 2003.
- [21] Z.-Q. LUO, J. F. STURM, AND S. ZHANG, *Superlinear convergence of a symmetric primal-dual path following algorithm for semidefinite programming*, SIAM J. Optim., 8 (1998), pp. 59–81.
- [22] L. McLINDEN, *An analogue of Moreau’s approximation theorem, with application to the non-linear complementarity problem*, Pacific J. Math., 88 (1980), pp. 101–161.
- [23] L. McLINDEN, *The complementarity problem for maximal monotone multifunctions*, in Variational Inequalities and Complementarity Problems, R. Cottle, F. Giannessi, and J.-L. Lions, eds., Wiley, New York, 1980, pp. 251–270.
- [24] N. MEGIDDO, *Pathways to the optimal set in linear programming*, in Progress in Mathematical Programming: Interior Point and Related Methods, N. Megiddo, ed., Springer-Verlag, New York, 1989, pp. 131–158.
- [25] J. MILNOR, *Singular Points of Complex Hypersurfaces*, Ann. of Math. Stud. 61, Princeton University Press, Princeton, NJ, 1968.
- [26] R. D. C. MONTEIRO, *Convergence and boundary behavior of the projective scaling trajectories for linear programming*, Math. Oper. Res., 16 (1991), pp. 842–858.
- [27] R. D. C. MONTEIRO AND J.-S. PANG, *Properties of an interior-point mapping for mixed complementarity problems*, Math. Oper. Res., 21 (1996), pp. 629–654.
- [28] R. D. C. MONTEIRO AND J.-S. PANG, *On two interior-point mappings for nonlinear semidefinite complementarity problems*, Math. Oper. Res., 23 (1998), pp. 39–60.
- [29] R. D. C. MONTEIRO AND T. TSUCHIYA, *Limiting behavior of the derivatives of certain trajectories associated with a monotone horizontal linear complementarity problem*, Math. Oper. Res., 21 (1996), pp. 793–814.

- [30] R. D. C. MONTEIRO AND T. TSUCHIYA, *Polynomial convergence of a new family of primal-dual algorithms for semidefinite programming*, SIAM J. Optim., 9 (1999), pp. 551–577.
- [31] R. D. C. MONTEIRO AND P. ZANJÁCOMO, *General interior-point maps and existence of weighted paths for nonlinear semidefinite complementarity problems*, Math. Oper. Res., 25 (2000), pp. 381–399.
- [32] R. D. C. MONTEIRO AND F. ZHOU, *On the existence and convergence of the central path for convex programming and some duality results*, Comput. Optim. Appl., 10 (1998), pp. 51–77.
- [33] F. A. POTRA AND R. SHENG, *Superlinear Convergence of Interior-Point Algorithms for Semidefinite Programming*, Reports on Computational Mathematics 86, Department of Mathematics, The University of Iowa, Iowa City, IA, 1996.
- [34] F. A. POTRA AND R. SHENG, *A superlinearly convergent primal-dual infeasible-interior-point algorithm for semidefinite programming*, SIAM J. Optim., 8 (1998), pp. 1007–1028.
- [35] M. PREIB AND J. STOER, *Analysis of infeasible-interior-point paths arising with semidefinite linear complementarity problems*, Math. Program., 99 (2003), pp. 499–520.
- [36] G. SPORRE AND A. FORSGREN, *Characterization of the Limit Point of the Central Path in Semidefinite Programming*, Technical report TRITA-MAT-2002-OS12, Department of Mathematics, Royal Institute of Technology, Stockholm, Sweden, 2002.
- [37] J. STOER AND M. WECHS, *Infeasible-interior-point paths for sufficient linear complementarity problems*, Math. Programming, 83 (1998), pp. 403–423.
- [38] J. STOER AND M. WECHS, *On the analyticity properties of infeasible-interior point paths for monotone linear complementarity problems*, Numer. Math., 81 (1999), pp. 631–645.
- [39] M. WECHS, *The analyticity of interior-point-paths at strictly complementary solutions of linear programs*, Optim. Methods Softw., 9 (1998), pp. 209–243.
- [40] C. WITZGALL, P. T. BOGGS, AND P. D. DOMICH, *On the convergence behavior of trajectories for linear programming*, in Mathematical Developments Arising from Linear Programming: Proceedings of a Joint Summer Research Conference, Bowdoin College, Brunswick, ME, 1988, Contemp. Math. 114, J. C. Lagarias and M. J. Todd, eds., AMS, Providence, RI, 1990, pp. 161–187.

KNOWLEDGE-BASED LINEAR PROGRAMMING*

O. L. MANGASARIAN†

Abstract. We introduce a class of linear programs with constraints in the form of implications. Such linear programs arise in support vector machine classification, where in addition to explicit datasets to be classified, prior knowledge such as expert experience in the form of logical implications are imposed on the classifier. The overall problem can be viewed either as a semi-infinite linear program or as a linear program with equilibrium constraints which, in either case, can be solved by an equivalent simple linear program under mild assumptions.

Key words. linear programming, prior knowledge, implication constraints

AMS subject classifications. 90C05, 90C33, 90C34

DOI. 10.1137/S1052623403432839

1. Introduction. Recently, support vector machines (SVMs) [17, 3, 18, 16, 4, 15], which constitute the method of choice for classification and other approximation problems, were applied to problems where prior expert knowledge, such as doctors' experience, was incorporated into the classification problem in order to improve the correctness of the classifier generated [6, 5]. This resulted in a linear program where constraints corresponding to prior knowledge were made up of logical implications. In this work, we generalize this type of constraint and call the corresponding linear program a *knowledge-based linear program* (KBLP). It is interesting to note that such knowledge-based logical implications are equivalent to an infinite number of constraints, thus resulting in a semi-infinite linear program [7], or alternatively they can be represented as constraints in a linear program with equilibrium constraints [8]. However, we will show in section 3 that such logical implications can be replaced by a finite number of inequality constraints, thus leading to a simple linear programming formulation. In section 2 of the paper we state our general KBLP and motivate it by a knowledge-based linear SVM classifier. In section 4 we establish the equivalence of KBLP to a single linear program, give some duality results, and establish the equivalence of KBLP to a linear program with equilibrium constraints. Section 5 concludes the paper with a summary and an outlook.

A word about our notation: All vectors will be column vectors unless transposed to a row vector by a prime superscript $'$. The scalar product of two vectors x and y in the n -dimensional real space R^n will be denoted by $x'y$. For $x \in R^n$ and $p \in [1, \infty)$, the norm $\|x\|_p$ will denote the p -norm $(\sum_{i=1}^n |x_i|^p)^{\frac{1}{p}}$ and $\|x\|_\infty$ will denote $\max_{1 \leq i \leq n} |x_i|$. For $x \in R^n$, $(x_+)_i = \max \{0, x_i\}$, $i = 1, \dots, n$, while the sign function $sign(x)$ is defined as $sign(x)_i = 1$ if $x_i > 0$ else $sign(x)_i = -1$ if $x_i \leq 0$ for $i = 1, \dots, n$. For $x \in R^n$ and $y \in R^n$, $x \perp y$ denotes orthogonality, that is, $x'y = 0$. For an $m \times n$ matrix A , A_i will denote the i th row of A , A_j will denote the j th column of A , and A_{ij} will denote the element in row i and column j . The identity matrix in a real space of arbitrary dimension will be denoted by I , while a column vector of ones of

*Received by the editors August 5, 2003; accepted for publication (in revised form) March 11, 2004; published electronically December 30, 2004. Data Mining Institute Technical Report 03-04, July 2003. Revised March 2004. This research was supported by NSF grant CR-0138308 and by Microsoft.

<http://www.siam.org/journals/siopt/15-2/43283.html>

†Computer Sciences Department, University of Wisconsin, Madison, WI 53706, and Department of Mathematics, University of California at San Diego, La Jolla, CA 92093 (olvi@cs.wisc.edu).

arbitrary dimension will be denoted by e .

2. The knowledge-based linear program. We consider the following linear program with a logical implication constraint:

$$(1) \quad \begin{array}{ll} \min_x & c'x \\ \text{s.t.} & Hx \geq h, \\ & Ls \leq d \implies (Ex - a)'s \geq \mu. \end{array}$$

Here, $H \in R^{m \times n}$, $L \in R^{\ell \times k}$, and $E \in R^{k \times n}$ are given matrices; $c \in R^n$, $h \in R^m$, $d \in R^\ell$, and $a \in R^k$ are given vectors; and μ is a given real number. An alternate way to write this problem is the following semi-infinite program:

$$(2) \quad \begin{array}{ll} \min_x & c'x \\ \text{s.t.} & Hx \geq h, \\ & (Ex - a)'s \geq \mu \quad \forall s \in \{s | Ls \leq d\}. \end{array}$$

We now describe knowledge-based classification problems [6] that lead to the KBLP (1)–(2). These problems, in addition to classifying conventional datasets consisting of a finite number of points in R^n , attempt to classify polyhedral sets in R^n which are known to belong to one of two classes. We thus consider the problem of classifying m points in the n -dimensional input space R^n , represented by the $m \times n$ matrix A , according to membership of each point A_i in the class $A+$ or $A-$ as specified by a given $m \times m$ diagonal matrix D with plus ones or minus ones along its diagonal. For this problem, the linear programming SVM [11, 1] with a linear kernel, which is a variant of the standard SVM [18, 2], is given by the following linear program with parameter $\nu > 0$:

$$(3) \quad \begin{array}{ll} \min_{(w, \gamma, y, t) \in R^{n+1+m}} & \nu e'y + e't \\ \text{s.t.} & D(Aw - e\gamma) + y \geq e, \\ & t \geq w \geq -t, \\ & y \geq 0. \end{array}$$

Here, w is the normal to the bounding planes

$$(4) \quad \begin{array}{l} x'w = \gamma + 1, \\ x'w = \gamma - 1 \end{array}$$

that bound most of the points belonging to the sets $A+$ and $A-$, respectively. The constant γ determines their location relative to the origin. When the two classes are strictly linearly separable, that is, when the error variable $y = 0$ in (3), the plane $x'w = \gamma + 1$ bounds all of the class $A+$ points, while the plane $x'w = \gamma - 1$ bounds all of the class $A-$ points as follows:

$$(5) \quad \begin{array}{ll} A_i w \geq \gamma + 1 & \text{for } D_{ii} = 1, \\ A_i w \leq \gamma - 1 & \text{for } D_{ii} = -1. \end{array}$$

Consequently, the plane

$$(6) \quad x'w = \gamma,$$

midway between the bounding planes in (4), is a separating plane that separates points belonging to $A+$ from those belonging to $A-$ completely if $y = 0$, else only approximately. The 1-norm term $\|w\|_1$ in (3), that is the term $e't$ at a solution point, which is half the reciprocal of the distance $\frac{2}{\|w\|_1}$ measured using the ∞ -norm distance [10] between the two bounding planes of (4), maximizes this distance, often called the *margin*. Maximizing the margin enhances the generalization capability of a support vector machine [18, 2]. If the two classes are linearly inseparable, then the two planes bound the two classes with a *soft margin* (i.e., bound approximately with some error) determined by the nonnegative error variable y , that is,

$$(7) \quad \begin{aligned} A_i w + y_i &\geq \gamma + 1 & \text{for } D_{ii} = 1, \\ A_i w - y_i &\leq \gamma - 1 & \text{for } D_{ii} = -1. \end{aligned}$$

The 1-norm of the error variable y is minimized parametrically with weight ν in (3), resulting in an approximate separating plane. This plane acts as a linear classifier as follows:

$$(8) \quad \text{sign}(x'w - \gamma) \begin{cases} = 1, & \text{then } x \in A+, \\ = -1, & \text{then } x \in A-, \end{cases}$$

where $\text{sign}(\cdot)$ is the sign function defined in the introduction. Suppose now that we have prior information of the following type. All points $x \in R^n$ lying in the polyhedral set determined by the linear inequalities

$$(9) \quad Bx \leq b,$$

where $B \in R^{k \times n}$ and $b \in R^k$, belong to class $A+$. Such inequalities generalize simple box constraints such as $a \leq x \leq d$. Looking at the inequalities (9), we conclude that the following implication must hold:

$$(10) \quad Bx \leq b \implies x'w \geq \gamma + 1.$$

That is, the *knowledge set* $\{x \mid Bx \leq b\}$ lies on the $A+$ side of the bounding plane $x'w = \gamma + 1$. This leads to the following knowledge-based support vector machine classification problem:

$$(11) \quad \begin{aligned} \min_{(w, \gamma, y, t) \in R^{n+1+m}} \quad & \nu e'y + e't \\ \text{s.t.} \quad & D(Aw - e\gamma) + y \geq e, \\ & t \geq w \geq -t, \\ & y \geq 0, \\ & Bs \leq b \implies s'w \geq \gamma + 1, \end{aligned}$$

which is a special case of the KBLP (1) upon making the appropriate identifications for $x, c, H, h, L, d, E, a, \mu$ in (1) in terms of $w, \gamma, y, t, \nu, D, A, B, b$ of (11).

Another consequence of minimizing the 1-norm of w in the linear programming formulation (3) is feature suppression. For example, in [1] as many as 86% of input space features were eliminated from the problem by a linear kernel SVM formulation such as (3). This raises the possibility of a similar dimensionality reduction for KBLP (1) by adding a perturbation term $\sigma\|x\|_1$ to the objective function of (1) and making use of the perturbation theory result of [13] to show that such a perturbation will

lead to a least 1-norm solution of (1) for $\sigma \in (0, \bar{\sigma}]$ for some $\bar{\sigma} > 0$. Based on SVM experience, this will often lead to substantial dimensionality reduction.

We further note that the implication constraint employed in the linear programming formulation of (1) can also be added to the constraints of any mathematical programming problem and converted to a set of ordinary linear constraints by the techniques described in section 3. In particular, for example, knowledge-based implication constraints can be added to the standard quadratic programming formulation of SVMs [17, 3, 18]. However, as shown in [1], the standard quadratic programming SVM formulation, which also minimizes $\|w\|_2^2$, induces no feature suppression and takes a longer time to solve than a 1-norm SVM such as (3) that minimizes $\|w\|_1$. For this reason and for the sake of simplicity, we confine ourselves here to knowledge-based *linear programs*.

We show now that the implication constraint (10) is equivalent to a finite number of inequality constraints and hence KBLP (1) is equivalent to a simple linear program.

3. Equivalence of KBLP to a linear program with a finite number of linear inequalities. The equivalence result can be obtained as a direct consequence of the set containment characterizations on [12, Proposition 2.1] or more directly by Motzkin's theorem of the alternative [9, Theorem 2.4.2], which we utilize here.

PROPOSITION 3.1 (implication constraint as linear inequalities). *Let the set $\{s \mid Ls \leq d\}$ be nonempty. Then, for a fixed $x \in R^n$, the implication constraint*

$$(12) \quad Ls \leq d \implies (Ex - a)'s \geq \mu$$

is equivalent to the following system of linear inequalities having a solution $u \in R^\ell$:

$$(13) \quad \begin{aligned} L'u + Ex &= a, \\ -d'u &\geq \mu, \\ u &\geq 0. \end{aligned}$$

Proof. For a fixed $x \in R^n$, the implication (12) is equivalent to the following system having *no* solution $(s, \zeta) \in R^{k+1}$:

$$(14) \quad Ls - d\zeta \leq 0, \quad (Ex - a)'s - \mu\zeta < 0, \quad -\zeta < 0.$$

By Motzkin's theorem [9, Theorem 2.4.2], this is equivalent to the following system *having* a solution $(u, \xi, \eta) \in R^{\ell+2}$:

$$(15) \quad L'u + (Ex - a)\xi = 0, \quad -d'u - \mu\xi - \eta = 0, \quad (u, \xi, \eta) \geq 0, \quad (\xi, \eta) \neq 0.$$

If $\xi = 0$, then $\eta > 0$ and the system (15) having such a solution contradicts the nonemptiness assumption on $\{s \mid Ls \leq d\}$, because for $s \in \{s \mid Ls \leq d\}$, we obtain the contradiction

$$(16) \quad 0 = s'L'u \leq u'd = -\eta < 0.$$

Hence the system (15) must have a solution with $\xi > 0$. Dividing throughout by ξ and redefining $\frac{u}{\xi} \rightarrow u$, $\frac{\eta}{\xi} \rightarrow \eta$, the system (15) becomes

$$(17) \quad L'u + (Ex - a) = 0, \quad -d'u - \mu = \eta \geq 0, \quad u \geq 0,$$

which is equivalent to (13). $\square$

With the above proposition, the following equivalence becomes evident.

COROLLARY 3.2 (equivalent KBLP). *Under the assumption that the set $\{s \mid Ls \leq d\}$ is nonempty, the KBLP (1) is equivalent to the following linear program with a finite number of linear inequalities:*

$$(18) \quad \begin{array}{ll} \min_{x,u} & c'x \\ \text{s.t.} & Hx \geq h, \\ & Ex + L'u = a, \\ & -d'u \geq \mu, \\ & u \geq 0. \end{array}$$

We turn now to establishing the existence of a solution to KBLP (1) through the above equivalence and to some duality results.

4. Existence of a solution and duality for KBLP. We will first establish that KBLP (1) has a solution provided that it is feasible and that the underlying linear program without the knowledge constraint is solvable.

PROPOSITION 4.1 (existence of a solution to KBLP). *Let KBLP (1) be feasible, that is,*

$$(19) \quad X = \{x \mid Hx \geq h, \quad (Ex - a)'s \geq \mu \quad \forall s : Ls \leq d\} \neq \emptyset,$$

and if the underlying linear program

$$(20) \quad \begin{array}{ll} \min_x & c'x \\ \text{s.t.} & Hx \geq h, \end{array}$$

is solvable, then KBLP (1) is solvable.

Proof. The dual of the underlying linear program (20) is

$$(21) \quad \begin{array}{ll} \max_t & h't \\ \text{s.t.} & H't = c, \\ & t \geq 0, \end{array}$$

while the dual of the equivalent KBLP (18) is

$$(22) \quad \begin{array}{ll} \max_{v,r,\zeta} & h'v + a'r + \mu\zeta \\ \text{s.t.} & H'v + E'r = c, \\ & Lr - d\zeta \leq 0, \\ & (v, \zeta) \geq 0. \end{array}$$

The assumption that KBLP (1) is feasible, that is, $X \neq \emptyset$, is equivalent to the fact that the equivalent KBLP (18) is feasible. Since the underlying linear program (20) is solvable, its dual (21) is feasible. It follows that the dual (22) of the equivalent KBLP (18) is also feasible by taking $v = t$, $r = 0$, $\zeta = 0$. Hence the equivalent KBLP (18) is solvable because it is feasible and its dual (22) is feasible. Hence KBLP (1) is solvable. $\square$

Having established the existence of a solution to KBLP (1), we pose the question: When is the dual (22), to the equivalent KBLP (18), also a KBLP? That is, can the

constraints of (22) be reduced to logical implication constraints and linear inequalities? In its present general form, the dual problem does not seem to be amenable to such a reduction. However, if we take a special case of $h = 0$, $d = 0$, $\mu = 0$, then indeed both the primal KBLP (1) and its dual (22) are KBLPs as follows.

PROPOSITION 4.2 (dual KBLPs). *Let*

$$(23) \quad Y = \{x \mid Hx \geq 0, \quad (Ex - a)'s \geq 0 \quad \forall s : Ls \leq 0\} \neq \emptyset.$$

Then the following problems are dual KBLPs for which standard linear programming duality results hold:

$$(24) \quad \begin{array}{ll} \min_x & c'x \\ \text{s.t.} & Hx \geq 0, \\ & Ls \leq 0 \implies (Ex - a)'s \geq 0; \end{array}$$

$$(25) \quad \begin{array}{ll} \max_r & a'r \\ \text{s.t.} & Lr \leq 0, \\ & Hy \leq 0 \implies (-E'r + c)'y \leq 0. \end{array}$$

Proof. Since (24) is a special case of KBLP (1), its dual is given as a special case of (22), the dual of KBLP (1), as

$$(26) \quad \begin{array}{ll} \max_{v,r,\zeta} & a'r \\ \text{s.t.} & H'v + E'r = c, \\ & Lr \leq 0, \\ & v \geq 0. \end{array}$$

For a fixed $r \in R^k$, it follows, by the Farkas theorem [9, Theorem 2.4.6], that the first and last constraints of (26) are equivalent to the following system *not* having a solution $y \in R^n$:

$$(27) \quad Hy \leq 0, \quad (c - E'r)'y > 0,$$

which in turn is equivalent to the implication

$$(28) \quad Hy \leq 0 \implies (-E'r + c)'y \leq 0.$$

Hence if we replace the first and last constraints of (26) by the equivalent implication of (28), we obtain the desired KBLP dual (25). $\square$

One final interesting result is a connection to linear programs with equilibrium constraints (LPEQs) [8]. We will show that our KBLP (1) is equivalent to an LPEQ that can be solved as a single linear program (18) under the simple assumption that the set $\{s \mid Ls \leq d\}$ is nonempty. We first write the KBLP (1) in the equivalent form

$$(29) \quad \min_x \{c'x \mid Hx \geq h, \quad \min_s \{(Ex - a)'s \mid Ls \leq d\} \geq \mu\}$$

and note that the dual to the inner linear program in (29) is given by

$$(30) \quad \max_u -d'u \quad \text{s.t.} \quad -L'u = Ex - a, \quad u \geq 0.$$

Hence, the formulation (29) can be rewritten, by using the KKT conditions for the inner linear program of (29), as the following *equivalent* LPEQ:

$$(31) \quad \min_{x,s,u} \{c'x \mid Hx \geq h, \quad Ex + L'u = a, \quad d'u + \mu \leq 0, \quad 0 \leq u \perp (-Ls + d) \geq 0\}.$$

However, we have established in Corollary 3.2 that under the assumption that $\{s \mid Ls \leq d\} \neq \emptyset$, the KBLP (1), and hence its equivalent LPEQ formulation (31), are both equivalent to the linear program (18). That is,

$$(32) \quad \min_{x,u} \{c'x \mid Hx \geq h, \quad Ex + L'u = a, \quad d'u + \mu \leq 0, \quad u \geq 0\},$$

provided $\{s \mid Ls \leq d\} \neq \emptyset$,

which is precisely the LPEQ (31) with the orthogonality condition $u \perp (-Ls + d) \geq 0$ removed. So the LPEQ (31) can be solved as the single linear program (32). A possible intuitive justification for this is that primal and dual feasibility for the inner linear program of (29) are sufficient, without the orthogonality condition $u \perp (-Lu + d) \geq 0$, to induce the condition $(Ex - a)'s \geq -d'u \geq \mu$ for any $s \in R^k$ such that $Ls \leq d$, which is the implication constraint of KBLP (1).

5. Conclusion. We have established the equivalence of linear programs with logical implication constraints to conventional linear programs with a finite number of inequalities. We refer to such linear programs as *knowledge-based linear programs* because they arise naturally in classification problems where these implication constraints are generated by prior knowledge. Since KBLPs can also be written as semi-infinite linear programs or as LPEQs, it follows from our equivalence result that such semi-infinite linear programs and LPEQs can be solved as single linear programs with a finite number of constraints. KBLPs can be generalized in a number of ways including the handling implication constraints involving nonlinear terms in the primal variable x . One such generalization using nonlinear kernels has been proposed in [5]. Another interesting application is to approximation problems where prior knowledge can be used to obtain better approximations. This avenue of research is currently being investigated [14]. It would be interesting to find other real-world applications that can be modeled as knowledge-based linear or convex programs.

REFERENCES

- [1] P. S. BRADLEY AND O. L. MANGASARIAN, *Feature selection via concave minimization and support vector machines*, in Proceedings of the Fifteenth International Conference Machine Learning, J. Shavlik, ed., Morgan Kaufmann, San Francisco, CA, 1998, pp. 82–90; also available from <ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-03.ps>.
- [2] V. CHERKASSKY AND F. MULIER, *Learning from Data—Concepts, Theory, and Methods*, John Wiley, New York, 1998.
- [3] N. CRISTIANINI AND J. SHAWE-TAYLOR, *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, UK, 2000.
- [4] T. EVGENIOU, M. PONTIL, AND T. POGGIO, *Regularization networks and support vector machines*, in Advances in Large Margin Classifiers, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, eds., MIT Press, Cambridge, MA, 2000, pp. 171–203.
- [5] G. FUNG, O. L. MANGASARIAN, AND J. SHAVLIK, *Knowledge-based nonlinear kernel classifiers*, Technical report 03-02, Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, WI, 2003; Conference on Learning Theory and Workshop on Kernel Machines, Washington, DC, 2003, M. Warmuth and B. Schölkopf, eds., Springer-Verlag, Berlin, Germany, 2003, pp. 102–113; also available from <ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/02-03.ps>.

- [6] G. FUNG, O. L. MANGASARIAN, AND J. SHAVLIK, *Knowledge-based support vector machine classifiers*, in Adv. Neural Inf. Process. Syst. 15, S. Thrun, S. Becker, and K. Obermayer, eds., MIT Press, Cambridge, MA, 2003, pp. 521–528; also available from <ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/01-09.ps>.
- [7] M. A. GOBERNA AND M. A. LÓPEZ, *Linear Semi-Infinite Optimization*, John Wiley, New York, 1998.
- [8] Z.-Q. LUO, J.-S. PANG, AND D. RALPH, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, Cambridge, UK, 1996.
- [9] O. L. MANGASARIAN, *Nonlinear Programming*, Classics Appl. Math. 10, SIAM, Philadelphia, PA, 1994.
- [10] O. L. MANGASARIAN, *Arbitrary-norm separating plane*, Oper. Res. Lett., 24 (1999), pp. 15–23; also available from <ftp://ftp.cs.wisc.edu/math-prog/tech-reports/97-07r.ps>.
- [11] O. L. MANGASARIAN, *Generalized support vector machines*, in Advances in Large Margin Classifiers, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, eds., MIT Press, Cambridge, MA, 2000, pp. 135–146; also available from <ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-14.ps>.
- [12] O. L. MANGASARIAN, *Set containment characterization*, Technical report 01-10, Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, WI, 2001; J. Global Optim., 24 (2002), pp. 473–480; also available from <ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/01-10.ps>.
- [13] O. L. MANGASARIAN AND R. R. MEYER, *Nonlinear perturbation of linear programs*, SIAM J. Control Optim., 17 (1979), pp. 745–752.
- [14] O. L. MANGASARIAN, J. W. SHAVLIK, AND E. W. WILD, *Knowledge-based kernel approximation*, Technical report 03-05, Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, WI, 2003; available from <ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/03-05.ps>.
- [15] B. SCHÖLKOPF, C. BURGESS, AND A. SMOLA, EDS., *Advances in Kernel Methods: Support Vector Machines*, MIT Press, Cambridge, MA, 1999.
- [16] A. SMOLA, P. L. BARTLETT, B. SCHÖLKOPF, AND J. SCHUURMANN, EDS., *Advances in Large Margin Classifiers*, MIT Press, Cambridge, MA, 2000.
- [17] A. SMOLA AND B. SCHÖLKOPF, *Learning with Kernels*, MIT Press, Cambridge, MA, 2002.
- [18] V. N. VAPNIK, *The Nature of Statistical Learning Theory*, 2nd ed., Springer-Verlag, New York, 2000.

POLYNOMIAL PROGRAMMING: LP-RELAXATIONS ALSO CONVERGE*

JEAN B. LASSERRE[†]

Abstract. We consider the global minimization of a multivariate polynomial on a compact semialgebraic set. Using a result of Krivine [*J. Analyse Math.*, 12 (1964), pp. 307–326; *C.R. Acad. Sci. Paris*, 258 (1964), pp. 3417–3418], it is shown that LP-relaxations based on products of the original constraints, in the spirit of the reformulation-linearization technique procedure of Sherali and Adams, converge to the global optimum. This extends to compact semialgebraic sets, previous results known for polytopes, and 0-1 programs. The asymptotic behavior of the LP-relaxations is also analyzed, and it is shown that in important cases, finite convergence cannot occur.

Key words. real algebraic geometry, global optimization, semidefinite programming relaxations, linear programming relaxations

AMS subject classifications. 90C22, 90C25

DOI. 10.1137/040603942

1. Introduction. Semidefinite programming relaxations (or SDP-relaxations) for the polynomial programming (PP) problem

$$(1.1) \quad \mathbb{P} \mapsto f^* := \min_{x \in \mathbb{R}^n} \{f(x) \mid g_i(x) \geq 0, \ i = 1, \dots, m\}$$

(where $f, g_i : \mathbb{R}^n \rightarrow \mathbb{R}$ are real-valued polynomials for all $i = 1, \dots, m$) have attracted a lot of attention in recent years. Apart from the broad spectrum of potential applications (e.g., in control [6] and mathematical finance [14]) this attention is also justified for two main reasons:

(a) *Convergence.* The sequence of optimal values of the SDP-relaxations associated with $\mathbb{P}$, as defined in Lasserre [10], converges to f^* under relatively weak assumptions on the feasible set $\mathbb{K} = \{x \in \mathbb{R}^n \mid g_i(x) \geq 0, \ i = 1, \dots, m\}$. Convergence to a global minimizer (if unique) is even proved in Schweighofer [23].

(b) *Efficiency.* Theoretically, convergence in (a) is asymptotic as the size of the relaxation increases. In fact, so far, practice seems to reveal that in most cases, the convergence is *fast* and even *finite*, which is good news in view of the rapidly growing size of the relaxations. See, for instance, the numerical results on a sample of problems in Henrion and Lasserre [5]. In addition, finite convergence has been proved when some equality constraints $\{g_j(x) = 0\}_{j \in J}$ in the definition of $\mathbb{K}$ define a zero-dimensional radical ideal in $\mathbb{R}[x_1, \dots, x_n]$; see, e.g., Lasserre [11], Laurent [15], and Parrilo [19].

However, LP-relaxations defined earlier in the pioneering work of Sherali and Adams [24, 25], Ceria, Balas, and Cornuejols [3], and Lovász and Schrijver [17], were also proved to have finite convergence for 0-1 programs. As mentioned in Lasserre [12] and Parrilo and Sturmfels [20], and invoking a representation result of Cassier [2] and Handelman [4], LP-relaxations also converge when the g_j 's are linear (i.e., $\mathbb{K}$ is a convex polytope). A comparison of SDP- and LP-relaxations from two dual points of view in Lasserre [12] and Laurent [15], respectively, shows that in principle,

*Received by the editors February 7, 2004; accepted for publication (in revised form) May 24, 2004; published electronically December 30, 2004.

<http://www.siam.org/journals/siopt/15-2/60394.html>

[†]LAAS-CNRS, 7 Avenue du Colonel Roche, 31077 Toulouse Cédex 4, France (lasserre@laas.fr).

SDP-relaxations are preferable to LP-relaxations, for theoretical as well as numerical reasons. But as SDP is still in its infancy, efficient software packages for *large-scale* SDPs are still awaited, which is not the case for LP packages. Therefore, in view of the large size of SDP- (as well as LP-) relaxations to solve, LP-relaxations could provide a viable alternative.

Contribution. In this paper, we show that LP-relaxations in the spirit of the Sherali–Adams reformulation-linearization technique also converge asymptotically for PP problems on compact semialgebraic sets, extending previous results on asymptotic convergence for *convex polytopes* in [12] and [20], and finite convergence for 0-1 programs in [24]. This result follows from a representation result originally due to Krivine [8, 9], later restated in Becker and Schwartz [1], and more recently in Marshall [18] and Vasilescu [26]. We also provide some qualitative analysis of the asymptotic behavior of optimal solutions when the size of the LP-relaxation increases, and we identify cases where none of the LP-relaxations can be exact.

2. LP-relaxations. Let $\mathbb{R}[x_1, \dots, x_n]$ be the ring of real-valued polynomials, and consider the problem $\mathbb{P}$ defined in (1.1) with $f \in \mathbb{R}[x_1, \dots, x_n]$, and feasible set

$$(2.1) \quad \mathbb{K} := \{x \in \mathbb{R}^n \mid g_j(x) \geq 0, \ j = 1, \dots, m\},$$

where $\{g_j\}_{j=1}^m \subset \mathbb{R}[x_1, \dots, x_n]$. Let $g_0 \equiv 1 \in \mathbb{R}[x_1, \dots, x_n]$. We make the following assumption.

Assumption 2.1. (a) The set $\mathbb{K}$ in (2.1) is nonempty and compact.

(b) The polynomials $\{g_j\}_{j=0}^m$ generate the $\mathbb{R}$ -algebra $\mathbb{R}[x_1, \dots, x_n]$.

Assumption 2.1(b) states that in fact $\mathbb{R}[x_1, \dots, x_n] = \mathbb{R}[g_0, g_1, \dots, g_m]$. If Assumption 2.1(b) is not satisfied, one can replace the set of polynomials $\{g_1, \dots, g_m\}$ with another set of polynomials representing $\mathbb{K}$ and generating $\mathbb{R}[x_1, \dots, x_n]$, in the following way. Let $\underline{x}_k \leq \min\{x_k \mid x \in \mathbb{K}\}$ for all $k = 1, \dots, n$. Then, with $x \mapsto g_{m+k}(x) := x_k - \underline{x}_k$, introduce the additional (redundant) constraints $g_{m+k}(x) \geq 0$, $k = 1, \dots, n$, in the definition (2.1) of $\mathbb{K}$, and reset $m := m + n$. With this new equivalent definition, Assumption 2.1 holds. In many cases, lower bounds $\{\underline{x}_k\}$ on x_k are either readily available or can be computed. If not, then it suffices to solve some SDP-relaxation $\mathbb{Q}_r$ (as introduced in Lasserre [10] and restated below in (2.13)), now with $f \equiv x_k$. Indeed, under a weak assumption on the g_j 's (see [10]), every SDP-relaxation $\mathbb{Q}_r$ will provide a finite lower bound, provided that r is sufficiently large.

For every $j = 1, \dots, m$, let $\bar{g}_j := \max_{x \in \mathbb{K}} g_j(x)$ (well defined because $\mathbb{K}$ is compact), and let $\{\hat{g}_j\}_{j=1}^m$ be the polynomials g_j 's, *normalized* with respect to $\mathbb{K}$, that is,

$$(2.2) \quad \hat{g}_j := \begin{cases} g_j/\bar{g}_j & \text{if } \bar{g}_j > 0, \\ g_j & \text{if } \bar{g}_j = 0, \end{cases} \quad j = 1, \dots, m.$$

In fact, one may also take for $\bar{g}_j$ any finite upper bound of g_j on $\mathbb{K}$. Such an upper bound can be obtained as for the lower bound $\underline{x}_k$, i.e., by running some SDP-relaxation $\mathbb{Q}_r$ in (2.13), now with $f := -g_j$ (see the comment just after Assumption 2.1).

Next, let $\mathcal{P} := \{0, 1, \hat{g}_1, \dots, \hat{g}_m, 1 - \hat{g}_1, \dots, 1 - \hat{g}_m\} \subset \mathbb{R}[x_1, \dots, x_n]$, and let $\Delta_{\mathcal{P}} \subset \mathbb{R}[x_1, \dots, x_n]$ be the set of all products of the form $q_1 \cdots q_k$, for polynomials $\{q_j\}_{j=1}^k \subset \mathcal{P}$, and integer $k \geq 1$. Denote by $\mathcal{C}_{\mathbb{K}}$ the *cone* generated by $\Delta_{\mathcal{P}}$, i.e., $p \in \mathcal{C}_{\mathbb{K}}$ if

$$(2.3) \quad p = \sum_{\alpha, \beta \in \mathbb{N}^m} c_{\alpha\beta} \hat{g}_1^{\alpha_1} \cdots \hat{g}_m^{\alpha_m} (1 - \hat{g}_1)^{\beta_1} \cdots (1 - \hat{g}_m)^{\beta_m},$$

for *finitely many nonnegative* coefficients $\{c_{\alpha\beta}\} \subset \mathbb{R}_+$ (and with $\widehat{g}_k^0 \equiv 1$), or using the vector notation

$$\widehat{g} = \begin{bmatrix} \widehat{g}_1 \\ \vdots \\ \widehat{g}_m \end{bmatrix}, \quad (1 - \widehat{g}) = \begin{bmatrix} (1 - \widehat{g}_1) \\ \vdots \\ (1 - \widehat{g}_m) \end{bmatrix},$$

$p \in \mathcal{C}_{\mathbb{K}}$ if p has the compact form

$$p = \sum_{\alpha, \beta \in \mathbb{N}^m} c_{\alpha\beta} \widehat{g}^\alpha (1 - \widehat{g})^\beta.$$

Equivalently, $p \in \mathcal{C}_{\mathbb{K}}$ if p is a polynomial of $\mathbb{R}[\widehat{g}_1, \dots, \widehat{g}_m, 1 - \widehat{g}_1, \dots, 1 - \widehat{g}_m]$, with nonnegative coefficients.

For $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{N}^m$ let $|\alpha| := \sum_j \alpha_j$, and define the following LP problem $\mathbb{L}_r$, for $r \in \mathbb{N}$:

$$(2.4) \quad \mathbb{L}_r \begin{cases} \max_{\lambda, c_{\alpha\beta}} \lambda, \\ \text{s.t.} \quad f - \lambda = \sum_{\alpha, \beta \in \mathbb{N}^m} c_{\alpha\beta} \widehat{g}^\alpha (1 - \widehat{g})^\beta, \quad \deg \widehat{g}^\alpha (1 - \widehat{g})^\beta \leq 2r, \\ c_{\alpha\beta} \geq 0. \end{cases}$$

The relaxation $\mathbb{L}_r$ is indeed a linear program because identifying coefficients of the polynomials $f - \lambda$, and $\sum_{\alpha, \beta \in \mathbb{N}^m} c_{\alpha\beta} \widehat{g}^\alpha (1 - \widehat{g})^\beta$ (with $\deg \widehat{g}^\alpha (1 - \widehat{g})^\beta \leq 2r$), yields a finite system of linear equalities on the nonnegative variables $c_{\alpha\beta}$'s. Denote by $\max \mathbb{L}_r$ the optimal value of the LP $\mathbb{L}_r$, including $-\infty$ whenever $\mathbb{L}_r$ has no feasible solution.

THEOREM 2.2. *Let $\{f, g_j\}_{j=1}^m \subset \mathbb{R}[x_1, \dots, x_n]$. With $\mathbb{K}$ as in (2.1), let Assumption 2.1 hold. Then*

$$(2.5) \quad \max \mathbb{L}_r \uparrow \min \mathbb{P} \quad \text{as } r \rightarrow \infty.$$

Proof. First, observe that every feasible solution of $\mathbb{L}_r$ is feasible for $\mathbb{L}_{r+1}$ and therefore the optimal value $\max \mathbb{L}_r$ is nondecreasing in r , i.e., $\max \mathbb{L}_r \uparrow$. Assume that $\max \mathbb{L}_r$ has a finite optimal value λ_r^* so that the polynomial $f - \lambda_r^*$ can be written as p in (2.3). This clearly implies $f - \lambda_r^* \geq 0$ on $\mathbb{K}$. Hence, it follows that $\lambda_r^* \leq \min \mathbb{P}$ for all $r \in \mathbb{N}$. Conversely, let $\lambda^* := \min \mathbb{P}$, and let $\epsilon > 0$ be fixed, arbitrary. The polynomial $f - \lambda^* + \epsilon$ is strictly positive on $\mathbb{K}$. Therefore, following Krivine [8, 9] (see also Vasilescu [26, Thm. 1.3.1, p. 183]), $f - \lambda^* + \epsilon \in \mathcal{C}_{\mathbb{K}}$, that is,

$$(2.6) \quad f - \lambda^* + \epsilon = \sum_{\alpha, \beta \in \mathbb{N}^m} c_{\alpha\beta}(\epsilon) \widehat{g}^\alpha (1 - \widehat{g})^\beta,$$

for finitely many nonnegative coefficients $\{c_{\alpha\beta}(\epsilon)\}$. Take

$$r(\epsilon) := \max_{\alpha, \beta \in \mathbb{N}^m} \{\deg \widehat{g}^\alpha (1 - \widehat{g})^\beta \mid c_{\alpha\beta}(\epsilon) > 0 \text{ in (2.6)}\},$$

and let $r \in \mathbb{N}$ be the smallest integer such that $2r \geq r(\epsilon)$. Then $\{\lambda^* - \epsilon, \{c_{\alpha\beta}(\epsilon)\}\}$ is a feasible solution of the linear program $\mathbb{L}_r$, with value $\lambda^* - \epsilon$. Therefore, $\lambda^* - \epsilon \leq \max \mathbb{L}_r \leq \lambda^*$, and as $\epsilon > 0$ was arbitrary, the result follows. $\square$

Remark 2.3. (i) Theorem 2.2 was proved in Lasserre [12] for the case where the g_j 's are linear and define a convex polytope $\mathbb{K}$. We used Cassier [2] and Handelman [4]

results on the representation of polynomials, positive on a polytope. This convergence result is also stated in Parrilo and Sturmfels [20].

(ii) For nonlinear 0-1 programs, i.e., when the additional constraints $h_j(x) = 0$ are added to the definition (2.1) $\mathbb{K}$, with

$$x \mapsto h_j(x) := x_j^2 - x_j, \quad j = 1, \dots, n,$$

then Sherali and Adams have shown the finite convergence of LP-relaxations related to the LP-relaxations $\{\mathbb{L}_r\}$ defined in (2.4). In fact, they do not include products of the polynomials g_j 's that define the set $\mathbb{K}$ (without the integrality constraints). They involve only products of the form $g_j \prod_{i \in I} x_i \prod_{k \in J} (1 - x_k)$ for disjoint $I, J \subset \{1, \dots, m\}$. For a detailed analysis and comparison of SDP- and LP-relaxations for 0-1 programs, the interested reader is referred to Lasserre [12] and Laurent [15]. See also Lasserre and Prieto Rumeau [13] for a comparison in some performance evaluation problems in mathematical finance.

2.1. Dual LP-relaxations. The dual LP-relaxations have an interpretation in terms of the $\mathbb{K}$ -moment problem, that is, the characterization of sequences of scalars that are moments of some measure with support contained in $\mathbb{K}$.

Let $\mathbb{K} \subset \mathbb{R}^n$ be the semialgebraic set defined in (2.1), and let $\gamma := \{\gamma_\alpha\}_{\alpha \in \mathbb{N}^n}$ be a sequence of scalars. Then γ is a $\mathbb{K}$ -moment sequence if there exists a so-called *representing* Borel measure μ on $\mathbb{R}^n$, whose *support* is contained in $\mathbb{K}$, that is, $\mu(\mathbb{R}^n \setminus \mathbb{K}) = 0$, and

$$\gamma_\alpha = \int x^\alpha d\mu, \quad \alpha \in \mathbb{N}^n.$$

Consider the linear form $L_\gamma : \mathbb{R}[x_1, \dots, x_n] \rightarrow \mathbb{R}$ given by

$$(2.7) \quad f \left(:= \sum_{\alpha \in \mathbb{N}^n} f_\alpha x^\alpha \right) \mapsto L_\gamma(f) = \sum_{\alpha \in \mathbb{N}^n} f_\alpha \gamma_\alpha, \quad f \in \mathbb{R}[x_1, \dots, x_n].$$

Then, under Assumption 2.1, Theorem 1.2.2 in Vasilescu [26] states that γ has a representing measure μ with support contained in $\mathbb{K}$, if and only if

$$(2.8) \quad L_\gamma(p) \geq 0 \quad \forall p \in \mathcal{C}_{\mathbb{K}},$$

or, equivalently,

$$(2.9) \quad L_\gamma(\widehat{g}^\alpha(1 - \widehat{g})^\beta) \geq 0, \quad \alpha, \beta \in \mathbb{N}^m.$$

Observe that (2.9) defines countably many linear inequality constraints on the variables γ_α . For instance, with $n = 1$, $\mathbb{K} = [0, 1]$, and $g_1 = 1 - g_2 \equiv x$, (2.9) reads

$$\sum_{j=0}^m (-1)^j \binom{m}{j} \gamma_{k+j} \geq 0, \quad m, k = 0, 1, \dots,$$

which are the well-known *Hausdorff moment conditions* for γ to have a representing measure supported on $[0, 1]$.

It is not difficult to see that the (linear program) dual $\mathbb{L}_r^*$ of the LP-relaxation $\mathbb{L}_r$ is the (finite) linear program

$$(2.10) \quad \min_{\gamma} \{L_\gamma(f) \mid L_\gamma(\widehat{g}^\alpha(1 - \widehat{g})^\beta) \geq 0, \deg \widehat{g}^\alpha(1 - \widehat{g})^\beta \leq 2r, \gamma_0 = 1\}.$$

2.2. Comparing with SDP-relaxations. We have seen that the primal LP-relaxations $\{\mathbb{L}_r\}$ are based on the representation (2.3) of the polynomial $f - \lambda$, positive on $\mathbb{K}$, whereas the dual LP-relaxations $\{\mathbb{L}_r^*\}$ are based on the *moment conditions* (2.9).

On the other hand, the analogue primal and dual SDP-relaxations introduced in Lasserre [10] use a different representation (and associated dual moment conditions), which is an extension of Schmüdgen's Positivstellensatz [22] due to Putinar [21], and Jacobi and Prestel [7]. Namely, under some assumption on the g_j 's that define the set $\mathbb{K}$ in (2.1), a polynomial $h \in \mathbb{R}[x_1, \dots, x_n]$, positive on $\mathbb{K}$, can be written (again denoting $g_0 \equiv 1$)

$$(2.11) \quad h = \sum_{j=0}^m q_j g_j$$

for some polynomials $\{q_j\}_{j=0}^m \subset \mathbb{R}[x_1, \dots, x_n]$, all *sums of squares* (s.o.s.). Alternatively, the associated moment conditions state that an infinite sequence $\gamma = \{\gamma_\alpha\}$ has a representing measure μ on $\mathbb{R}^n$, with support contained in $\mathbb{K}$, if

$$(2.12) \quad L_\gamma(h^2 g_j) \geq 0 \quad \forall h \in \mathbb{R}[x_1, \dots, x_n], \quad j = 0, \dots, m.$$

For more details, the interested reader is referred to Lasserre [10].

Hence, the primal SDP analogues of the LP-relaxations $\{\mathbb{L}_r\}$ are the semidefinite programs

$$(2.13) \quad \mathbb{Q}_r \begin{cases} \max_{\lambda, \{q_j\}} \lambda \\ \text{s.t.} \quad f - \lambda = \sum_{j=0}^m q_j g_j, \\ q_j \text{ s.o.s.} \quad \forall j = 0, \dots, m, \\ \deg(q_j g_j) \leq 2r \quad \forall j = 0, \dots, m \end{cases}$$

(to compare with (2.4)), whereas the dual SDP analogues of the LP-relaxations $\{\mathbb{L}_r^*\}$ read

$$(2.14) \quad \mathbb{Q}_r^* \begin{cases} \min_{\gamma} L_\gamma(f) \\ \text{s.t.} \quad \gamma_0 = 1; \quad L_\gamma(h^2 g_j) \geq 0 \quad \forall h, \deg(h^2 g_j) \leq 2r, \quad j = 0, \dots, m \end{cases}$$

(to compare with (2.10)).

The relaxation $\mathbb{Q}_r$ is indeed a SDP because the constraint that q_j is a s.o.s. polynomial is equivalent to stating that a symmetric matrix associated with q_j is positive semidefinite. Similarly, the relaxation $\mathbb{Q}_r^*$ is also a SDP because the constraint $L_\gamma(h^2 g_j) \geq 0$ for all $h \in \mathbb{R}[x_1, \dots, x_n]$ with $\deg(h^2 g_j) \leq 2r$, is equivalent to state that the *localizing* matrix associated with g_j and γ (and the *moment* matrix associated with $g_0 \equiv 1$) is positive semidefinite; for more details the reader is referred to Lasserre [10].

Hence, for the same value of $r \in \mathbb{N}$, both dual LP- and SDP-relaxations $\mathbb{L}_r^*$ and $\mathbb{Q}_r^*$ have the same number $\binom{n+2r}{n}$ of variables γ_α (associated with the monomials x^α with $|\alpha| \leq 2r$). The conditions $L_\gamma(\hat{g}^\alpha(1-\hat{g})^\beta) \geq 0$ define linear inequality constraints in $\mathbb{L}_r^*$, each associated with a $2m$ -uplet $\alpha_1, \dots, \alpha_m, \beta_1, \dots, \beta_m$ such that $\deg \hat{g}^\alpha(1-\hat{g})^\beta \leq 2r$. For instance, when the g_j 's are linear we have $O((2r)^{2m})$ such inequalities. On the other hand, the conditions $L_\gamma(h^2 g_j) \geq 0$ in the SDP-relaxation $\mathbb{Q}_r^*$ yield $m+1$ semidefinite constraints on the moment and localizing matrices associated with γ and the g_j 's. The size of the moment matrix is $\binom{n+r}{n}$.

Similarly, both $\mathbb{L}_r$ and $\mathbb{Q}_r$ have the same number of equality constraints on their respective variables (equating two polynomials on degree at most $2r$). The LP-relaxations variables are the nonnegative coefficients $\{c_{\alpha\beta}\}$ in the representation of $f - \lambda$ in (2.4), whereas the SDP-relaxations variables are the $m+1$ positive semidefinite matrices associated with the s.o.s. polynomials $\{q_j\}_{j=0}^m$ in the representation of $f - \lambda$ in (2.13). The comparison is summarized in Table 1.

In the context of 0-1 programs, the SDP-relaxations $\{\mathbb{Q}_r^*\}$ have been analyzed in detail in Laurent [15] and compared with the LP-relaxations of Sherali and Adams and Lovász and Schrijver (see also Lasserre [12]). In [15] it is shown that the SDP-relaxations are stronger than their LP counterpart (that is, when comparing $\mathbb{L}_r$ and $\mathbb{Q}_r$, which use the same degree bound $2r$ in the representation of $f - \lambda$). However, the computational burden involved in $\mathbb{L}_r$ and $\mathbb{Q}_r$ is more difficult to compare.

TABLE 1
Relaxations $\mathbb{L}_r$, $\mathbb{Q}_r$ and their respective duals.

Primal LP-relaxation $\mathbb{L}_r$	Primal SDP-relaxation $\mathbb{Q}_r$
$\begin{aligned} &\max_{\lambda, \{c_{\alpha\beta}\}} \lambda \\ &f - \lambda = \sum_{\alpha, \beta \in \mathbb{N}^m} c_{\alpha\beta} \widehat{g}^\alpha (1 - \widehat{g})^\beta \\ &\deg \widehat{g}^\alpha (1 - \widehat{g})^\beta \leq 2r \\ &c_{\alpha\beta} \geq 0 \quad \forall \alpha, \beta \end{aligned}$	$\begin{aligned} &\max_{\lambda, \{q_j\}} \lambda \\ &f - \lambda = \sum_{j=0}^m q_j g_j \\ &\deg(q_j g_j) \leq 2r \\ &q_j \text{ s.o.s.} \\ &j \in \{0, \dots, m\} \end{aligned}$
Dual LP-relaxation $\mathbb{L}_r^*$	Dual SDP-relaxation $\mathbb{Q}_r^*$
$\begin{aligned} &\min_{\gamma} L_\gamma(f) \\ &\gamma_0 = 1; \quad L_\gamma(\widehat{g}^\alpha (1 - \widehat{g})^\beta) \geq 0 \\ &\forall \alpha, \beta, \quad \deg \widehat{g}^\alpha (1 - \widehat{g})^\beta \leq 2r \end{aligned}$	$\begin{aligned} &\min_{\gamma} L_\gamma(f) \\ &\gamma_0 = 1; \quad L_\gamma(h^2 g_j) \geq 0 \\ &\forall h, \deg(h g_j) \leq 2r \\ &j \in \{0, \dots, m\} \end{aligned}$

3. Asymptotic behavior of the LP-relaxations. In this section, we analyze the behavior of optimal solutions of the LP-relaxations when $r \rightarrow \infty$. Notice that for large values of r , both LP-relaxations $\{\mathbb{L}_r\}$ and $\{\mathbb{L}_r^*\}$ will suffer a numerical ill-conditioning due to large binomials coefficients in the expansion of $(1 - \widehat{g}_j)^{\beta_j}$. Therefore, it is important to identify cases where the LP-relaxations do *not* have finite convergence, in which case this numerical ill-conditioning eventually appears.

Notice that the LP-relaxations can still be defined even if Assumption 2.1 does not hold. However, if $f - \lambda$ cannot be represented as in (2.4) no matter how large r can be, they will not have finite value. Or, even if they have finite value, they may not converge to the global optimum f^* .

THEOREM 3.1. *Let $\mathbb{K}$ be as in (2.1), and assume that the set*

$$(3.1) \quad \overset{\circ}{\mathbb{K}} = \{x \in \mathbb{R}^n \mid g_j(x) > 0, \bar{g}_j - g_j(x) > 0, j = 1, \dots, m\}$$

is nonempty. Let $f \in \mathbb{R}[x_1, \dots, x_n]$ have a global minimum f^* on $\mathbb{K}$, and let

$$(\lambda_r, \{c_{\alpha\beta}^r\}_{\alpha, \beta \in \mathbb{N}^m} \subset \mathbb{R}_+)$$

be an optimal solution of the LP-relaxation $\mathbb{L}_r$ (assumed to exist provided that $r \in \mathbb{N}$ is sufficiently large, say $r \geq r_0$ for some $r_0 \in \mathbb{N}$). Then we have the following:

(a) There is a subsequence $\{r_j\} \subset \mathbb{N}$ and an infinite nonnegative sequence $\{c_{\alpha\beta}^*\}_{\alpha,\beta \in \mathbb{N}^m} \subset \mathbb{R}_+$ such that, as $j \rightarrow \infty$,

$$(3.2) \quad c_{\alpha\beta}^{r_j} \rightarrow c_{\alpha\beta}^*, \quad \alpha, \beta \in \mathbb{N}^m,$$

and everywhere on $\mathbb{K}$,

$$(3.3) \quad f - \lim_{r \rightarrow \infty} \lambda_r \geq \sum_{\alpha, \beta \in \mathbb{N}^m} c_{\alpha\beta}^* \widehat{g}^\alpha (1 - \widehat{g})^\beta.$$

(b) In addition, under Assumption 2.1, let $x^* \in \mathbb{K}$ be a global minimizer of f on $\mathbb{K}$. Then

$$(3.4) \quad f - f^* \geq \sum_{\alpha, \beta \in \mathbb{N}^m} c_{\alpha\beta}^* \widehat{g}^\alpha (1 - \widehat{g})^\beta$$

everywhere on $\mathbb{K}$, and

$$(3.5) \quad c_{\alpha\beta}^* > 0 \Rightarrow \begin{cases} g_j(x^*) = 0 & \text{for some } \alpha_j > 0, \\ \overline{g_k} - g_k(x^*) = 0 & \text{for some } \beta_k > 0. \end{cases} \quad \text{or}$$

(c) Under Assumption 2.1, if f has a global minimizer $x^* \in \overset{\circ}{\mathbb{K}}$ so that $\overline{g}_j - g_j(x^*) > 0$ for all $j = 1, \dots, m$, then $c_{\alpha\beta}^* = 0$ for all $\alpha, \beta \in \mathbb{N}^m$.

(d) The conclusions (a)–(c) are true for any accumulation point of the sequence $\{\{c_{\alpha\beta}^r\}_{\alpha,\beta}\}_r$ as $r \rightarrow \infty$ (each element $\{c_{\alpha\beta}^r\}_{\alpha,\beta}$ being considered as an element of the space l_∞ of bounded sequences).

Proof. (a) By assumption, provided $r \in \mathbb{N}$ is sufficiently large, say $r \geq r_0$ for some $r_0 \in \mathbb{N}$, the LP-relaxations $\mathbb{L}_r$ have finite optimal value λ_r , and some optimal solution $\{c_{\alpha\beta}^r\}_{\alpha,\beta \in \mathbb{N}^m} \subset \mathbb{R}_+$ (a finite sequence). Fix $x_0 \in \overset{\circ}{\mathbb{K}}$, arbitrary. Then

$$(3.6) \quad f(x_0) - \lambda_r = \sum_{\alpha, \beta \in \mathbb{N}^m} c_{\alpha\beta}^r \widehat{g}(x_0)^\alpha (1 - \widehat{g}(x_0))^\beta, \quad r \geq r_0.$$

In addition, as $\lambda_r \uparrow$, and with

$$0 < \mu_{\alpha\beta} := \widehat{g}(x_0)^\alpha (1 - \widehat{g}(x_0))^\beta, \quad \alpha, \beta \in \mathbb{N}^m,$$

(3.6) yields

$$(3.7) \quad 0 \leq c_{\alpha\beta}^r \leq \frac{f(x_0) - \lambda_{r_0}}{\mu_{\alpha\beta}}, \quad \alpha, \beta \in \mathbb{N}^m.$$

For each $r \geq r_0$, complete with zeros to make the sequence $\{\widehat{c}_{\alpha\beta}^r := \mu_{\alpha\beta} c_{\alpha\beta}^r\}_{\alpha,\beta \in \mathbb{N}^m} \subset \mathbb{R}_+$, an element of the space l_∞ of bounded sequences, equipped with the sup-norm. From (3.7) notice that

$$\widehat{c}_{\alpha\beta}^r \leq f(x_0) - \lambda_{r_0}, \quad \alpha, \beta \in \mathbb{N}^m, \quad r \geq r_0;$$

that is, the sequences $\{\widehat{c}_{\alpha\beta}^r\}_{\alpha,\beta \in \mathbb{N}^m}$, $r \geq r_0$, are all in some ball of l_∞ . Therefore, there is a sequence $\{\widehat{c}_{\alpha\beta}^*\}_{\alpha,\beta \in \mathbb{N}^m} \in l_\infty$, and a subsequence $\{r_j\} \subset \mathbb{N}$, such that, as $j \rightarrow \infty$,

$\{\widehat{c}_{\alpha\beta}^{r_j}\} \rightarrow \{\widehat{c}_{\alpha\beta}^*\}$ for the $\sigma(l_\infty, l_1)$ weak-star topology of l_∞ . In particular, $\widehat{c}_{\alpha\beta}^{r_j} \rightarrow \widehat{c}_{\alpha\beta}^*$ for all $\alpha, \beta \in \mathbb{N}^m$, or, equivalently, as $j \rightarrow \infty$,

$$(3.8) \quad c_{\alpha\beta}^{r_j} \rightarrow c_{\alpha\beta}^*, \quad \alpha, \beta \in \mathbb{N}^m,$$

with $c_{\alpha\beta}^* := \widehat{c}_{\alpha\beta}^* / \mu_{\alpha\beta}$ for all $\alpha, \beta \in \mathbb{N}^m$. This proves (3.2).

Fix $x \in \mathbb{K}$ arbitrary. By Fatou's lemma we obtain

$$\begin{aligned} f(x) - \lim_{j \rightarrow \infty} \lambda_{r_j} &= \lim_{j \rightarrow \infty} (f(x) - \lambda_{r_j}) \\ &= \liminf_{j \rightarrow \infty} (f(x) - \lambda_{r_j}) \\ &\geq \sum_{\alpha, \beta \in \mathbb{N}^m} c_{\alpha\beta}^* \widehat{g}(x)^\alpha (1 - \widehat{g}(x))^\beta, \end{aligned}$$

and so (3.3) holds, which proves (a).

(b) From Theorem 2.2, $\lambda_r \uparrow \min \mathbb{P} = f^*$ as $r \rightarrow \infty$. So, (3.3) becomes (3.4), and with $x = x^*$, we get

$$f(x^*) - f^* = 0 \geq \sum_{\alpha, \beta \in \mathbb{N}^m} c_{\alpha\beta}^* \widehat{g}(x^*)^\alpha (1 - \widehat{g}(x^*))^\beta,$$

from which we deduce

$$c_{\alpha\beta}^* = 0 \quad \text{whenever } \widehat{g}(x^*)^\alpha (1 - \widehat{g}(x^*))^\beta > 0.$$

In other words, the only nontrivial coefficients $c_{\alpha\beta}^*$ are those associated with products that contain at least one polynomial g_j or $\bar{g}_k - g_k$, active at the global minimizer x^* .

This proves (b), and (c) follows from (b) when $x^* \in \overset{\circ}{\mathbb{K}}$. Finally, (d) follows because the converging subsequence $\{r_j\}$ was arbitrary. $\square$

3.1. Discussion. We next analyze some features of the LP-relaxations, already observed for the case where $\mathbb{K}$ is a polytope. So let $\mathbb{K}$ be as in (2.1), and again, for every $j = 1, \dots, m$, let $\widehat{g}_j$ be the polynomial associated with g_j and normalized with respect to $\mathbb{K}$, as in (2.2).

We say that the LP-relaxations $\{\mathbb{L}_r\}$ are *exact* if $\max \mathbb{L}_{r_0} = \min \mathbb{P}$ (hence $\max \mathbb{L}_r = \min \mathbb{P}$ for all $r \geq r_0$) for some index $r_0 \in \mathbb{N}$. Given a global minimizer $x^* \in \mathbb{K}$ of $\mathbb{P}$, let

$$I(x^*) := \{j \in \{1, \dots, m\} \mid \widehat{g}_j(x^*) = 0\}$$

and

$$J(x^*) := \{j \in \{1, \dots, m\} \mid (1 - \widehat{g}_j(x^*)) = 0\}.$$

PROPOSITION 3.2. *Let $\{f, g_j\}_{j=1}^m \subset \mathbb{R}[x_1, \dots, x_n]$, and $\mathbb{K}$ be as in (2.1). Then none of the LP-relaxations $\{\mathbb{L}_r\}$ defined in (2.4) can be exact if there exists a global minimizer $x^* \in \mathbb{K}$, and a nonoptimal point $x \in \mathbb{K}$ such that $\widehat{g}_j(x) = 0$ for all $j \in I(x^*)$ and $\widehat{g}_k(x) = 1$ for all $k \in J(x^*)$.*

Proof. The proof is by contradiction. Assume that the LP-relaxations $\{\mathbb{L}_r\}$ are exact; i.e., $\max \mathbb{L}_r = f^* = \min \mathbb{P}$ for some index $r \in \mathbb{N}$. This implies that

$$f - f^* = \sum_{\alpha, \beta \in \mathbb{N}^m} c_{\alpha\beta} \widehat{g}^\alpha (1 - \widehat{g})^\beta$$

for some nonnegative coefficients $\{c_{\alpha\beta}\} \subset \mathbb{R}_+$. Evaluating $f - f^*$ at an arbitrary point $x \in \mathbb{K}$ yields

$$(3.9) \quad f(x) - f^* = \sum_{\alpha, \beta \in \mathbb{N}^m} c_{\alpha\beta} \widehat{g}(x)^\alpha (1 - \widehat{g}(x))^\beta.$$

In particular, at the global minimizer $x^* \in \mathbb{K}$,

$$(3.10) \quad 0 = \sum_{\alpha, \beta \in \mathbb{N}^m} c_{\alpha\beta} \widehat{g}(x^*)^\alpha (1 - \widehat{g}(x^*))^\beta.$$

From (3.9) we must have $c_{\alpha\beta} > 0$ for some $\alpha, \beta \in \mathbb{N}^m$. Otherwise $f(x) - f^* = 0$ for all $x \in \mathbb{K}$.

Next, in view of (3.10), for all strictly positive coefficients $\{c_{\alpha\beta}\}$, we also must have $\alpha_j > 0$ for some $j \in I(x^*)$, or $\beta_k > 0$ for some $k \in J(x^*)$. But then, evaluating $f - f^*$ at some nonoptimal $x_0 \in \mathbb{K}$ such that $\widehat{g}_j(x_0) = 0$ for all $j \in I(x^*)$ and $\widehat{g}_k(x_0) = 1$ for all $k \in J(x^*)$ yields the contradiction

$$(3.11) \quad 0 < f(x_0) - f^* = \sum_{\alpha, \beta \in \mathbb{N}^m} c_{\alpha\beta} \widehat{g}(x_0)^\alpha (1 - \widehat{g}(x_0))^\beta = 0. \quad \square$$

For instance, Proposition 3.2 holds if $\mathbb{K}$ contains a global minimizer x^* such that $I(x^*) = J(x^*) = \emptyset$, that is, $x^* \in \overset{\circ}{\mathbb{K}}$, with $\overset{\circ}{\mathbb{K}}$ as in (3.1). Indeed, this easily follows from Theorem 3.1(c). In this latter case, in the limit as $r \rightarrow \infty$, all coefficients $\{c_{\alpha\beta}^r\}$ vanish, another source of ill-conditioning for the LP-relaxations $\mathbb{L}_r$, for large values of r .

Finally, in the particular case when $\mathbb{K}$ is a convex polytope, the only case where the relaxations may be exact is if *every* global minimizer is a *vertex* of $\mathbb{K}$ (e.g., if f is *concave* or linear) or if $f - f^*$ vanishes on every face of $\mathbb{K}$ that contains a global minimizer x^* .

Example 3.3. Let $n = 1$ and $\mathbb{K} = [0, 1] = \{x \in \mathbb{R} \mid x \geq 0; (1 - x) \geq 0\}$, and $x \mapsto f(x) := x - x^2$. Take $x \mapsto g_1(x) = (1 - g_2(x)) = x$. In this case, f is concave and we have the nonconvex optimization problem $\mathbb{P}_1 : \min\{x - x^2 \mid g_1(x), g_2(x) \geq 0\}$, for which $x_1^* = 1$ and $x_2^* = 0$ are both global minimizers, with optimal value $f^* = \min \mathbb{P}_1 = 0$. Notice that

$$f(x) - f^* = x(1 - x) = g_1(x)g_2(x), \quad x \in \mathbb{R},$$

which proves that the LP-relaxation $\mathbb{L}_1$ is exact (notice that each global minimizer is a vertex of the polytope $[0, 1]$). Similarly, we have

$$f(x) - f^* = x(1 - x) = (1 - x)^2 g_1(x) + x^2 g_2(x), \quad x \in \mathbb{R},$$

which also proves that the SDP-relaxation $\mathbb{Q}_2$ is also exact.

On the other hand, if we now replace f with $-f$, we obtain the convex optimization problem $\mathbb{P}_2 : \min\{x^2 - x \mid g_1(x), g_2(x) \geq 0\}$, with global minimum $f^* = -1/4$. Observe that

$$f(x) - f^* = x^2 - x + 0.25 = (0.5 - x)^2, \quad x \in \mathbb{R},$$

which proves that the SDP-relaxation $\mathbb{Q}_1$ is exact. In contrast, *none* of the LP-relaxations $\{\mathbb{L}_r\}$ can be exact (notice that the global minimizer $x^* = 0.5$ is in the

interior of $\mathbb{K}$). As g_1, g_2 generate the $\mathbb{R}$ -algebra $\mathbb{R}[x]$, Assumption 2.1 holds, and from Theorem 2.2, the LP-relaxations $\{\mathbb{L}_r\}$ converge; that is, $\max \mathbb{L}_r \uparrow -1/4$ as $r \rightarrow \infty$. But, from Theorem 3.1(c), the coefficients $\{c_{ij}^r\}$ in the representation

$$x^2 - x - \lambda_r = \sum_{i,j \in \mathbb{N}} c_{ij}^r x^i (1-x)^j, \quad x \in \mathbb{R},$$

vanish as $r \rightarrow \infty$.

So we have the paradox that the LP-relaxations $\mathbb{L}_r$ are more efficient for the nonconvex optimization $\mathbb{P}_1$ than for its convex counterpart $\mathbb{P}_2$; even more, for the latter, the convergence $\lambda_r \uparrow f^* = -1/4$ is very slow. For instance,

$$\lambda_1 = \lambda_2 = -1/3; \quad \lambda_3 = -0.3; \quad \lambda_5 = -0.27, \quad \dots$$

3.2. Linear programming. Let $A \in \mathbb{R}^{m \times n}$, $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$. We now show that the linear program

$$(3.12) \quad \mathbb{P} : \min_x \{c'x \mid Ax = b; x \geq 0\}$$

can be rephrased as

$$(3.13) \quad \max_{\mu \geq 0, \lambda, \gamma} \left\{ \gamma \mid c'x - \gamma = \sum_{j=1}^m \lambda_j (Ax - b)_j + \sum_{k=1}^n \mu_k x_k \quad \forall x \in \mathbb{R}^n \right\};$$

that is, the LP relaxation $\mathbb{L}_1$ of $\mathbb{P}$ is exact (which was to be expected) and yields the linear program $\mathbb{P}^*$, the dual of $\mathbb{P}$.

Indeed, if

$$c'x - \gamma = \sum_{j=1}^m \lambda_j (Ax - b)_j + \sum_{k=1}^n \mu_k x_k \quad \forall x \in \mathbb{R}^n,$$

then equating coefficients of terms with the same powers in both sides yields

$$A'\lambda + \mu = c; \quad b'\lambda = \gamma,$$

and so, as $\mu \geq 0$, (3.13) is just $\max\{b'\lambda \mid A'\lambda \leq c\}$, the usual LP dual $\mathbb{P}^*$ of $\mathbb{P}$.

Similarly, the dual $\mathbb{P}^*$ can be rephrased as

$$(3.14) \quad \mathbb{P} : -\max_{x \geq 0, \gamma} \left\{ \gamma \mid -b'\lambda - \gamma = \sum_{j=1}^n x_j (c - A'\lambda)_j \quad \forall \lambda \in \mathbb{R}^m \right\}$$

or, equivalently,

$$\min_{x \geq 0, \gamma} \left\{ \gamma \mid \gamma - b'\lambda = \sum_{j=1}^n x_j (c - A'\lambda)_j \quad \forall \lambda \in \mathbb{R}^m \right\},$$

which, by the same arguments, yields $\mathbb{P}$.

Acknowledgment. The author wishes to thank M. Schweighofer and M. Laurent for several fruitful discussions at MSRI (Berkeley) in April 2004, which helped greatly to improve the initial version of the paper (in particular to add missing important references).

REFERENCES

- [1] E. BECKER AND N. SCHWARTZ, *Zum Darstellungssatz von Kadison-Dubois*, Arch. Math. (Basel), 40 (1983), pp. 421–428.
- [2] G. CASSIER, *Problème des moments sur un compact de $\mathbb{R}^n$ et décomposition de polynômes à plusieurs variables*, J. Funct. Anal., 58 (1984), pp. 254–266.
- [3] S. CERIA, E. BALAS, AND G. CORNUJOLS, *A lift-and-project cutting plane algorithm for mixed 0-1 programs*, Math. Programming, 58 (1993), pp. 295–324.
- [4] D. HANDELMAN, *Representing polynomials by positive linear functions on compact convex polyhedra*, Pacific J. Math., 132 (1988), pp. 35–62.
- [5] D. HENRION AND J. B. LASSERRE, *GloptiPoly: Global optimization over polynomials with Matlab and SeDuMi*, ACM Trans. Math. Software, 29 (2003), pp. 165–194.
- [6] D. HENRION AND J. B. LASSERRE, *Solving nonconvex optimization problems*, IEEE Control Systems Magazine, 24 (2004), pp. 72–83.
- [7] T. JACOBI AND A. PRESTEL, *Distinguished representations of strictly positive polynomials*, J. Reine Angew. Math., 532 (2001), pp. 223–235.
- [8] J. L. KRIVINE, *Anneaux préordonnés*, J. Analyse Math., 12 (1964), pp. 307–326.
- [9] J. L. KRIVINE, *Quelques propriétés des préordres dans les anneaux commutatifs unitaires*, C.R. Acad. Sci. Paris, 258 (1964), pp. 3417–3418.
- [10] J. B. LASSERRE, *Global optimization with polynomials and the problem of moments*, SIAM J. Optim., 11 (2001), pp. 796–817.
- [11] J. B. LASSERRE, *An explicit equivalent positive semidefinite program for nonlinear 0-1 programs*, SIAM J. Optim., 12 (2002), pp. 756–769.
- [12] J. B. LASSERRE, *Semidefinite programming vs. LP relaxations for polynomial programming*, Math. Oper. Res., 27 (2002), pp. 347–360.
- [13] J. B. LASSERRE AND T. PRIETO RUMEAU, *SDP vs LP-relaxations for the moment approach in some performance evaluation problems*, Stoch. Models, 20 (2004), pp. 439–456.
- [14] J. B. LASSERRE, T. PRIETO RUMEAU, AND M. ZERVOS, *Pricing a Class of Exotic Options Via Moments and SDP Relaxations*, Tech. report 03125, LAAS-CNRS, Toulouse, France, 2003.
- [15] M. LAURENT, *A comparison of the Sherali-Adams, Lovász-Schrijver and Lasserre relaxations for 0-1 programming*, Math. Oper. Res., 28 (2003), pp. 470–496.
- [16] M. LAURENT, *Semidefinite representations for finite varieties*, Math. Progr., to appear.
- [17] L. LOVÁSZ AND A. SCHRIJVER, *Cones of matrices and set-functions and 0-1 optimization*, SIAM J. Optim., 1 (1991), pp. 166–190.
- [18] M. MARSHALL, *A general representation theorem for partially ordered commutative rings*, Math. Z., 242 (2002), pp. 217–225.
- [19] P. PARRILO, *An Explicit Construction of Distinguished Representations of Polynomials Non-negative over Finite Sets*, Tech. report AUTO 2-02, ETH, Zurich, Switzerland, 2002.
- [20] P. PARRILO AND B. STURMFELS, *Minimizing polynomial functions*, DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 60, AMS, Providence, RI, 2003, pp. 83–100.
- [21] M. PUTINAR, *Positive polynomials on compact semi-algebraic sets*, Indiana Univ. Math. J., 42 (1993), pp. 969–984.
- [22] K. SCHMÜDGEN, *The K -moment problem for compact semi-algebraic sets*, Math. Ann., 289 (1991), pp. 203–206.
- [23] M. SCHWEIGHOFER, *Optimization of polynomials on compact semialgebraic sets*, SIAM J. Optim., to appear.
- [24] H. D. SHERALI AND W. P. ADAMS, *A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems*, SIAM J. Discrete Math., 3 (1990), pp. 411–430.
- [25] H. D. SHERALI AND W. P. ADAMS, *A Reformulation-Linearization Technique for Solving Discrete and Continuous Nonconvex Problems*, Kluwer Academic Publishers, Boston, 1999.
- [26] F.-H. VASILESCU, *Spectral measures and moment problems*, in Spectral Theory and Its Applications, Theta Foundation, Bucharest, Romania, 2003, pp. 173–215.

ON THE CLASSICAL NECESSARY SECOND-ORDER OPTIMALITY CONDITIONS IN THE PRESENCE OF EQUALITY AND INEQUALITY CONSTRAINTS*

ABDELJELIL BACCARI[†] AND ABDELHAMID TRAD[‡]

Abstract. For nonconvex optimization problems in $\mathbf{R}^n$, a counterexample was recently given by Anitescu [*SIAM J. Optim.*, 10 (2000), pp. 1116–1135] that showed that the Mangasarian–Fromovitz constraint qualification (MFCQ) is not sufficient for the classical necessary second-order optimality conditions to hold. We prove these optimality conditions with the assumptions that the set of Lagrange multipliers is a bounded line segment and a relaxed strict complementary slackness (SCS) condition holds. A new constraint qualification is presented in this paper: We assume that the MFCQ holds and the active constraint rank deficiency is 1. This assumption relaxes the linear independence constraint qualification and enforces the MFCQ. In addition, the set of Lagrange multipliers is a bounded line segment.

Key words. nonconvex optimization, necessary conditions, constraint qualifications, pair of quadratic forms

AMS subject classifications. 49, 49K

DOI. 10.1137/S105262340342122X

1. Introduction. This paper is concerned with necessary second-order optimality conditions for nonconvex optimization problems of the form

$$(P) \quad \min f(x) \quad \text{such that} \quad g(x) \leq 0, \quad h(x) = 0.$$

The functions

$$f : \mathbf{R}^n \longrightarrow \mathbf{R}; \quad g : \mathbf{R}^n \longrightarrow \mathbf{R}^p; \quad h : \mathbf{R}^n \longrightarrow \mathbf{R}^q$$

are twice continuously differentiable. Problem (P) may have no equality or inequality constraints. We recall the main notations, definitions, and classical results that will be used in what follows.

The generalized Lagrangian function of (P) is defined on $\mathbf{R}^n \times \mathbf{R}_+^{p+1} \times \mathbf{R}^q$ by

$$\mathcal{L}(x, \lambda_0, \lambda, \mu) = \lambda_0 f(x) + \sum_{i=1}^{i=p} \lambda_i g_i(x) + \sum_{j=1}^{j=q} \mu_j h_j(x).$$

The Lagrangian function of (P) is defined on $\mathbf{R}^n \times \mathbf{R}_+^p \times \mathbf{R}^q$ by

$$L(x, \lambda, \mu) = \mathcal{L}(x, 1, \lambda, \mu).$$

The gradient and the Hessian matrix of $\mathcal{L}$ with respect to x are denoted, respectively, by $\nabla_x \mathcal{L}(x, \lambda, \mu)$ and $\nabla_{xx}^2 \mathcal{L}(x, \lambda, \mu)$. The gradient of f with respect to x is the column vector $\nabla f(x)$, and $\nabla f(x)^t$ is its transpose. The feasible set is

$$F = \{x \mid g(x) \leq 0, \quad h(x) = 0\}.$$

*Received by the editors January 13, 2003; accepted for publication (in revised form) March 11, 2004; published electronically December 30, 2004.

<http://www.siam.org/journals/siopt/15-2/42122.html>

[†]Ecole Supérieure des Sciences et Techniques de Tunis, 5 Avenue Taha Hussein, 1008 Tunis, Tunisia (Abdjil.Baccari@fst.rnu.tn).

[‡]Faculté des Sciences de Tunis, 1060 Tunis, Tunisia (abdelhamid.trad@fst.rnu.tn).

For $x \in F$, the active index set $I(x)$, the critical cone $C(x)$, the set of generalized Lagrange multipliers $\Lambda_0(x)$, and the set of Lagrange multipliers $\Lambda(x)$ are defined as follows:

$$I(x) = \{i \mid g_i(x) = 0\},$$

$$C(x) = \{d \mid \nabla f(x)^t d \leq 0, \nabla g_i(x)^t d \leq 0, i \in I(x), \nabla h_j(x)^t d = 0, j = 1, 2, \dots, q\},$$

$$\Lambda_0(x) = \{(\lambda_0, \lambda, \mu) \neq 0 \mid \nabla_x \mathcal{L}(x, \lambda_0, \lambda, \mu) = 0, (\lambda_0, \lambda) \in \mathbf{R}_+^{p+1}, \lambda_i g_i(x) = 0 \forall i\},$$

$$\Lambda(x) = \{(\lambda, \mu) \mid \lambda_0 = 1, (\lambda_0, \lambda, \mu) \in \Lambda_0(x)\}.$$

A quadratic form Q on $\mathbf{R}^n$ is defined by $Q(x) = B(x, x)$, where B is a bilinear and symmetric form on $\mathbf{R}^n$.

DEFINITION 1.1. Let Q be a quadratic form on $\mathbf{R}^n$, and let S be a subset of $\mathbf{R}^n$. Q is said to be positive semidefinite on S if

$$Q(s) \geq 0 \quad \forall s \in S.$$

This can be written as $Q \succeq 0$ on S .

DEFINITION 1.2. A nonempty subset $L \subset \mathbf{R}^m$ is a line segment if there exists $X \in \mathbf{R}^m$, $Y \in \mathbf{R}^m$, and an interval $J \subset \mathbf{R}$ such that

$$L = X + JY = \{l = X + \theta Y \mid \theta \in J\}.$$

Remark 1.1. If $Y \neq 0$ and L is closed and bounded, then J is closed and bounded and L can be written as $L = X + [a, b]Y$ for some real numbers $a \leq b$. Moreover, if L is not a singleton, then L has exactly two extreme points $X + aY$ and $X + bY$.

DEFINITION 1.3. A first-order cone K is a cone of the form $K = E + \mathbf{R}_+ d^0$, where E is a vector subspace in $\mathbf{R}^n$ and $d^0 \in \mathbf{R}^n$.

Note that every vector subspace $E \subset \mathbf{R}^n$ is a first-order cone. We recall the most classical constraint qualifications that will be used.

DEFINITION 1.4. A feasible point $x^* \in F$ is said to satisfy the strict complementary slackness (SCS) condition if $\Lambda(x^*)$ is not empty and, for every $i \in I(x^*)$, there exists $(\lambda, \mu) \in \Lambda(x^*)$ such that $\lambda_i > 0$.

Remark 1.2. If $x^* \in F$, the SCS condition holds, and p^* is the number of active inequality constraints, then the following hold:

- (i) For any $i \in I(x^*)$, there exists $(\lambda^i, \mu^i) \in \Lambda(x^*)$ such that $\lambda_i^i > 0$,

$$(\lambda^*, \mu^*) = \frac{1}{p^*} \sum_{i \in I(x^*)} (\lambda^i, \mu^i) \in \Lambda(x^*)$$

and satisfies $\lambda_i^* > 0$ for all $i \in I(x^*)$.

- (ii) The critical cone $C(x^*)$ is a vector space. To see this, let $d \in C(x^*)$ and $J = \{1, 2, \dots, q\}$; then

- (a) $\nabla_x L(x^*, \lambda^*, \mu^*) = 0 \implies \nabla_x L(x^*, \lambda^*, \mu^*)^t d = 0$,
- (b) $\nabla g_i(x^*)^t d \leq 0, \lambda_i^* > 0, \forall i \in I(x^*), \nabla h_j(x)^t d = 0, \forall j \in J$,
- (c) $\nabla f(x^*)^t d \leq 0 \implies \nabla g_i(x^*)^t d = 0, \forall i \in I(x^*)$.

So

$$C(x^*) = \{d \mid \nabla g_i(x^*)^t d = 0, i \in I(x^*), \nabla h_j(x^*)^t = 0, j \in J\}.$$

DEFINITION 1.5. A feasible point $x^* \in F$ is said to satisfy the linear independence constraint qualification (LICQ) if the vectors

$$\nabla g_i(x^*) \quad \forall i \in I(x^*), \quad \nabla h_j(x^*), \quad j = 1, 2, \dots, q,$$

are linearly independent.

The well-known Mangasarian–Fromovitz constraint qualification (MFCQ) is defined as follows [1].

DEFINITION 1.6. We say that the MFCQ holds at a feasible point $x^* \in F$, if the two following conditions hold:

- (i) The q vectors $\nabla h_j(x^*)$ are linearly independent, and
- (ii) there exists a vector d^* such that

$$\nabla g_i(x^*)^t d^* < 0 \quad \forall i \in I(x^*) ; \quad \nabla h_j(x^*)^t d^* = 0, \quad j = 1, 2, \dots, q.$$

The following lemma is well known (see, for instance, [2, p. 241]).

LEMMA 1.7. The MFCQ holds if and only if there is no $(\lambda, \mu) \neq 0$ such that

$$(1.1) \quad \lambda_i \geq 0 \quad \forall i \in I(x^*),$$

$$(1.2) \quad \sum_{i \in I(x^*)} \lambda_i \nabla g_i(x^*) + \sum_{j=1}^{j=q} \mu_j \nabla h_j(x^*) = 0.$$

Remark 1.3. Let $x^* \in F$ and satisfy one of the following conditions:

- (a) There are no active inequality constraints.
- (b) There is only one active inequality constraint.

Then, using (1.2), the MFCQ implies the LICQ.

DEFINITION 1.8. A feasible point $x^* \in F$ satisfies the sufficient second-order optimality conditions (SC2) if $\Lambda(x^*)$ is not empty and

$$(1.3) \quad \sup_{(\lambda, \mu) \in \Lambda(x^*)} (d)^t \nabla_{xx}^2 L(x^*, \lambda, \mu) d > 0 \quad \forall d \in C(x^*), \quad d \neq 0.$$

The following result is known as the Karush–Kuhn–Tucker necessary optimality conditions: Assume that x^* is a local optimal solution of (P) and satisfies the LICQ; then $\Lambda(x^*)$ is a singleton (i.e., $\Lambda(x^*) = \{(\lambda, \mu)\}$) and satisfies the following classical necessary second-order optimality conditions:

$$(CN2) \quad (d)^t \nabla_{xx}^2 L(x^*, \lambda, \mu) d \geq 0 \quad \forall d \in C(x^*).$$

(CN2) has the important property that the Lagrange multiplier (λ, μ) is the same for all critical vectors $d \in C(x^*)$. If the LICQ does not hold, $\Lambda(x^*)$ fails to be a singleton and (CN2) fails to be satisfied [4]. However, (CN2) holds if one of the following conditions is satisfied (see, for instance, [3, p. 211]; [5, p. 230]):

- (i) All constraint functions g and h are affine.
- (ii) The functions f and g are convex, h is affine, and x^* satisfies the Slater condition.

- (iii) There exists $(\lambda, \mu) \in \mathbf{R}_+^p \times \mathbf{R}^q$ such that (x^*, λ, μ) is a saddle point for the Lagrangian function of (P) .

Without any constraint qualification, a local optimal solution x^* of (P) satisfies the following Fritz John necessary first- and second-order optimality conditions (see, for example, [5, p. 443]):

$$(1.4) \quad \Lambda_0(x^*) \neq \emptyset,$$

$$(1.5) \quad \forall d \in C(x^*) \quad \exists (\lambda_0, \lambda, \mu) \in \Lambda_0(x^*) : (d)^t \nabla_{xx}^2 \mathcal{L}(x^*, \lambda_0, \lambda, \mu) d \geq 0.$$

The necessary second-order optimality condition (1.5) has two drawbacks: The first component λ_0 of λ , in (1.5), may vanish, and the multiplier $(\lambda_0, \lambda, \mu)$ in (1.5) is not necessarily the same for all critical vectors.

Assume that a local optimal solution x^* of (P) satisfies the MFCQ; then $\Lambda(x^*)$ is not empty and is bounded [6], convex, and compact. Every $(\lambda_0, \lambda, \mu) \in \Lambda_0(x^*)$ satisfies $\lambda_0 > 0$. The condition (1.5) can be written as

$$(GN2) \quad \max_{(\lambda, \mu) \in \Lambda(x^*)} (d)^t \nabla_{xx}^2 L(x^*, \lambda, \mu) d \geq 0 \quad \forall d \in C(x^*),$$

and (GN2) is free from the first drawback. For the second drawback, the first example was given in [4] and showed that the MFCQ does not imply (CN2). However, in each of the following cases, the MFCQ implies (CN2) [7]:

- (i) $n \leq 2$.
- (ii) There are, at most, two active inequality constraints.

A counterexample is given in [7] for $n = 3$ and three active inequality constraints.

This paper is a continuation of [7] and is devoted to the case in which the set of Lagrange multipliers is a convex hull of two extreme points (i.e., a bounded line segment). In this case the “max” in (GN2) can be written as the “max” of two quadratic forms. If the critical cone is a vector subspace or a first-order cone, then classical results on a pair of quadratic forms [2, 8, 9, 10, 11] are generalized and used. In section 2, two lemmas by Hestenes [2] are generalized to some closed cones (Lemma 2.1). In section 3, it is proved that Yuan’s lemma and its extension (Theorem 3.1) can be easily deduced from Hestenes’s lemmas. In section 5, we prove (CN2) for first-order cones included in the critical cone (Theorem 4.1). The main result of this paper (Theorem 5.1) is proved in section 6. It states that (CN2) holds if the SCS condition, or a weaker assumption (that only one index $i \in I(x^*)$ satisfies $\lambda_i = 0$ for all $(\lambda, \mu) \in \Lambda(x^*)$) is assumed. In section 7, the counterexample of [7] is used to show that (CN2) does not hold if $\Lambda(x^*)$ is bounded but is not a line segment. In section 8, a criterion for $\Lambda(x^*)$ to be a line segment is given (Lemma 7.2), a new constraint qualification (called the modified MFCQ) for $\Lambda(x^*)$ to be a bounded line segment is proposed, and Theorem 5.1 is rewritten (Theorem 7.7). In section 8, some remarks are made for further progress.

2. Extensions of Hestenes’s lemmas. Let P and Q be two quadratic forms on $\mathbf{R}^n$, and let K be a closed cone in $\mathbf{R}^n$. Consider the following statements:

$$(2.1) \quad d \in K, P(d) \leq 0, Q(d) \leq 0 \implies d = 0;$$

$$(2.2) \quad P \succeq 0 \quad \text{or} \quad Q \succeq 0 \quad \text{on } K;$$

$$(2.3) \quad \exists (d^1, d^2) \in K \times K : P(d^1) < 0, Q(d^2) < 0;$$

$$(2.4) \quad \exists \theta > 0 : P(d) + \theta Q(d) > 0 \quad \forall d \in K, d \neq 0.$$

Two lemmas by Hestenes [2, pp. 113–114] show that if K is a vector space, then

- (i) (2.1) and (2.2) imply (2.4), and
- (ii) (2.1) and (2.3) imply (2.4).

It can be seen that the statement (2.2) or (2.3) is always true. So, for a vector space K , (2.1) implies (2.4). We extend Hestenes's lemmas as follows.

LEMMA 2.1.

- (a) For every closed cone K in $\mathbf{R}^n$, (2.1) and (2.2) imply (2.4).
- (b) If K is a first-order cone, then (2.1) implies (2.4).

Proof. We prove (a): Assume first that $Q \succeq 0$ on K . We prove, by contradiction, that there exists a real number $c > 0$ such that, for every $\theta > c$, (2.4) holds. Suppose that, for every $c = n \in \mathbf{N}^*$, there exists $\theta_n > n$ and $d_n \in K - \{0\}$ such that $P(d_n) + \theta_n Q(d_n) \leq 0$. It follows that $P(d_n) \leq 0$, $y_n = d_n / \|d_n\| \in K$, $(y_n)_n$ has a subsequence $(y_{n_k})_k$ convergent to some $d \in K$, $\|d\| = 1$, $P(d) \leq 0$, $\theta_{n_k} \rightarrow +\infty$, and

$$P(d_{n_k}) + \theta_{n_k} Q(d_{n_k}) \leq 0 \implies Q(d) \leq 0.$$

From $Q(d) \leq 0$ and $P(d) \leq 0$, we get $d = 0$ and this contradicts $\|d\| = 1$. So c exists and any $\theta > c$ satisfies (2.4). A similar argument is used in the case $P \succeq 0$ on K .

We prove (b): From (a) it suffices to prove (b) in the case where (2.2) does not hold, that is, (2.3) holds. Let E be a vector space and $d^0 \in \mathbf{R}^n$ such that $K = E + \mathbf{R}_+ d^0$; then the vector subspace $E + \mathbf{R} d^0$ satisfies (2.1). To see this, suppose that $d = d_0 - r d^0$, with $r > 0$ and $d_0 \in E$, satisfies $P(d) \leq 0$ and $Q(d) \leq 0$. It follows that $P(-d) \leq 0$, $Q(-d) \leq 0$, $-d \in K$, and $-d = 0$. So we can suppose that $K = E + \mathbf{R} d^0$ and use the proof of [2, pp. 114–116]. To make the paper self-contained, we add this proof with some minor modifications.

For d, x, y in $\mathbf{R}^n$ and $\mu \in \mathbf{R}$, let $J(d, \mu) = P(d) + \mu Q(d)$, $I(x, y, \mu) = (1/2)(J(x + y, \mu) - J(x, \mu) - J(y, \mu))$, $BQ(x, y, \mu) = (1/2)(Q(x + y, \mu) - Q(x, \mu) - Q(y, \mu))$, and $S = \{d \in K \mid Q(d) \leq 0\}$. We apply (a) of Lemma 2.1 to the pair (P, Q) of quadratic forms on the closed cone S , and there exists a real number $\theta > 0$ such that

$$J(d, \theta) = P(d) + \theta Q(d) > 0 \quad \forall d \in S, d \neq 0.$$

Let

$$b = \sup \{\theta > 0 \mid J(d, \theta) > 0 \quad \forall d \in S, d \neq 0\}.$$

We have $Q(d^2) < 0$ and $d^2 \in S$. So b is finite and

$$J(d, b) \geq 0 \quad \forall d \in S.$$

The definition of b implies that, for every $n \in \mathbf{N}^*$, there exists $d_n \in S$, $d_n \neq 0$, such that $J(d_n, b + \frac{1}{n}) \leq 0$ and $(d_n / \|d_n\|)_n$ has a subsequence convergent to some $d^* \in S$ such that $\|d^*\| = 1$ and $J(d^*, b) \leq 0$. So, $J(d^*, b) = 0$. If $Q(d^*) = 0$ or $b = 0$, then $P(d^*) = 0$ and, using (2.1), $d^* = 0$, which is impossible. We get $b > 0$, $Q(d^*) < 0$, and $P(d^*) > 0$.

We claim that

$$J(d, b) \geq 0 \quad \forall d \in K; \quad I(d^*, d, b) = 0 \quad \forall d \in K.$$

Let $d \in K$. For $t \in \mathbf{R}$ and $|t|$ small enough, $Q(d^* + td) < 0$, $d^* + td \in S$, and $J(d^* + td, b) \geq 0$. The function f , defined by $f(t) = J(d^* + td, b) = J(d^*, b) + 2tI(d^*, d, b) + t^2J(d, b)$, has a local minimum at $t = 0$. So $f'(0) = 0 = 2I(d^*, d, b)$ and $f''(0) = 2J(d, b) \geq 0$.

We conclude that there exists $b > 0$ and $d^* \in K$ such that

- (1) $\|d^*\| = 1$, $J(d^*, b) = P(d^*) + bQ(d^*) = 0$, $P(d^*) > 0$, and $Q(d^*) < 0$.
- (2) $J(d, b) \geq 0$ for all $d \in K$ and $I(d^*, d, b) = 0$ for all $d \in K$.

In the same way, replacing Q by P in S , we find $c > 0$ and $d^{**} \in K$ such that

- (i) $\|d^{**}\| = 1$, $cP(d^{**}) + Q(d^{**}) = 0$, $P(d^{**}) < 0$, and $Q(d^{**}) > 0$.
- (ii) $cP(d) + Q(d) \geq 0$ for all $d \in K$ and $I(d^{**}, d, 1/c) = 0$ for all $d \in K$.

Let $a = 1/c$; then a and d^{**} satisfy

- (3) $\|d^{**}\| = 1$, $J(d^{**}, a) = P(d^{**}) + aQ(d^{**}) = 0$, $P(d^{**}) < 0$, and $Q(d^{**}) > 0$,
- (4) $J(d, a) \geq 0$ for all $d \in K$ and $I(d^{**}, d, a) = 0$ for all $d \in K$.

As $Q(d^*) < 0$ and $Q(d^{**}) > 0$, the equation in t , $Q(td^* + d^{**}) = t^2Q(d^*) + 2tBQ(d^*, d^{**}) + Q(d^{**}) = 0$, admits a real solution t_0 . So $d_0 = t_0d^* + d^{**} \in K - \{0\}$ satisfies $Q(d_0) = 0$ and $J(d_0, b) = P(d_0) > 0$. It follows that

- (iii) $J(d_0, b) = t_0^2J(d^*, b) + 2t_0I(d^*, d^{**}, b) + J(d^{**}, b) = J(d^{**}, b)$.
- (iv) $0 < P(d_0) = J(d_0, b) = P(d^{**}) + bQ(d^{**})$.

From $J(d^{**}, a) = 0 = P(d^{**}) + aQ(d^{**})$ and (iv), we get $(b - a)Q(d^{**}) > 0$ and $a < b$. Let $\theta \in]a, b[$ and $d \in K$; then

$$J(d, \theta) = J(d, a) + (\theta - a)Q(d) \geq (\theta - a)Q(d)$$

and

$$J(d, \theta) = J(d, b) + (\theta - b)Q(d) \geq -(b - \theta)Q(d).$$

This implies that $J(d, \theta) \geq 0$. If $J(d, \theta) = 0$, then $Q(d) = 0$, $P(d) = 0$, and $d = 0$. So

$$J(d, \theta) > 0 \quad \forall d \in K, d \neq 0. \quad \square$$

3. Extension of Yuan's lemma. Let K be a closed cone in $\mathbf{R}^n$, and let P and Q be two quadratic forms on $\mathbf{R}^n$. Consider the following statements:

$$(3.1) \quad \max(P(d), Q(d)) \geq 0 \quad \forall d \in K;$$

$$(3.2) \quad \exists (t_1, t_2) \in \mathbf{R}_+^2, t_1 + t_2 = 1 \text{ such that } t_1P(d) + t_2Q(d) \geq 0 \quad \forall d \in K;$$

$$(3.3) \quad \max(P(d), Q(d)) > 0 \quad \forall d \in K, d \neq 0;$$

$$(3.4) \quad \exists (t_1, t_2) \in \mathbf{R}_+^2, t_1 + t_2 = 1 \text{ such that } t_1P(d) + t_2Q(d) > 0 \quad \forall d \in K, d \neq 0.$$

It was shown in [8] that, for $K = \mathbf{R}^n$, (3.1) implies (3.2). In fact, these two statements are equivalent for $K = \mathbf{R}^n$. For the cone $K = \mathbf{R}_+^n$, (3.1) and (3.2) are not equivalent [9]. For three quadratic forms and $K = \mathbf{R}^n$, (3.1) and (3.2) are not equivalent [10]. A recent result in [11] shows that, for $n \geq 3$ and $K = \mathbf{R}^n$, (3.3) and (3.4) are equivalent. A little progress is made, in this section, by proving that (3.3) and (3.4) are equivalent for first-order cone $K = E + \mathbf{R}_+d^0 \subset \mathbf{R}^n$ with n unrestricted.

The main result of this section is as follows.

THEOREM 3.1. *For a first-order cone K , (3.3) and (3.4) are equivalent.*

Proof. We only prove that (3.3) implies (3.4). Inequality (3.3) is equivalent to the condition (2.1) and, from (b) of Lemma 2.1, (2.4) holds. So (3.4) is satisfied by $t_1 = 1/(1 + \theta)$ and $t_2 = \theta/(1 + \theta)$. $\square$

We get Yuan's lemma as a corollary.

COROLLARY 3.2. *For a first-order cone K , (3.1) and (3.2) are equivalent.*

Proof. To prove that (3.1) implies (3.2), let $n \in \mathbf{N}^*$. The quadratic forms $P_n(d) = P(d) + (1/n)\|d\|^2$ and $Q_n(d) = Q(d) + (1/n)\|d\|^2$ satisfy (3.3). By Theorem 3.1, there exists $t_1^n \geq 0$ and $t_2^n \geq 0$ such that $t_1^n + t_2^n = 1$ and

$$t_1^n P_n(d) + t_2^n Q_n(d) = t_1^n P(d) + t_2^n Q(d) + (1/n)\|d\|^2 > 0 \quad \forall d \in K, d \neq 0.$$

$(t_1^n, t_2^n)_n$ has a subsequence $(t_1^{n_k}, t_2^{n_k})_k$ convergent to (t_1, t_2) , $t_1 \geq 0$, $t_2 \geq 0$, $t_1 + t_2 = 1$ and, for every fixed $d \in K$, $d \neq 0$,

$$t_1^n P(d) + t_2^n Q(d) + (1/n)\|d\|^2 > 0 \implies t_1 P(d) + t_2 Q(d) \geq 0. \quad \square$$

4. Abstract main result. Theorem 3.1 and its corollary are used to prove the following abstract main result.

THEOREM 4.1. *Let x^* be a local optimal solution of (P) and such that $\Lambda(x^*)$ is a bounded line segment; then, for every first-order cone K included in the critical cone $C(x^*)$, there exists $(\lambda^K, \mu^K) \in \Lambda(x^*)$ such that*

$$(4.1) \quad (d)^t \nabla_{xx}^2 L(x^*, \lambda^K, \mu^K) d \geq 0 \quad \forall d \in K.$$

Moreover, if x^* satisfies the sufficient SC2, then (λ^K, μ^K) can be chosen so that

$$(4.2) \quad (d)^t \nabla_{xx}^2 L(x^*, \lambda^K, \mu^K) d > 0 \quad \forall d \in K, d \neq 0.$$

Proof. We begin with (4.1): x^* is a local optimal solution of (P) and satisfies (MFCQ), and $(GN2)$ holds. If $\Lambda(x^*)$ is a singleton, then $(GN2)$ implies (4.1). Suppose that the closed and bounded line segment $\Lambda(x^*)$ is not a singleton; then $\Lambda(x^*)$ has exactly two extreme points $(\lambda^1, \mu^1) \in \Lambda(x^*)$ and $(\lambda^2, \mu^2) \in \Lambda(x^*)$. Let

$$P(d) = (d)^t \nabla_{xx}^2 L(x^*, \lambda^1, \mu^1) d; \quad Q(d) = (d)^t \nabla_{xx}^2 L(x^*, \lambda^2, \mu^2) d.$$

$(d)^t \nabla^2 L(x^*, \lambda, \mu) d$ is linear in (λ, μ) and the “max” in $(GN2)$ is attained at an extreme point. So we have

$$0 \leq \max_{(\lambda, \mu) \in \Lambda(x^*)} (d)^t \nabla_{xx}^2 L(x^*, \lambda, \mu) d = \max(P(d), Q(d)) \quad \forall d \in K,$$

and (3.1) holds. From Corollary 3.2, (3.1) implies (3.2); that is, there exist $t_1 \geq 0$ and $t_2 \geq 0$ such that $t_1 + t_2 = 1$,

$$t_1 P(d) + t_2 Q(d) \geq 0 \quad \forall d \in K,$$

$$(\lambda^K, \mu^K) = t_1(\lambda^1, \mu^1) + t_2(\lambda^2, \mu^2) \in \Lambda(x^*),$$

$$t_1 P(d) + t_2 Q(d) = (d)^t \nabla_{xx}^2 L(x^*, \lambda^K, \mu^K) d \geq 0 \quad \forall d \in K.$$

To prove (4.2), we use Theorem 3.1 and the same above arguments. $\square$

5. Main result. Classical necessary second-order optimality conditions (CN2), without the LICQ, convexity assumptions, or SCS condition are not easy to get. A first step for doing this is to consider optimization problems in which the critical cone is a first-order cone and the set of Lagrange multipliers is a bounded line segment.

The main result of this paper is as follows.

THEOREM 5.1. *Let x^* be a local optimal solution of (P) and be such that the following hold:*

- (i) *The set of Lagrange multipliers $\Lambda(x^*)$ is a bounded line segment.*
- (ii) *There exists, at most, only one index $i_0 \in I(x^*)$ such that*

$$(5.1) \quad (\lambda, \mu) \in \Lambda(x^*) \implies \lambda_{i_0} = 0;$$

then there exists a Lagrange multiplier $(\lambda^*, \mu^*) \in \Lambda(x^*)$ such that

$$(d)^t \nabla_{xx}^2 L(x^*, \lambda^*, \mu^*) d \geq 0 \quad \forall d \in C(x^*).$$

Moreover, if x^* satisfies the sufficient second-order optimality condition (SC2), then (λ^*, μ^*) can be chosen so that

$$(d)^t \nabla_{xx}^2 L(x^*, \lambda^*, \mu^*) d > 0 \quad \forall d \in C(x^*), d \neq 0.$$

Proof. We prove first that $C(x^*)$ is a first-order cone. We have the following two cases:

(1) There is no $i_0 \in I(x^*)$ such that (5.1) holds. This means that x^* satisfies the SCS condition and $C(x^*)$ is a vector subspace (see Remark 1.2).

(2) There exists only one index $i_0 \in I(x^*)$ such that (5.1) holds. It follows that $d \in C(x^*)$ if and only if

$$\nabla g_i(x^*)^t d = 0 \quad \forall i \in I(x^*) - \{i_0\}, \quad \nabla g_{i_0}(x^*)^t d \leq 0,$$

$$\nabla h_j(d)^t d = 0, \quad j = 1, 2, \dots, q.$$

If $C(x^*)$ is not a subspace, there exists $d^0 \in C(x^*)$ such that $\nabla g_{i_0}(x^*)^t d^0 < 0$ and, for every other $d \in C(x^*)$, there exists $r \geq 0$ such that

$$\nabla g_{i_0}(x^*)^t d = r \nabla g_{i_0}(x^*)^t d^0, \quad \nabla g_{i_0}(x^*)^t (d - r d^0) = 0,$$

$$\nabla g_i(x^*)^t (d - r d^0) = 0 \quad \forall i \in I(x^*) - \{i_0\}; \quad \nabla h_j(x^*)^t (d - r d^0) = 0, \quad j = 1, 2, \dots, q,$$

and $w = d - r d^0$ is in the vector space

$$E = \{w | \nabla g_i(x^*)^t w = 0, \quad i \in I(x^*); \quad \nabla h_j(x^*)^t w = 0, \quad j = 1, 2, \dots, q\} \subset C(x^*).$$

So $C(x^*) = E + \mathbf{R}_+ d^0$.

We conclude that (ii) implies that $C(x^*)$ is a first-order cone and, from Theorem 4.1, we have the desired result. $\square$

Remark 5.1.

(i) For Theorem 5.1, a direct proof, with Hestenes's lemmas, is possible, but the use of Yuan's extended lemma (Theorem 3.1) and the abstract main result (Theorem 4.1) make this proof more clear.

(ii) If (5.1) holds for many active indexes $i \in I(x^*)$, then the critical cone $C(x^*)$ fails to be a first-order cone and Theorem 4.1 cannot be used to prove Theorem 5.1. To see this fact, assume, for example, that $1 \in I(x^*)$ and $2 \in I(x^*)$ are the only active indexes which satisfy (5.1), and suppose that there exists $d^1 \in C(x^*)$ and $d^2 \in C(x^*)$ such that

$$\nabla g_1(x^*)^t d^1 < 0, \quad \nabla g_2(x^*)^t d^1 = 0,$$

$$\nabla g_2(x^*)^t d^2 < 0, \quad \nabla g_1(x^*)^t d^2 = 0.$$

For any $d \in C(x^*)$, we can find $r_1 \geq 0$ and $r_2 \geq 0$ such that

$$\nabla g_1(x^*)^t (d - r_1 d^1 - r_2 d^2) = \nabla g_1(x^*)^t (d - r_1 d^1) = \nabla g_1(x^*)^t (d) - r_1 \nabla g_1(x^*)^t (d^1) = 0,$$

$$\nabla g_2(x^*)^t (d - r_1 d^1 - r_2 d^2) = \nabla g_2(x^*)^t (d - r_2 d^2) = \nabla g_2(x^*)^t (d) - r_2 \nabla g_2(x^*)^t (d^2) = 0.$$

This means that $w = d - r_1 d^1 - r_2 d^2$ is in the vector space

$$E = \{w \mid \nabla g_i(x^*)^t w = 0, \quad i \in I(x^*); \quad \nabla h_j(x^*)^t w = 0, \quad j = 1, 2, \dots, q\} \subset C(x^*),$$

$$C(x^*) = E + \mathbf{R}_+ d^1 + \mathbf{R}_+ d^2,$$

and $C(x^*)$ is not a first-order cone.

Example 5.1. The nonconvex optimization problem, in $\mathbf{R}^4$, is

$$\min\{-x_1 \mid g_i(x) \leq 0, \quad i = 1, 2, 3\},$$

where

$$g_1(x) = 2x_1^2 - x_2^2 + 2x_3^2 + x_1, \quad g_2(x) = -x_2^2 + x_1 - 2x_4, \quad g_3(x) = -x_1^2 + x_2^2 - x_3^2 + x_1 + x_4$$

for a feasible point x , $g_1(x) + g_2(x) + 2g_3(x) = 4x_1 \leq 0$. It can be seen that $x^* = (0, 0, 0, 0)$ is a global optimal solution,

$$\Lambda(x^*) = \{(\lambda_1, \lambda_2, \lambda_3) \geq 0 \mid \lambda_1 + \lambda_2 + \lambda_3 = 1; \quad 2\lambda_2 = \lambda_3\}$$

is a bounded line segment, x^* satisfies the SCS condition, and $C(x^*) = \{d = (d_1, d_2, d_3, d_4)^t \mid d_1 = d_4 = 0\}$. $\lambda = (1/4, 1/4, 1/2)$ is the unique multiplier which satisfies (CN2).

A simple criterion for $\Lambda(x^*)$ to be a bounded line segment is given in the following lemma.

LEMMA 5.2. *Let x^* be a local optimal solution of (P) , satisfy the MFCQ, and be such that the number of active inequality constraints is at most two; then $\Lambda(x^*)$ is a bounded line segment.*

Proof. Let p_{x^*} be the number of active inequality constraints. If $p_{x^*} \leq 1$, then the LICQ holds and $\Lambda(x^*)$ is a singleton. If $p_{x^*} = 2$, we can suppose, without loss of generality, that $I(x^*) = \{1, 2\}$.

Assume that one of the following conditions holds:

- (i) There exists $(\lambda^0, \mu^0) \in \Lambda(x^*)$ such that $\lambda_1^0 = 0 = \lambda_2^0$.
- (ii) All $(\lambda, \mu) \in \Lambda(x^*)$ satisfy $\lambda_i > 0$, $i = 1, 2$.
- (iii) All $(\lambda, \mu) \in \Lambda(x^*)$ satisfy $\lambda_1 = 0$.
- (iv) All $(\lambda, \mu) \in \Lambda(x^*)$ satisfy $\lambda_2 = 0$.

We claim that $\Lambda(x^*)$ is a singleton: Suppose that (i) holds; then

$$\nabla f(x^*) + \sum_{j=1}^{j=q} \mu_j^0 \nabla h_j(x^*) = 0$$

and for any $(\lambda, \mu) \in \Lambda(x^*)$ we have

$$\nabla f(x^*) + \lambda_1 \nabla g_1(x^*) + \lambda_2 \nabla g_2(x^*) + \sum_{j=1}^{j=q} \mu_j \nabla h_j(x^*) = 0.$$

The difference between these two equations and Lemma 1.7 yield

$$\lambda_1 = 0 = \lambda_2, \quad \mu = \mu^0.$$

Similar arguments are used for other cases.

So, if $\Lambda(x^*)$ is not a singleton, there exists $(\lambda^1, \mu^1) \in \Lambda(x^*)$ and $(\lambda^2, \mu^2) \in \Lambda(x^*)$ such that

$$\lambda_1^1 > 0, \quad \lambda_2^1 = 0; \quad \lambda_1^2 = 0, \quad \lambda_2^2 > 0,$$

$$(a) \quad \nabla f(x^*) + \lambda_1^1 \nabla g_1(x^*) + \sum_{j=1}^{j=q} \mu_j^1 \nabla h_j(x^*) = 0,$$

$$(b) \quad \nabla f(x^*) + \lambda_2^2 \nabla g_2(x^*) + \sum_{j=1}^{j=q} \mu_j^2 \nabla h_j(x^*) = 0.$$

x^* satisfies the MFCQ, the vectors of

$$\{\nabla g_1(x^*), \nabla h_j(x^*), j = 1, 2, \dots, q\}$$

are linearly independent, and (λ^1, μ^1) is unique. Also, (λ^2, μ^2) is unique. Let $(\lambda, \mu) \in \Lambda(x^*)$; then

$$(c) \quad \nabla f(x^*) + \lambda_1 \nabla g_1(x^*) + \lambda_2 \nabla g_2(x^*) + \sum_{j=1}^{j=q} \mu_j \nabla h_j(x^*) = 0.$$

If $\lambda_1 = 0$ or $\lambda_2 = 0$, then $(\lambda, \mu) = (\lambda^1, \mu^1)$ or $(\lambda, \mu) = (\lambda^2, \mu^2)$. Suppose that $\lambda_1 > 0$ and $\lambda_2 > 0$. For $t = \lambda_1 / \lambda_1^1$, (c) - $t(a)$ gives

$$(d) \quad (1-t) \nabla f(x^*) + \lambda_2 \nabla g_2(x^*) + \sum_{j=1}^{j=q} (\mu_j - t \mu_j^1) \nabla h_j(x^*) = 0.$$

If $t = 1$, then

$$\lambda_2 \nabla g_2(x^*) + \sum_{j=1}^{j=q} (\mu_j - \mu_j^1) \nabla h_j(x^*) = 0,$$

$\lambda_2 = 0$, and this is impossible. If $t > 1$, then (d) can be written as

$$(e) \quad \nabla f(x^*) + (\lambda_2/(1-t))\nabla g_2(x^*) + \sum_{j=1}^{j=q} (\mu_j - t\mu_j^1)/(1-t)\nabla h_j(x^*) = 0.$$

$\lambda_2^2 = \lambda_2/(1-t) < 0$, and this is impossible, too. So $0 < t < 1$, $\lambda_2^2 = \lambda_2/(1-t)$, $\mu_j^2 = (\mu_j - t\mu_j^1)/(1-t)$, $(\lambda^2, \mu^2) = [(\lambda, \mu) - t(\lambda^1, \mu^1)]/(1-t)$, and

$$(\lambda, \mu) = t(\lambda^1, \mu^1) + (1-t)(\lambda^2, \mu^2).$$

This means that every $(\lambda, \mu) \in \Lambda(x^*)$ is a convex combination of (λ^1, μ^1) and (λ^2, μ^2) and $\Lambda(x^*)$ is a bounded line segment. $\square$

Remark 5.2. We list some cases in which Theorem 5.1 holds without (ii):

- (i) The working space is $\mathbf{R}$ or $\mathbf{R}^2$ [7].
- (ii) The number of active inequality constraints is, at most, two [7].
- (iii) The critical cone is included in a one-dimensional vector space (use (GN2)).

6. Theorem 5.1 does not hold without line segment assumption. The counterexample in [7] can be used to show that condition (i) of Theorem 5.1 cannot be replaced by the MFCQ.

The optimization problem, in $\mathbf{R}^3$, is as follows.

Example 6.1.

$$\min \{x_3 \mid g_i(x) \leq 0, i = 1, 2, 3\},$$

where

$$g_1(x) = 2\sqrt{3}x_1x_2 - 2x_2^2 - x_3, \quad g_2(x) = x_2^2 - 3x_1^2 - x_3, \quad g_3(x) = -2\sqrt{3}x_1x_2 - 2x_2^2 - x_3.$$

It can be seen that $x^* = (0, 0, 0)$ is a global optimal solution and satisfies the MFCQ. The set of Lagrange multipliers

$$\Lambda(x^*) = \{\lambda \in \mathbf{R}^3 \mid \lambda_i \geq 0, i = 1, 2, 3; \lambda_1 + \lambda_2 + \lambda_3 = 1\}$$

is not a line segment. The critical cone

$$C(x^*) = \{d = (d_1, d_2, d_3)^t \mid d_3 = 0\}$$

is a vector space. For the critical vectors $d^1 = (1, 0, 0)^t$ and $d^2 = (0, 1, 0)^t$, there is no $\lambda \in \Lambda(x^*)$ such that (see [7])

$$(d^1)^t \nabla_{xx}^2 L(x^*, \lambda) d^1 \geq 0 \quad \text{and} \quad (d^2)^t \nabla_{xx}^2 L(x^*, \lambda) d^2 \geq 0.$$

7. A new constraint qualification. Let x^* be a feasible point. Without loss of generality, we suppose that there exists an integer $p^* \leq p$ such that $g_i(x^*) = 0$, $i = 1, 2, \dots, p^*$, and $g_i(x^*) < 0$, $i = p^* + 1 \dots p$.

DEFINITION 7.1. We say that a feasible point x^* satisfies the rank condition (RC) if the vectors of

$$\{\nabla g_i(x^*), i = 1, 2, \dots, p^*; \nabla h_j(x^*), j = 1, 2, \dots, q\}$$

have $p^* + q - 1$ linearly independent vectors.

LEMMA 7.2. Let x^* be a feasible point such that

(i) $\Lambda(x^*)$ is not empty, and
(ii) x^* satisfies the RC;
then $\Lambda(x^*)$ is a closed line segment.

Proof. It suffices to prove the lemma in the case $\Lambda(x^*)$ is not a singleton and $p^* = p$. Let $(\lambda^*, \mu^*) \in \Lambda(x^*)$; then

$$(7.1) \quad \nabla f(x^*) + \sum_{i=1}^{i=p} \lambda_i^* \nabla g_i(x^*) + \sum_{j=1}^{j=q} \mu_j^* \nabla h_j(x^*) = 0.$$

For any other $(\lambda, \mu) \in \Lambda(x^*)$, we have

$$(7.2) \quad \nabla f(x^*) + \sum_{i=1}^{i=p} \lambda_i \nabla g_i(x^*) + \sum_{j=1}^{j=q} \mu_j \nabla h_j(x^*) = 0;$$

$$(7.3) \quad \sum_{i=1}^{i=p} (\lambda_i - \lambda_i^*) \nabla g_i(x) + \sum_{j=1}^{j=q} (\mu_j - \mu_j^*) \nabla h_j(x) = 0.$$

Equation (7.3) means that the Jacobian matrix, $Dc(x^*)$, of active constraints

$$c = (g_1, g_2, \dots, g_p, h_1, h_2, \dots, h_q)$$

satisfies

$$Dc(x^*)^t((\lambda, \mu) - (\lambda^*, \mu^*)) = 0; (\lambda, \mu) - (\lambda^*, \mu^*) \in \text{Ker}(Dc(x^*)^t),$$

where $Dc(x^*)^t$ is the transpose of the matrix $Dc(x^*)$. It is well known that

$$\text{Ker}(Dc(x^*)^t) = (R(Dc(x^*)))^\perp,$$

where $(R(Dc(x^*)))^\perp$ is the orthogonal space to the range of $Dc(x^*)$. The dimension of $R(Dc(x^*))$ is $p^* + q - 1$, $(R(Dc(x^*)))^\perp$ is a one-dimensional space, and there exists $z \in \mathbf{R}^{p+q}$, $z \neq 0$, such that

$$\text{Ker}(Dc(x^*)^t) = (R(Dc(x^*)))^\perp = \mathbf{R}z$$

and any $(\lambda, \mu) \in \Lambda(x^*)$ satisfies $(\lambda, \mu) - (\lambda^*, \mu^*) \in \mathbf{R}z$, so

$$\Lambda(x^*) \subset (\lambda^*, \mu^*) + \mathbf{R}z.$$

Let $J = \{\alpha \in \mathbf{R} \mid (\lambda^*, \mu^*) + \alpha z \in \Lambda(x^*)\}$. $\Lambda(x^*)$ is closed and convex and J is closed and convex, too, so it is a closed interval and $\Lambda(x^*) = (\lambda^*, \mu^*) + Jz$ is a closed line segment. $\square$

LEMMA 7.3. Let x^* be a feasible point for (P) and such that

- (i) $\Lambda(x^*)$ is not empty and bounded, and
- (ii) x^* satisfies the RC;

then there exist $b \geq 0$, $z \in \mathbf{R}^{p+q}$, and an extreme point $(\lambda^*, \mu^*) \in \Lambda(x^*)$ such that

$$\Lambda(x^*) = (\lambda^*, \mu^*) + [0, b]z.$$

Proof. $\Lambda(x^*)$ is convex and compact and has at least one extreme point (λ^*, μ^*) . It can be seen, from Lemma 7.2 and Remark 1.1, that $\Lambda(x^*) = (\lambda^*, \mu^*) + [a, b]z$ for

some vector $z \in \mathbf{R}^{p+q}$ and some real numbers $a \leq b$. Supposing that $a < b$, we claim that $0 \notin]a, b[$: Otherwise, for $\varepsilon \in]0, b[$, $\varepsilon < -a$, we have

$$(\lambda^1, \mu^1) = (\lambda^*, \mu^*) + \varepsilon z \in \Lambda(x^*); \quad (\lambda^2, \mu^2) = (\lambda^*, \mu^*) - \varepsilon z \in \Lambda(x^*),$$

and

$$(\lambda^*, \mu^*) = [(\lambda^1, \mu^1) + (\lambda^2, \mu^2)]/2$$

contradicts the fact that (λ^*, μ^*) is an extreme point of $\Lambda(x^*)$. It follows that $0 \notin]a, b[$. $(\lambda^*, \mu^*) \in \Lambda(x^*)$ and $a = 0$ or $b = 0$. Replacing z by $-z$, if necessary, we can assume that $a = 0$. $\square$

DEFINITION 7.4. *We say that a feasible point x^* satisfies the generalized strict complementary slackness (GSCS) condition if there exists, at most, only one index $i_0 \in I(x^*)$ such that*

$$(\lambda, \mu) \in \Lambda(x^*) \implies \lambda_{i_0} = 0.$$

DEFINITION 7.5. *We say that a feasible point x^* satisfies the modified MFCQ (MMF) if x^* satisfies the RC and the MFCQ.*

Remark 7.1. It will be helpful to have some equivalent forms of the MMF. Note that the following hold.

- (i) x^* satisfies the LICQ if and only if there is no $(\lambda, \mu) \neq 0$ such that (1.2) holds.
- (ii) x^* satisfies the MFCQ if and only if there is no $(\lambda, \mu) \neq 0$ such that (1.1) and (1.2) hold.
- (iii) We look for a criterion, like (1.1) or (1.2), to characterize the MMF.

Let $x^* \in F$ and $\{i_1, i_2\} \subset I(x^*)$; then x^* is a feasible point for the optimization problem

$$(P_2) \quad \min\{f(x) \mid g_i(x) \leq 0, i = i_1, i_2, g_i(x) = 0, i \in I(x), i \neq i_1, i_2, h(x) = 0\}.$$

In (P_2) , all active inequalities of (P) , except i_1 and i_2 , are converted into equalities. The following lemma shows that x^* satisfies the MMF if and only if there exists $\{i_1, i_2\} \subset I(x^*)$ such that x^* satisfies the MFCQ for (P_2) .

LEMMA 7.6. *For a feasible point x^* , with more than one active inequality constraint, the following conditions are equivalent:*

- (i) x^* satisfies the MMF.
- (ii) *There exist i_1 and i_2 in $I(x^*)$ such that there is no $(\lambda, \mu) \neq 0$ such that*

$$(7.4) \quad \lambda_{i_1} \geq 0, \lambda_{i_2} \geq 0,$$

and

$$(7.5) \quad \sum_{i \in I(x^*)} \lambda_i \nabla g_i(x^*) + \sum_{j=1}^{j=q} \mu_j \nabla h_j(x^*) = 0.$$

- (iii) *There exist i_1 and i_2 in $I(x^*)$ such that the vectors*

$$\nabla g_i(x^*), i \in I(x^*) - \{i_1, i_2\}, \nabla h_j(x^*) \quad \forall j,$$

are linearly independent and there is a vector d^* such that

$$(7.6) \quad \nabla g_{i_1}(x^*)^t d^* < 0, \nabla g_{i_2}(x^*)^t d^* < 0,$$

$$(7.7) \quad \nabla g_{i_1}(x^*)^t d^* = 0, i \in I(x^*) - \{i_1, i_2\}, \nabla h_j(x^*)^t d^* = 0 \quad \forall j.$$

Proof. We begin with (i) implies (ii): If the vectors of

$$C = \{\nabla g_i(x^*), i \in I(x^*), \nabla h_j(x^*), j = 1, 2, \dots, q\}$$

are linearly independent, then any i_1 and i_2 in $I(x^*)$ satisfy (ii). Suppose that the vectors of C are not linearly independent; then there exists $i_1 \in I(x^*)$ such that

$$\nabla g_{i_1}(x^*) = \sum_{i \in I(x^*) - \{i_1\}} \lambda_i^* \nabla g_i(x^*) + \sum_{j=1}^{j=q} \mu_j^* \nabla h_j(x^*).$$

If all $i \in I(x^*) - \{i_1\}$ satisfy $\lambda_i^* \leq 0$, then

$$\nabla g_{i_1}(x^*) - \sum_{i \in I(x^*) - \{i_1\}} \lambda_i^* \nabla g_i(x^*) - \sum_{j=1}^{j=q} \mu_j^* \nabla h_j(x^*) = 0.$$

Because of the MFCQ and Lemma 1.7, we have

$$\lambda_i^* = 0 \quad \forall i \in I(x^*); \quad \mu_j^* = 0, j = 1, 2, \dots, q,$$

and this contradicts $\lambda_{i_1}^* = 1$. So there exists $i_2 \in I(x^*) - \{i_1\}$ such that $\lambda_{i_2}^* > 0$. It can be seen, from the RC, that the vectors of

$$B = \{\nabla g_i(x^*), i \in I(x^*) - \{i_1\}, \nabla h_j(x^*), j = 1, 2, \dots, q\}$$

are linearly independent and the components (λ^*, μ^*) of $\nabla g_{i_1}(x^*)$ with respect to B are unique. Suppose that (λ, μ) satisfies (7.4) and (7.5). If $\lambda_{i_1} > 0$, then $-\lambda_{i_2}/\lambda_{i_1} \leq 0$ and must be equal to $\lambda_{i_2}^* > 0$; this is impossible, so $\lambda_{i_1} = 0$ and all other components of (λ, μ) vanish.

We prove that (ii) implies (iii): by Lemma 1.7, (ii) means that, for (P_2) , x^* satisfies the MFCQ which is (iii). So (ii) and (iii) are equivalent. To prove that (ii) implies (i), note that, by Lemma 1.7, x^* satisfies the MFCQ for the problem (P) . From (7.5), the vectors of

$$B = \{\nabla g_i(x^*), i \in I(x^*) - \{i_1\}, \nabla h_j(x^*), j = 1, 2, \dots, q\}$$

are linearly independent and the RC holds. $\square$

The main result (Theorem 5.1) of this paper can be written, with the MMF and GSCS condition, in the following form.

THEOREM 7.7. *Let x^* be a local optimal solution of (P) and be such that*

- (i) x^* satisfies the MMF, and
- (ii) x^* satisfies the GSCS condition;

then there exists $(\lambda^*, \mu^*) \in \Lambda(x^*)$ such that

$$(d)^t \nabla_{xx}^2 L(x^*, \lambda^*, \mu^*) d \geq 0 \quad \forall d \in C(x^*).$$

Moreover, if x^* satisfies the sufficient SC2, then (λ^*, μ^*) can be chosen so that

$$(d)^t \nabla_{xx}^2 L(x^*, \lambda^*, \mu^*) d > 0 \quad \forall d \in C(x^*), d \neq 0.$$

8. Conclusion. We have proved the classical necessary second-order optimality conditions (CN2) without using the LICQ. Our result is concerned with nonconvex optimization problems in which the set of Lagrange multipliers $\Lambda(x^*)$, at the local optimal solution x^* , is a bounded line segment and the GSCS condition holds. The GSCS condition is not necessary for some special structures of the critical cone. A condition for $\Lambda(x^*)$ to be a bounded line segment is given. If the LICQ is assumed, then (CN2) holds without the SCS or GSCS condition. Our result is limited by the GSCS assumption which seems not to be necessary, but we have no counterexample.

REFERENCES

- [1] O. L. MANGASARIAN AND S. FROMOVITZ, *The Fritz John necessary optimality conditions in the presence of equality and inequality constraints*, J. Math. Anal. Appl., 17 (1967), pp. 37–47.
- [2] M. R. HESTENES, *Optimization Theory: The Finite Dimensional Case*, Robert E. Krieger Publishing Company, Huntington, NY, 1975.
- [3] M. S. BAZARAA, H. D. SHERALI, AND C. M. SHETTY, *Nonlinear Programming: Theory and Algorithms*, John Wiley, New York, 1993.
- [4] M. ANITESCU, *Degenerate nonlinear programming with a quadratic growth condition*, SIAM J. Optim., 10 (2000), pp. 1116–1135.
- [5] J. F. BONNANS AND A. SHAPIRO, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, Berlin, 2000.
- [6] J. GAUVIN, *A necessary and sufficient regularity condition to have bounded multipliers in nonconvex programming*, Math. Program., 12 (1977), pp. 136–138.
- [7] A. BACCARI, *On the classical necessary second-order optimality conditions*, J. Optim. Theory Appl., 123 (2004), pp. 213–221.
- [8] Y. YUAN, *On a subproblem of trust region algorithms for constrained optimization*, Math. Program., 47 (1990), pp. 53–63.
- [9] J.-P. CROUZEIX, J.-E. MARTINEZ-LEGAZ, AND A. SEEGER, *Un théorème de l’alternative pour des formes quadratiques et quelques extensions*, C. R. Acad. Sci. Paris Sér. I Math., 314 (1992), pp. 505–506.
- [10] J.-E. MARTINEZ-LEGAZ AND A. SEEGER, *Yuan’s alternative theorem and the maximization of the minimum eigenvalue function*, J. Optim. Theory Appl., 82 (1994), pp. 159–167.
- [11] J.-B. HIRIART-URRUTY AND M. TORKI, *Permanently going back and forth between the “quadratic world” and the “convexity world” in optimization*, Appl. Math. Optim., 45 (2002), pp. 169–184.

A CLASS OF ACTIVE-SET NEWTON METHODS FOR MIXED COMPLEMENTARITY PROBLEMS*

A. N. DARYINA[†], A. F. IZMAILOV[‡], AND M. V. SOLODOV[§]

Abstract. Based on the identification of indices active at a solution of the mixed complementarity problem (MCP), we propose a class of Newton methods for which local superlinear convergence holds under extremely mild assumptions. In particular, the error bound condition needed for the identification procedure and the nondegeneracy condition needed for the convergence of the resulting Newton method are individually and collectively strictly weaker than the property of semistability of a solution. Thus the local superlinear convergence conditions of the presented method are weaker than conditions required for the semismooth (generalized) Newton methods applied to MCP reformulations. Moreover, they are also weaker than convergence conditions of the linearization (Joseph–Newton) method. For the special case of optimality systems with primal-dual structure, we further consider the question of superlinear convergence of primal variables. We illustrate our theoretical results with numerical experiments on some specially constructed MCPs whose solutions do not satisfy the usual regularity assumptions.

Key words. mixed complementarity problem, semistability, 2-regularity, weak regularity, error bound, Newton method

AMS subject classifications. 90C30, 65K05

DOI. 10.1137/S105262340343590X

1. Introduction. The *mixed complementarity problem* (MCP) [8] is the variational inequality on a generalized box; that is,

$$(1.1) \quad \text{find } x \in B \text{ such that } \langle F(x), y - x \rangle \geq 0 \quad \forall y \in B,$$

where $F : \mathbf{R}^n \rightarrow \mathbf{R}^n$ and

$$B = \{x \in \mathbf{R}^n \mid l_i \leq x_i \leq u_i, \quad i = 1, \dots, n\},$$

$l_i \in \mathbf{R} \cup \{-\infty\}$, $u_i \in \mathbf{R} \cup \{+\infty\}$, $l_i < u_i \quad \forall i = 1, \dots, n$. Equivalently, this can be stated as

$$\text{find } x \in B \text{ such that } F_i(x) \begin{cases} \geq 0 & \text{if } x_i = l_i, \\ = 0 & \text{if } x_i \in (l_i, u_i), \\ \leq 0 & \text{if } x_i = u_i. \end{cases} \quad i = 1, \dots, n.$$

As is well known, many important problems can be cast in the format of MCP [10, 8]. As a special case of MCP, we mention the nonlinear complementarity problem (NCP), which corresponds to setting $l_i = 0$, $u_i = +\infty$, $i = 1, \dots, n$. The systems of nonlinear

*Received by the editors October 3, 2003; accepted for publication (in revised form) June 10, 2004; published electronically December 30, 2004. The research of the first two authors was supported by Russian Foundation for Basic Research grant 01-01-00810 and Council for the State Support of Leading Scientific Schools grant 00-15-96141. The third author was supported in part by CNPq grants 300734/95-6(RN) and 471780/2003-0, by PRONEX-Optimization, and by FAPERJ.

<http://www.siam.org/journals/siopt/15-2/43590.html>

[†]Russian Peoples Friendship University, Miklukho-Maklaya Str. 6, 117198 Moscow, Russia (anna.daryina@mail.ru).

[‡]Moscow State University, Faculty of Computational Mathematics and Cybernetics, Department of Operations Research, Leninskiye Gori, GSP-2, 119899 Moscow, Russia (izmaf@ccas.ru).

[§]Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro, RJ 22460-320, Brazil (solodov@impa.br).

equations are obtained by choosing $l_i = -\infty$, $u_i = +\infty$, $i = 1, \dots, n$. Another important example is the primal-dual Karush–Kuhn–Tucker (KKT) optimality system: find $z \in \mathbf{R}^p$ and $\mu \in \mathbf{R}^m$ such that

$$(1.2) \quad \begin{aligned} g(z) - (G'(z))^T \mu &= 0, \\ \mu &\geq 0, \quad G(z) \geq 0, \quad \langle \mu, G(z) \rangle = 0, \end{aligned}$$

where $g : \mathbf{R}^p \rightarrow \mathbf{R}^p$ and $G : \mathbf{R}^p \rightarrow \mathbf{R}^m$. The KKT system (1.2) can be written as an MCP if we set $n = p + m$ and

$$F(x) = \begin{pmatrix} g(z) - (G'(z))^T \mu \\ G(z) \end{pmatrix}, \quad x = (z, \mu) \in \mathbf{R}^p \times \mathbf{R}^m,$$

$l_i = -\infty$, $i = 1, \dots, p$, $l_i = 0$, $i = p + 1, \dots, n$, $u_i = +\infty$, $i = 1, \dots, n$. Under well-known assumptions, (1.2) represents the first-order primal-dual necessary conditions characterizing solutions in variational inequality or constrained optimization problems. We note that the inclusion of pure equality constraints in the KKT system does not add anything conceptually important in the setting of this paper. For this reason, when talking about the KKT systems we shall consider only the format of (1.2).

This paper follows the development of Newton methods based on the identification of active constraints for KKT systems, presented in [15]. Apart from extending the ideas to MCP, this paper contains a number of improvements and refinements, as will be pointed out in what follows. In particular, the MCP regularity condition introduced here, even when reduced to the special case of KKT systems, is strictly weaker than the one in [15]. Also, some numerical experiments will be reported to illustrate the local behavior of the proposed method under weak assumptions.

As another somewhat related recent work on active-set methods, we mention [4]. In that reference, the special case of NCP is considered. We note that our assumptions for local superlinear convergence are neither stronger nor weaker than those for the method of [4]. On the one hand, [4] can deal with nonisolated solutions, while our assumptions do imply that the given solution is locally unique. On the other hand, monotonicity of F is essential in all of the constructions in [4], while no assumptions of this type are being made in this paper. Also, it seems that the error bound assumed in [4] cannot be directly compared with more standard types of bounds, such as ours. Finally, we consider a more general class of Newton methods, not restricted to the Gauss–Newton method for the (overdetermined) system of nonlinear equations obtained via some identification procedure.

We start in section 2 with deriving a new error bound for MCP (an upper estimate for the distance from a given point to a solution of MCP), based on a smooth reformulation of MCP and a 2-regularity condition [13, 14]. It has been shown previously that in the context of NCP [13] and KKT [15], this construction leads to error bounds which hold under weaker conditions than the alternatives for the corresponding problems (such as b -regularity, semistability, R_0 -property, quasi-regularity, etc.). Here, we extend the analysis and the comparisons to MCP. In addition, we further prove that 2-regularity is both a sufficient and necessary condition for the associated error bound to hold.

Error bounds have many applications [24], among which is identifying active constraints in constrained optimization [7] (see also [8, Ch. 6.7]). In the context of

MCP, those ideas correspond to identifying the sets of indices

$$\begin{aligned} A &= A(\bar{x}) = \{i = 1, \dots, n \mid F_i(\bar{x}) = 0\}, \\ N &= N(\bar{x}) = \{i = 1, \dots, n \mid F_i(\bar{x}) \neq 0\}, \\ N_l &= N_l(\bar{x}) = \{i \in N \mid \bar{x}_i = l_i\}, \\ N_u &= N_u(\bar{x}) = \{i \in N \mid \bar{x}_i = u_i\}, \end{aligned}$$

where $\bar{x}$ is some solution of MCP. If the specified sets can be correctly identified using information available at a point x close enough to the solution $\bar{x}$, then locally MCP can be reduced to a system of nonlinear equations (which is structurally a much simpler problem to solve). In what follows, we shall also use the following partitioning of the set of active indices:

$$\begin{aligned} A_0 &= A_0(\bar{x}) = \{i \in A \mid \bar{x}_i = l_i \text{ or } \bar{x}_i = u_i\}, \\ A_+ &= A_+(\bar{x}) = \{i \in A \mid \bar{x}_i \in (l_i, u_i)\}, \\ A_{0l} &= A_{0l}(\bar{x}) = \{i \in A_0 \mid \bar{x}_i = l_i\}, \\ A_{0u} &= A_{0u}(\bar{x}) = \{i \in A_0 \mid \bar{x}_i = u_i\}. \end{aligned}$$

The analogue of the strict complementarity condition in NCP (or KKT) corresponds, in the setting of MCP, to saying that $A_0 = \emptyset$. Under this assumption, locally MCP trivially reduces to a system of nonlinear equations, which simplifies the local structure of MCP significantly. The condition of strict complementarity, however, is restrictive. We emphasize that this condition is not assumed anywhere in this paper.

In section 3, we propose a new class of active-set Newton methods for solving MCP. Each iteration of the method consists of solving one system of linear equations. We note that when specified to the setting of KKT, this class is different from what has been discussed in [15]. Moreover, the nondegeneracy condition that we introduce here is weaker than the corresponding condition in [15]. Also, the new condition permits a specific deterministic choice of parameters involved in reducing the MCP to a system of equations, while in [15] in general a generic choice of parameters had to be made (at least without strengthening somewhat the regularity assumptions). We show that the conditions needed for the identification of active sets and for convergence of the proposed local Newton method are weaker than semistability of the MCP solution [3, 8] (equivalently, the R_0 -property of the natural residual). This implies, in particular, that the proposed method attains local superlinear/quadratic convergence under assumptions considerably weaker than what is needed for semismooth Newton methods (SNMs) for MCP [1, 19, 9, 16, 18] (BD -regularity of the reformulation being used). Even more remarkably, our assumptions are also strictly weaker than those needed for the linearization (Joseph–Newton) method [17, 3, 12] (which are semistability *and* hemistability of the solution). It should also be noted that in the latter methods subproblems are linearized MCPs, which are computationally more complex than systems of linear equations in our methods.

In section 4, we turn our attention to the specific case of KKT, and in particular consider the issue of superlinear convergence of primal variables. Some comments on the comparison of convergence conditions for various Newton-type methods for MCP constitute section 5. Numerical experiments are presented and discussed in section 6.

A few words about our notation. Given a finite set I , $|I|$ stands for its cardinality. By $\mathbf{R}(m, n)$ we denote the space of $m \times n$ matrices with real entries. By E we shall denote the identity matrix whose dimension would always be clear from the context.

For $x \in \mathbf{R}^n$ and an index set $I \subset \{1, \dots, n\}$, x_I stands for the vector with components x_i , $i \in I$. For a linear operator Λ , $\text{im } \Lambda$ is its range (image space), and $\ker \Lambda$ is its kernel (null space). For a directionally differentiable mapping $\phi : \mathbf{R}^n \rightarrow \mathbf{R}^m$, by $\phi'(x; d)$ we denote the usual directional derivative of ϕ at $x \in \mathbf{R}^n$ in the direction $d \in \mathbf{R}^n$. If $\{z^k\}$ is a sequence in $\mathbf{R}^p$ and $\{t_k\}$ is a sequence in $\mathbf{R}$ such that $t_k \rightarrow 0+$ as $k \rightarrow \infty$, by $z^k = o(t_k)$ we mean that $\lim_{k \rightarrow \infty} \|z^k\|/t_k = 0$.

2. A new error bound for MCP. In this section, we are interested in estimating the distance to a solution of MCP in terms of some *computable* quantity. As is well known [1, 9], MCP can be equivalently reformulated as a system of nonlinear equations via the following transformation.

Let $\psi : \mathbf{R} \times \mathbf{R} \rightarrow \mathbf{R}$ be a *complementarity function*, i.e., a function such that

$$\psi(a, b) = 0 \quad \Leftrightarrow \quad a \geq 0, \quad b \geq 0, \quad ab = 0.$$

Assuming that ψ also satisfies the following additional assumptions:

$$\begin{aligned} a > 0, \quad b < 0 &\Rightarrow \psi(a, b) < 0, \\ a > 0, \quad b > 0 &\Rightarrow \psi(a, b) > 0, \end{aligned}$$

solutions of MCP coincide with solutions of the following system of nonlinear equations:

$$(2.1) \quad \Psi(x) = 0,$$

where

$$\Psi : \mathbf{R}^n \rightarrow \mathbf{R}^n, \quad \Psi_i(x) = \begin{cases} F_i(x) & \text{if } i \in I_F, \\ \psi(x_i - l_i, F_i(x)) & \text{if } i \in I_l, \\ -\psi(u_i - x_i, -F_i(x)) & \text{if } i \in I_u, \\ \psi(x_i - l_i, -\psi(u_i - x_i, -F_i(x))) & \text{if } i \in I_{lu}, \end{cases}$$

$$I_F = \{i = 1, \dots, n \mid -\infty = l_i, \quad u_i = +\infty\},$$

$$I_l = \{i = 1, \dots, n \mid -\infty < l_i, \quad u_i = +\infty\},$$

$$I_u = \{i = 1, \dots, n \mid -\infty = l_i, \quad u_i < +\infty\},$$

$$I_{lu} = \{i = 1, \dots, n \mid -\infty < l_i, \quad u_i < +\infty\}.$$

Complementarity functions to be mentioned in what follows are the natural residual $\psi_{NR}(a, b) = \min\{a, b\}$, the Fischer–Burmeister function $\psi_{FB}(a, b) = a + b - \sqrt{a^2 + b^2}$, and $\psi_S(a, b) = 2ab - (\min\{0, a + b\})^2$ (where S stands for “smooth”). All these functions satisfy the assumptions above. The corresponding reformulations of MCP would be denoted by Ψ_{NR} , Ψ_{FB} , and Ψ_S , respectively. For the purposes of this paper, the signs of the components of Ψ could be chosen differently from the above (see, however, [2, 9] for the justification of the choice adopted here).

Let $\bar{x}$ be a solution of MCP. It is known [8, Ch. 6.2] that the natural residual error bound

$$(2.2) \quad \|x - \bar{x}\| \leq M \|\Psi_{NR}(x)\| \quad \forall x \in U$$

holds for some neighborhood U of $\bar{x}$ and some constant $M > 0$ if and only if $\bar{x}$ is a *semistable* [3, 8] solution of MCP. It can be seen that this is also equivalent to an

error bound in terms of Ψ_{FB} . Furthermore, for MCP, semistability is equivalent to the R_0 -property of Ψ_{NR} at $\bar{x}$, which is $\{\xi \in \mathbf{R}^n \mid \Psi'_{NR}(\bar{x}; \xi) = 0\} = \{0\}$, and the latter is also equivalent to the corresponding property for Ψ_{FB} . As can be easily checked, the R_0 -property means that

$$L = \{0\},$$

where $L = L(\bar{x})$ denotes the solution set of the “linearized” MCP:

$$(2.3) \quad L = \left\{ \xi \in \mathbf{R}^n \left| \begin{array}{l} \xi_i \geq 0, \langle F'_i(\bar{x}), \xi \rangle \geq 0, \xi_i \langle F'_i(\bar{x}), \xi \rangle = 0, \quad i \in A_{0l}, \\ \xi_i \leq 0, \langle F'_i(\bar{x}), \xi \rangle \leq 0, \xi_i \langle F'_i(\bar{x}), \xi \rangle = 0, \quad i \in A_{0u}, \\ F'_{A_+}(\bar{x})\xi = 0, \quad \xi_N = 0 \end{array} \right. \right\}.$$

Semistability is one of the weakest conditions under which a computable error bound for MCP had been exhibited up to now. To our knowledge, alternative conditions are either stronger or different in nature and not comparable to semistability (e.g., analyticity [23] or subanalyticity of F). In what follows, we provide an error bound under a condition which we show to be strictly weaker than semistability.

DEFINITION 2.1. *Let $\Psi : \mathbf{R}^n \rightarrow \mathbf{R}^n$ be differentiable in a neighborhood of $\bar{x} \in \mathbf{R}^n$ and $\Psi' : \mathbf{R}^n \rightarrow \mathbf{R}(n, n)$ be directionally differentiable at $\bar{x}$. Then Ψ is 2-regular at $\bar{x}$ if*

$$T = \{0\},$$

where

$$(2.4) \quad \begin{aligned} T = T(\bar{x}) &= \{\xi \in \ker \Psi'(\bar{x}) \mid (\Psi')'(\bar{x}; \xi)\xi \in \text{im } \Psi'(\bar{x})\} \\ &= \{\xi \in \ker \Psi'(\bar{x}) \mid P(\Psi')'(\bar{x}; \xi)\xi = 0\}, \end{aligned}$$

with P being the orthogonal projector onto $(\text{im } \Psi'(\bar{x}))^\perp$.

The above is a special case of 2-regularity of a nonlinear mapping [14, 13], corresponding to the case when the mapping acts from some space into itself.

We next give an error bound result based on the smooth MCP reformulation Ψ_S . A comparison with semistability will be made later. We note that the fact that the error bound below is actually equivalent to 2-regularity is new.

THEOREM 2.2. *Let $F : \mathbf{R}^n \rightarrow \mathbf{R}^n$ be sufficiently smooth near a point $\bar{x} \in \mathbf{R}^n$, which is a solution of MCP.*

The mapping Ψ_S is 2-regular at $\bar{x}$ if and only if there exist a neighborhood U of $\bar{x}$ and a constant $M > 0$ such that

$$(2.5) \quad \|x - \bar{x}\| \leq M(\|(E - P)\Psi_S(x)\| + \|P\Psi_S(x)\|^{1/2}) \quad \forall x \in U.$$

Proof. Obviously, Ψ_S satisfies the smoothness assumptions in Definition 2.1. The sufficiency part of the assertion is a direct consequence of [13, Theorem 4]. We prove the necessity part.

Take any $\xi \in T$. For any $t \geq 0$, we have that

$$\|\Psi_S(\bar{x} + t\xi)\| = \|\Psi_S(\bar{x} + t\xi) - \Psi_S(\bar{x})\| = \|t\Psi'_S(\bar{x})\xi\| + o(t) = o(t),$$

where the last equality follows from $\xi \in \ker \Psi'_S(\bar{x})$. We further have that

$$\begin{aligned}
\|P\Psi_S(\bar{x} + t\xi)\| &= \|P(\Psi_S(\bar{x} + t\xi) - \Psi_S(\bar{x}))\| \\
&\leq t \sup_{\tau \in [0,1]} \|P\Psi'_S(\bar{x} + \tau t\xi)\| \|\xi\| \\
&= t \sup_{\tau \in [0,1]} \|P(\Psi'_S(\bar{x} + \tau t\xi) - \Psi'_S(\bar{x}))\| \|\xi\| \\
&= t \sup_{\tau \in [0,1]} \|\tau t P(\Psi'_S)'(\bar{x}; \xi) \xi\| \|\xi\| + o(t^2) \\
&= o(t^2),
\end{aligned}$$

where we have used the mean-value theorem, the fact that $P\Psi'_S(\bar{x}) = 0$, the positive homogeneity of the mapping $P(\Psi'_S)'(\bar{x}; \cdot)$, and the fact that $P(\Psi'_S)'(\bar{x}; \xi)\xi = 0$.

Therefore, (2.5) implies that

$$t\|\xi\| = \|\bar{x} + t\xi - \bar{x}\| \leq M(\|(E - P)\Psi_S(\bar{x} + t\xi)\| + \|P\Psi_S(\bar{x} + t\xi)\|^{1/2}) = o(t),$$

which means that $\xi = 0$. We have thus established that $T = \{0\}$, i.e., that Ψ_S is 2-regular at $\bar{x}$. $\square$

Adjusting M and U , if necessary, the error bound (2.5) can be simplified into the following relation (less accurate but possibly easier to use):

$$(2.6) \quad \|x - \bar{x}\| \leq M\|\Psi_S(x)\|^{1/2} \quad \forall x \in U.$$

Note that for NCP or KKT the error bound (2.6) is implied by error bound (2.2) (this follows from a comparison of growth rates for ψ_{NR} , ψ_{FB} , and ψ_S , given in [28]). However, for MCP it is not clear whether one can use the same comparison, as the definition of Ψ involves a superposition of the functions ψ . In any case, the more accurate estimate (2.5) does not follow from (2.2). Moreover, (2.5) and (2.6) can hold when (2.2) does not, as shown next.

PROPOSITION 2.3. *Semistability of a solution $\bar{x}$ of MCP (equivalently, error bound (2.2)) implies 2-regularity of Ψ_S at $\bar{x}$ (equivalently, error bound (2.5)) but not vice versa.*

Proof. Let $\bar{x}$ be a solution of MCP. The fact that 2-regularity of Ψ_S at $\bar{x}$ can hold when the error bound (2.2) (semistability of $\bar{x}$) does not has already been shown for two special cases of MCP, namely, NCP [13, Example 1] and KKT [15, Example 2]. Thus no further justification is needed for this assertion.

Let $\bar{x}$ be a semistable solution. To prove that this implies 2-regularity of Ψ_S at $\bar{x}$, it suffices to show that $T \subset L$, where T and L are defined in (2.4) and (2.3), respectively.

By direct computations, we have that

$$(\Psi'_S)_i(\bar{x}) = \begin{cases} 0 & \text{if } i \in A_0, \\ \alpha_i F'_i(\bar{x}) & \text{if } i \in A_+, \\ \beta_i e^i & \text{if } i \in N, \end{cases}$$

where $e^1, \dots, e^n$ is the canonical basis in $\mathbf{R}^n$, and

$$\alpha_i = \begin{cases} 2(\bar{x}_i - l_i) & \text{if } i \in I_l, \\ 2(u_i - \bar{x}_i) & \text{if } i \in I_u, \\ 1 & \text{if } i \in I_F, \\ 4(\bar{x}_i - l_i)(u_i - \bar{x}_i) & \text{if } i \in I_{lu}, \end{cases}$$

$$\beta_i = \begin{cases} 2F_i(\bar{x}) & \text{if } i \in I_l, \\ -2F_i(\bar{x}) & \text{if } i \in I_u, \\ 4F_i(\bar{x})(u_i - l_i) + 2(\min\{0, u_i - l_i - F_i(\bar{x})\})^2 & \text{if } i \in I_{lu} \cap N_l, \\ -4F_i(\bar{x})(u_i - l_i) & \text{if } i \in I_{lu} \cap N_u. \end{cases}$$

Observe that $\alpha_i \neq 0 \ \forall i \in A_+$ and $\beta_i \neq 0 \ \forall i \in N$. Hence,

$$\ker \Psi'_S(\bar{x}) = \{\xi \in \mathbf{R}^n \mid F'_{A_+}(\bar{x})\xi = 0, \ \xi_N = 0\}.$$

Then P , the orthogonal projector onto $(\text{im } \Psi'_S(\bar{x}))^\perp$ in $\mathbf{R}^n$, satisfies

$$(Py)_i = y_i \quad \forall y \in \mathbf{R}^n, \ \forall i \in A_0.$$

Since $P\Psi'_S = (P\Psi_S)'$, it now easily follows that

$$(2.7) \quad T \subset \{\xi \in \mathbf{R}^n \mid (\Psi'_S)'_i(\bar{x}; \xi)\xi = 0, \ i \in A_0, \ F'_{A_+}(\bar{x})\xi = 0, \ \xi_N = 0\}.$$

For any $i \in A_0$ and $\xi \in \mathbf{R}^n$, we further obtain that

$$(\Psi'_S)'_i(\bar{x}; \xi)\xi = 2 \begin{cases} \psi_S(\xi_i, \langle F'_i(\bar{x}), \xi \rangle) & \text{if } i \in I_l, \\ -\psi_S(-\xi_i, -\langle F'_i(\bar{x}), \xi \rangle) & \text{if } i \in I_u, \\ \psi_S(\xi_i, (u_i - l_i)\langle F'_i(\bar{x}), \xi \rangle) & \text{if } i \in I_{lu} \cup A_{0l}, \\ -(u_i - l_i)\psi_S(-\xi_i, -\langle F'_i(\bar{x}), \xi \rangle) & \text{if } i \in I_{lu} \cup A_{0u}. \end{cases}$$

Since ψ_S is a complementarity function, the right-hand side of the latter equality is zero if and only if

$$\begin{aligned} \xi_i &\geq 0, \quad \langle F'_i(\bar{x}), \xi \rangle \geq 0, \quad \xi_i \langle F'_i(\bar{x}), \xi \rangle = 0, \quad i \in A_{0l}, \\ \xi_i &\leq 0, \quad \langle F'_i(\bar{x}), \xi \rangle \leq 0, \quad \xi_i \langle F'_i(\bar{x}), \xi \rangle = 0, \quad i \in A_{0u}. \end{aligned}$$

Hence, the right-hand side in (2.7) coincides with L defined in (2.3). In particular, we have established that $T \subset L$, which completes the proof. $\square$

3. A class of active-set Newton methods. The following technique for identifying the relevant index sets is based on the ideas of [7]; see also [8, Ch. 6.7]. Define the *identification function*

$$\rho : \mathbf{R}_+ \rightarrow \mathbf{R}, \quad \rho(t) = \begin{cases} \bar{\rho} & \text{if } t \geq \bar{t}, \\ -1/\log t & \text{if } t \in (0, \bar{t}), \\ 0 & \text{if } t = 0, \end{cases}$$

where $\bar{t} \in (0, 1)$ and $\bar{\rho} > 0$ are fixed numbers (the choice of $\bar{t}$ and $\bar{\rho}$ does not affect theoretical analysis; in our numerical experiments reported in section 6, we use $\bar{t} = 0.9$

and $\bar{\rho} = -1/\log \bar{t}$, as suggested in [7]). For any $x \in \mathbf{R}^n$, define further the index sets

$$(3.1) \quad A(x) = \{i = 1, \dots, n \mid |F_i(x)| \leq \rho(\|\Psi_S(x)\|)\},$$

$$(3.2) \quad N(x) = \{1, \dots, n\} \setminus A(x),$$

$$(3.3) \quad N_l(x) = \{i \in N(x) \mid x_i - l_i \leq u_i - x_i\}, \quad N_u(x) = N(x) \setminus N_l(x),$$

$$(3.4) \quad A_0(x) = \{i \in A(x) \mid \min\{|x_i - l_i|, |u_i - x_i|\} \leq \rho(\|\Psi_S(x)\|)\},$$

$$(3.5) \quad A_+(x) = A(x) \setminus A_0(x),$$

$$(3.6) \quad A_{0l}(x) = \{i \in A_0(x) \mid x_i - l_i \leq u_i - x_i\}, \quad A_{0u}(x) = A_0(x) \setminus A_{0l}(x).$$

PROPOSITION 3.1. *If Ψ_S is 2-regular at a solution $\bar{x}$ of MCP (equivalently, the error bound (2.5) holds), then for any $x \in \mathbf{R}^n$ sufficiently close to $\bar{x}$, it holds that*

$$(3.7) \quad A(x) = A, \quad N(x) = N, \quad N_l(x) = N_l, \quad N_u(x) = N_u,$$

$$(3.8) \quad A_{0l}(x) = A_{0l}, \quad A_{0u}(x) = A_{0u}, \quad A_0(x) = A_0, \quad A_+(x) = A_+.$$

Proof. Take any $i \in A$. Then for any x close enough to $\bar{x}$, we have that

$$|F_i(x)| = |F_i(x) - F_i(\bar{x})| \leq K\|x - \bar{x}\| \leq KM\|\Psi_S(x)\|^{1/2} \leq \rho(\|\Psi_S(x)\|),$$

where the second inequality is by the local Lipschitz-continuity of F (with some modulus $K > 0$), the third inequality is by (2.5), and the last follows from the fact that $\lim_{t \rightarrow 0+} t^\nu \log t = 0$ for any $\nu > 0$. The above shows that $i \in A(x)$. Hence, $A \subset A(x)$.

Take any $i \in \{1, \dots, n\} \setminus A$. In that case, there exists some $\gamma > 0$ such that for any x close enough to $\bar{x}$ it holds that

$$|F_i(x)| \geq \gamma, \quad \rho(\|\Psi_S(x)\|) < \gamma.$$

It follows that $i \notin A(x)$, which shows that $A(x) \subset A$.

We have therefore established the first (and hence the second) equality in (3.7).

The other relations either hold trivially (e.g., (3.3)) or can be verified by considerations similar to the above. $\square$

We note that any other MCP reformulation Ψ with a corresponding valid error bound can be used in the identification procedure. But since it has been established above that Ψ_S requires the weakest assumptions for the error bound to hold, it is fair to say that this is the function which should be used for this purpose. However, different choices of the function ρ are possible under the same assumptions. For example,

$$\rho : \mathbf{R}_+ \rightarrow \mathbf{R}, \quad \rho(t) = t^\theta, \quad \theta \in (0, 1/2).$$

Observe also that in the implementation of the identification procedure, the following obvious relations can be taken into account: $I_F \subset A_+$, $I_l \subset (A_{0l} \cup A_+ \cup N_l)$, and $I_u \subset (A_{0u} \cup A_+ \cup N_u)$.

Once the index sets are identified, we have the following relations, which are guaranteed to be satisfied at a solution $\bar{x}$ of MCP:

$$F_A(x) = 0, \quad x_{A_{0l} \cup N_l} = l_{A_{0l} \cup N_l}, \quad x_{A_{0u} \cup N_u} = u_{A_{0u} \cup N_u}.$$

For simplicity of notation, suppose that the components of $x \in \mathbf{R}^n$ are ordered in such a way that $x = (x_{A_+}, x_{A_{0l} \cup N_l}, x_{A_{0u} \cup N_u})$. Then MCP locally reduces to the following system of nonlinear equations:

$$(3.9) \quad F_A(x_{A_+}, l_{A_{0l} \cup N_l}, u_{A_{0u} \cup N_u}) = 0.$$

Observe that in the absence of strict complementarity (when $A_0 \neq \emptyset$, i.e., $|A| > |A_+|$), the system is overdetermined (the number of equations is larger than the number of unknowns). This opens up a number of options. Of course, one can just solve the system by the Gauss–Newton method (GNM). This possibility will be considered. However, we prefer not to limit ourselves to GNM for the following reason: the Gauss–Newton approach can destroy structure present in F_A (for example, sparsity or the primal–dual structure in the case of KKT).

Our proposal is to consider the following system of nonlinear equations:

$$(3.10) \quad \Phi_C(x_{A_+}) = 0,$$

where

$$\Phi_C : \mathbf{R}^{|A_+|} \rightarrow \mathbf{R}^{|A_+|}, \quad \Phi_C(x_{A_+}) = C(x_{A_+})F_A(x_{A_+}, l_{A_0l \cup N_l}, u_{A_0u \cup N_u}),$$

with $C : \mathbf{R}^{|A_+|} \rightarrow \mathbf{R}^{|A_+|, |A|}$ being a smooth mapping (possibly constant). Clearly, $\bar{x}_{A_+}$ is a solution of (3.10) for any choice of C . The Jacobian of (3.10) at this solution is given by

$$(3.11) \quad \Phi'_C(\bar{x}_{A_+}) = C(\bar{x}_{A_+}) \frac{\partial F_A}{\partial x_{A_+}}(\bar{x}),$$

where we have taken into account that $F_A(\bar{x}) = 0$. Thus $\bar{x}_{A_+}$ can be found by applying Newton-type methods to (3.10) whenever the matrix in (3.11) is nonsingular.

Note that GNM for (3.9) would essentially correspond to choosing in (3.10)

$$(3.12) \quad C(x_{A_+}) = \left(\frac{\partial F_A}{\partial x_{A_+}}(x_{A_+}, l_{A_0l \cup N_l}, u_{A_0u \cup N_u}) \right)^T$$

and applying to the resulting system an approximate version of the pure Newton method. Indeed, with the notation of (3.12), the Gauss–Newton iteration for (3.9) has the form

$$(3.13) \quad x_{A_+}^{k+1} = x_{A_+}^k - \left(C(x_{A_+}^k) \frac{\partial F_A}{\partial x_{A_+}}(x_{A_+}^k, l_{A_0l \cup N_l}, u_{A_0u \cup N_u}) \right)^{-1} \Phi_C(x_{A_+}^k).$$

Observe that the above formula is just an approximation of the standard Newton iteration for (3.10), where the Jacobian $\Phi'_C(x_{A_+}^k)$ is replaced by

$$C(x_{A_+}^k) \frac{\partial F_A}{\partial x_{A_+}}(x_{A_+}^k, l_{A_0l \cup N_l}, u_{A_0u \cup N_u}).$$

Due to (3.11), this change preserves the superlinear convergence of the pure Newton iteration for (3.10). Note finally that with the choice of (3.12), we have

$$(3.14) \quad \Phi'_C(\bar{x}_{A_+}) = \left(\frac{\partial F_A}{\partial x_{A_+}}(\bar{x}) \right)^T \frac{\partial F_A}{\partial x_{A_+}}(\bar{x}).$$

This immediately motivates the following definition.

DEFINITION 3.2. A solution $\bar{x}$ of MCP is referred to as weakly regular if

$$\text{rank } \frac{\partial F_A}{\partial x_{A_+}}(\bar{x}) = |A_+|.$$

Clearly, weak regularity is a necessary and sufficient condition for the matrix in (3.14) to be nonsingular, and hence for the superlinear convergence of GNM applied to (3.9) (or the approximate Newton method applied to (3.10) with the choice of (3.12)). It is also clear that weak regularity is necessary for the nonsingularity of the matrix in (3.11) corresponding to the more general scheme, and this is regardless of the choice of C .

We next show that weak regularity is implied by semistability, but not vice versa. Moreover, 2-regularity of Ψ_S at $\bar{x}$ and weak regularity, when combined, are still a weaker condition than semistability.

PROPOSITION 3.3. *Let $\bar{x}$ be a solution of MCP. Then semistability of $\bar{x}$ implies weak regularity of $\bar{x}$ but not vice versa.*

Proof. Suppose that $\bar{x}$ is a semistable solution. If it is not weakly regular, then there exists $\xi_{A_+} \in \ker \frac{\partial F_A}{\partial x_{A_+}}(\bar{x}) \setminus \{0\}$. But then setting $\xi_{A_0 \cup N} = 0$, we obtain $\xi \neq 0$ such that $\xi \in L$, where L is defined in (2.3). This contradicts semistability.

The lack of the reverse implication is established in Example 3.5. $\square$

The following result is of special importance.

PROPOSITION 3.4. *Let $\bar{x}$ be a solution of MCP. Then semistability of $\bar{x}$ implies the combination of 2-regularity of Ψ_S at $\bar{x}$ and weak regularity of $\bar{x}$ but not vice versa.*

Proof. The forward assertion is by Propositions 2.3 and 3.3. The lack of the reverse implication is shown in Example 3.5. $\square$

Example 3.5. Let $n = 2$, $l_i = 0$, $u_i = +\infty$, $i = 1, 2$, and let $F(x) = ((x_1 - 1)^2, x_1 + x_2 - 1)$.

The point $\bar{x} = (1, 0) \in \mathbf{R}^2$ is the solution of this NCP, and we have $A = \{1, 2\}$, $A_0 = A_{0l} = \{2\}$, $A_+ = \{1\}$, with all the other index sets being empty.

We first verify that semistability is violated. Noting that $F'_{A_+}(\bar{x}) = 0$, it can be seen that the cone L defined in (2.3) is

$$L = \{\xi \in \mathbf{R}^2 \mid \xi_2 \geq 0, \xi_1 + \xi_2 \geq 0, \xi_2(\xi_1 + \xi_2) = 0\} \neq \{0\}.$$

Thus $\bar{x}$ is not semistable.

Weak regularity certainly holds, as

$$\frac{\partial F_A}{\partial x_{A_+}}(\bar{x}) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \text{rank } \frac{\partial F_A}{\partial x_{A_+}}(\bar{x}) = 1 = |A_+|.$$

We proceed to show 2-regularity of Ψ_S at $\bar{x}$. It can be seen that $\Psi'_S(\bar{x}) = 0$. Hence,

$$\ker \Psi'_S(\bar{x}) = \mathbf{R}^2 = (\text{im } \Psi'_S(\bar{x}))^\perp, \quad P = E,$$

and further

$$T = \{\xi \in \mathbf{R}^2 \mid (\Psi'_S)'(\bar{x}; \xi)\xi = 0\}.$$

We obtain that

$$(\Psi'_S)'(\bar{x}; \xi)\xi = \begin{pmatrix} 4\xi_1^2 \\ 2\xi_2(\xi_1 + 2\xi_2) - \min\{0, \xi_2 + (\xi_1 + \xi_2)\} \end{pmatrix}.$$

Hence,

$$T = \{\xi \in L \mid \xi_1 = 0\} = \{0\},$$

and Ψ_S is 2-regular at $\bar{x}$.

We have thus constructed a local algorithm with superlinear convergence under assumptions weaker than semistability of the MCP solution. Specifically, we have the following.

THEOREM 3.6. *Let $F : \mathbf{R}^n \rightarrow \mathbf{R}^n$ be sufficiently smooth near a point $\bar{x} \in \mathbf{R}^n$, which is a solution of MCP. Suppose that this solution is weakly regular and Ψ_S is 2-regular at $\bar{x}$.*

For any $x^0 \in \mathbf{R}^n$ sufficiently close to $\bar{x}$, if the index sets $A = A(x^0)$, $A_+ = A_+(x^0)$, $A_{0l} = A_{0l}(x^0)$, $A_{0u} = A_{0u}(x^0)$, $N_l = N_l(x^0)$, and $N_u = N_u(x^0)$ are defined according to (3.1)–(3.6), then GNM applied to the system (3.9) (with $x_{A_+}^0$ as a starting point) is well defined and superlinearly convergent to $\bar{x}_{A_+}$.

Proof. By 2-regularity of Ψ_S at $\bar{x}$ and Proposition 3.1, for any $x = x^0$ close to $\bar{x}$ the index sets defined according to (3.1)–(3.6) correctly identify the index sets at the solution $\bar{x}$. Then $\bar{x}_{A_+}$ is the solution of (3.9). By weak regularity, we have that the matrix in (3.14) is nonsingular. Hence, GNM applied to (3.9) is locally superlinearly convergent to $\bar{x}_{A_+}$. $\square$

As already mentioned above (see also section 4), it sometimes can be useful to choose the mapping C differently from the Gauss–Newton option of (3.12). For example, we might want to take $C(\cdot) = C \in \mathbf{R}(|A_+|, |A|)$, a fixed matrix, in order to preserve in the matrix $C \frac{\partial F_A}{\partial x_{A_+}}(x_{A_+}, l_{A_{0l} \cup N_l}, u_{A_{0u} \cup N_u})$ the structure (primal-dual, sparsity, etc.) of the matrix $\frac{\partial F_A}{\partial x_{A_+}}(x_{A_+}, l_{A_{0l} \cup N_l}, u_{A_{0u} \cup N_u})$. This motivates the following considerations.

PROPOSITION 3.7. *Suppose that a solution $\bar{x}$ of MCP is weakly regular. Then the set of matrices $C \in \mathbf{R}(|A_+|, |A|)$ such that $\Phi'_C(\bar{x}_{A_+})$ is nonsingular is open and dense in $\mathbf{R}(|A_+|, |A|)$.*

Proof. The determinant $\det \Phi'_C(\bar{x}_{A_+})$ is a polynomial with respect to the elements of the matrix $C \in \mathbf{R}(|A_+|, |A|)$. By weak regularity (see Definition 3.2), this polynomial is not everywhere zero, because it is not zero for the choice $C = (\frac{\partial F_A}{\partial x_{A_+}}(\bar{x}))^T$. Hence, the set where this polynomial is not zero is obviously open and dense in $\mathbf{R}(|A_+|, |A|)$. $\square$

Proposition 3.7 justifies choosing C in any desirable way, as the chance that the resulting system would be degenerate is negligible (the set of matrices for which this would happen is of the Lebesgue measure zero). Of course, one should make reasonable choices. For example, it should hold that $\text{rank } C = |A_+|$.

4. KKT systems. In the case of the KKT system (1.2), the developments of section 3 give

$$F_A(x_{A_+}, l_{A_{0l} \cup N_l}, u_{A_{0u} \cup N_u}) = \begin{pmatrix} g(z) - (G'_{I_+}(z))^T \mu_{I_+} \\ G_I(z) \end{pmatrix},$$

$$x_{A_+} = (z, \mu_{I_+}) \in \mathbf{R}^p \times \mathbf{R}^{|I_+|},$$

where

$$I = I(\bar{z}) = \{i = 1, \dots, m \mid G_i(\bar{z}) = 0\}, \quad I_+ = I_+(\bar{z}) = \{i \in I \mid \bar{\mu}_i > 0\},$$

$$A = \{1, \dots, p\} \cup \{p+j \mid j \in I\}, \quad A_+ = \{1, \dots, p\} \cup \{p+j \mid j \in I_+\}.$$

Defining

$$\phi : \mathbf{R}^p \times \mathbf{R}^{|I_+|} \rightarrow \mathbf{R}^p, \quad \phi(z, \mu_{I_+}) = g(z) - (G'_{I_+}(z))^T \mu_{I_+},$$

Definition 3.2 of weak regularity consists of saying that

$$(4.1) \quad \text{rank} \begin{pmatrix} \frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+}) & -(G'_{I_+}(\bar{z}))^T \\ G'_I(\bar{z}) & 0 \end{pmatrix} = p + |I_+|.$$

In the case of optimization (g is a gradient mapping), $\frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+})$ coincides with the Hessian of the standard Lagrangian function at a KKT point $(\bar{z}, \bar{\mu})$.

We first show that (4.1) is weaker than the regularity condition for KKT systems introduced in [15, Definition 2]. Defining $I_0 = I \setminus I_+$, the latter states that there exist $D_1, D_2 \in \mathbf{R}(|I_0|, |I_0|)$ such that

$$(4.2) \quad \det \begin{pmatrix} \frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+}) & -(G'_{I_0}(\bar{z}))^T & -(G'_{I_+}(\bar{z}))^T \\ D_1 G'_{I_0}(\bar{z}) & D_2 & 0 \\ G'_{I_+}(\bar{z}) & 0 & 0 \end{pmatrix} \neq 0.$$

PROPOSITION 4.1. *Let $(\bar{z}, \bar{\mu})$ be a solution of the KKT system. Then (4.2) implies (4.1) but not vice versa.*

Proof. Assume (4.2). If (4.1) does not hold, then there exists $(\zeta, \nu_{I_+}) \neq 0$ in the kernel of the matrix in (4.1). But then $(\zeta, 0, \nu_{I_+}) \neq 0$ will be in the kernel of the matrix in (4.2) for any choice of $D_1, D_2 \in \mathbf{R}(|I_0|, |I_0|)$.

The lack of the reverse implication is shown in Example 4.2. $\square$

Example 4.2. Let $p = 2$, $m = 1$, $g(z) = (z_2, 0)$, $G(z) = z_1$.

The point $(\bar{z}, \bar{\mu}) = (0, 0)$ is a solution of the KKT system (1.2), and we have $I_+ = \emptyset$, $I_0 = \{1\}$.

The matrix in (4.1) takes the form

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \\ 1 & 0 \end{pmatrix},$$

and it has full column rank.

The matrix in (4.2) takes the form

$$\begin{pmatrix} 0 & 1 & -1 \\ 0 & 0 & 0 \\ D_1 & 0 & D_2 \end{pmatrix},$$

which is always singular.

Note that in Example 4.2, $\frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+})$ is asymmetric. It is an open question at this time whether the conditions (4.2) and (4.1) are different also in the symmetric (optimization) case.

We next provide a characterization of weak regularity for KKT systems in terms of a constraint qualification and a second-order condition. We say that the weak linear independence constraint qualification (WLICQ) holds at $\bar{z}$ if

$$\text{rank } G'_{I_+}(\bar{z}) = |I_+|.$$

We say that the second-order condition (SOC) holds if

$$\left\langle \frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+}) \zeta, \zeta \right\rangle \neq 0 \quad \forall \zeta \in \ker G'_I(\bar{z}) \setminus \{0\}.$$

Obviously, WLICQ is implied by the standard linear independence constraint qualification, while SOC is implied by the standard second-order sufficiency condition.

PROPOSITION 4.3. *Let $(\bar{z}, \bar{\mu})$ be a solution of the KKT system. Then WLICQ and SOC imply weak regularity of $(\bar{z}, \bar{\mu})$. Weak regularity implies WLICQ.*

Proof. The fact that weak regularity subsumes WLICQ is obvious.

In view of (4.1), it suffices to prove that under WLICQ and SOC, the equality

$$(4.3) \quad \begin{pmatrix} \frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+})\zeta - (G'_{I_+}(\bar{z}))^T \nu_{I_+} \\ G'_I(\bar{z})\zeta \end{pmatrix} = 0$$

implies that $(\zeta, \nu_{I_+}) = 0$. Indeed, from (4.3) we obtain that

$$\left\langle \frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+})\zeta, \zeta \right\rangle = \langle (G'_{I_+}(\bar{z}))^T \nu_{I_+}, \zeta \rangle = \langle \nu_{I_+}, G'_{I_+}(\bar{z})\zeta \rangle = 0.$$

Since $\zeta \in \ker G'_I(\bar{z})$, SOC implies that $\zeta = 0$. Thus $(G'_{I_+}(\bar{z}))^T \nu_{I_+} = 0$. By WLICQ, we have that $\nu_{I_+} = 0$, which concludes the proof. $\square$

Under the weak regularity condition (4.1) and 2-regularity of Ψ_S at $(\bar{z}, \bar{\mu})$, we can solve the reduced KKT system (3.9) by GNM, with the convergence result given by Theorem 3.6. We next show that for KKT systems with symmetric $\frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+})$, the latter two assumptions, when combined, are equivalent to semistability of $(\bar{z}, \bar{\mu})$. (Under the second-order necessity condition, the latter is further equivalent to the uniqueness of $\bar{\mu}$ associated to $\bar{z}$ and the second-order sufficiency condition; see [3, Proposition 6.2], [15, Proposition 1].)

PROPOSITION 4.4. *If $\frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+})$ is symmetric, then semistability of a solution $(\bar{z}, \bar{\mu})$ of the KKT system is equivalent to the combination of 2-regularity of Ψ_S at this solution and weak regularity of this solution.*

Proof. The fact that semistability implies the other two properties is given by Proposition 3.4.

We next re-examine the proof of Proposition 2.3 under the new assumptions. In addition to the sets of indices defined above, let $I_N = \{1, \dots, m\} \setminus I$ and $I_0 = I \setminus I_+$. Suppose, for simplicity of notation, that the ordering is such that in the set $\{1, \dots, m\}$, the indices in I_+ come first, then those in I_0 , and then those in I_N . With this convention, the matrix whose rows are comprised by $F'(\bar{x})$, $i \in A_+$, e^i , $i \in N$, is given by

$$\begin{pmatrix} \frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+}) & -(G'_{I_+}(\bar{z}))^T & -(G'_{I_0}(\bar{z}))^T & -(G'_{I_N}(\bar{z}))^T \\ G'_{I_+}(\bar{z}) & 0 & 0 & 0 \\ 0 & 0 & 0 & E_{I_N} \end{pmatrix}.$$

Under the assumption that $\frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+})$ is symmetric, the weak regularity condition (4.1) implies that the rows of this matrix are linearly independent. Revisiting the proof of Proposition 2.3, it is easily seen that in this case

$$\text{im } \Psi'_S(\bar{x}) = \{y \in \mathbf{R}^n \mid y_i = 0, i \in A_0\},$$

and

$$(Py)_i = \begin{cases} y_i & \text{if } i \in A_0, \\ 0 & \text{if } i \in A_+ \cup N, \end{cases} \quad y \in \mathbf{R}^n,$$

where

$$A_0 = A \setminus A_+ = \{p + j \mid j \in I_0\}, \quad N = \{i = 1, \dots, n\} \setminus A = \{p + j \mid j \in I_N\}.$$

It follows that the inclusion (2.7) holds as equality, in which case the proof of Proposition 2.3 establishes that $T = L$, where T and L are defined in (2.4) and (2.3), respectively. Semistability and 2-regularity are therefore equivalent in that case. $\square$

It is not difficult to observe that the Gauss–Newton iteration in a certain sense destroys the primal–dual structure present in a KKT system. For example, it does not seem possible to analyze the superlinear convergence of the primal variables separately from the convergence of the primal–dual pair. Proposition 3.7, on the other hand, allows us to make other choices of C in (3.10), with the expectation that they would still do the job. We next make one specific choice and analyze conditions for the superlinear convergence of primal variables. We refer the reader also to [15] (the discussion following Proposition 6) for some possibilities of how further assumptions about the problem can be taken into account in the framework of that paper. Similar options under similar assumptions could be analyzed here, but we shall not pursue the details.

Assuming again that the active constraints are ordered in such a way that the strongly active are first (i.e., for the first $|I_+|$ active constraints the corresponding multiplier is positive), let

$$(4.4) \quad C = \begin{pmatrix} C_1 & 0 & 0 \\ 0 & C_2 & 0 \end{pmatrix},$$

where $C_1 \in \mathbf{R}(p, p)$ and $C_2 \in \mathbf{R}(|I_+|, |I_+|)$ are arbitrary nonsingular matrices. Note that $\text{rank } C = p + |I_+| = |A_+|$ holds. We shall consider further the case where in the implementation of the Newton method for (3.10) the matrix $\frac{\partial \phi}{\partial z}(z^k, \mu_{I_+}^k)$ at iteration k is replaced by some (e.g., quasi-Newton) approximation H_k . The resulting iteration is then given by

$$(4.5) \quad C \begin{pmatrix} g(z^k) - (G'_{I_+}(z^k))^T \mu_{I_+}^k + H_k(z^{k+1} - z^k) - (G'_{I_+}(z^k))^T (\mu_{I_+}^{k+1} - \mu_{I_+}^k) \\ G_I(z^k) + G'_I(z^k)(z^{k+1} - z^k) \end{pmatrix} = 0,$$

and taking into account the chosen structure of C , it holds that

$$(4.6) \quad g(z^k) + H_k(z^{k+1} - z^k) - (G'_{I_+}(z^k))^T \mu_{I_+}^{k+1} = 0,$$

$$(4.7) \quad G_{I_+}(z^k) + G'_{I_+}(z^k)(z^{k+1} - z^k) = 0.$$

To establish a sufficient condition for the superlinear rate of convergence of primal variables, we shall need assumptions somewhat stronger than weak regularity. We say that the strong second-order sufficiency condition (SSOSC) holds if

$$\left\langle \frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+}) \zeta, \zeta \right\rangle \neq 0 \quad \forall \zeta \in \ker G'_{I_+}(\bar{z}) \setminus \{0\}.$$

THEOREM 4.5. *Suppose that the sequence $\{(z^k, \mu_{I_+}^k)\}$ generated according to (4.5) with C given by (4.4) converges to $(\bar{z}, \bar{\mu}_{I_+})$, a solution of the KKT system.*

If $\{z^k\}$ converges to $\bar{z}$ superlinearly, then

$$(4.8) \quad \Pi \left(\left(\frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+}) - H_k \right) (z^{k+1} - z^k) \right) = o(\|z^{k+1} - z^k\|),$$

where Π is the orthogonal projector onto $\ker G'_{I_+}(\bar{z})$ in $\mathbf{R}^p$.

Under *WLICQ* and *SSOSC*, the condition (4.8) is also sufficient for $\{z^k\}$ to converge to $\bar{z}$ at a superlinear rate.

Proof. By (4.6), we have that

$$\begin{aligned} -H_k(z^{k+1} - z^k) &= g(z^k) - (G'_{I_+}(z^k))^T \mu_{I_+}^{k+1} \\ &= g(\bar{z}) - (G'_{I_+}(\bar{z}))^T \mu_{I_+}^{k+1} + \frac{\partial \phi}{\partial z}(\bar{z}, \mu_{I_+}^{k+1})(z^k - \bar{z}) + o(\|z^k - \bar{z}\|) \\ &= (G'_{I_+}(\bar{z}))^T (\bar{\mu}_{I_+} - \mu_{I_+}^{k+1}) + \frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+})(z^k - \bar{z}) + o(\|z^k - \bar{z}\|), \end{aligned}$$

where in the third equality we have used the fact that $g(\bar{z}) = (G'_{I_+}(\bar{z}))^T \bar{\mu}_{I_+}$ and the assumption that $\{\mu_{I_+}^k\} \rightarrow \bar{\mu}_{I_+}$. Hence,

$$\begin{aligned} (4.9) \quad &\left(\frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+}) - H_k \right) (z^{k+1} - z^k) \\ &= (G'_{I_+}(\bar{z}))^T (\bar{\mu}_{I_+} - \mu_{I_+}^{k+1}) + \frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+})(z^{k+1} - \bar{z}) + o(\|z^k - \bar{z}\|). \end{aligned}$$

Suppose first that $z^{k+1} - \bar{z} = o(\|z^k - \bar{z}\|)$. Then we obtain from (4.9) that

$$(4.10) \quad \left(\frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+}) - H_k \right) (z^{k+1} - z^k) = (G'_{I_+}(\bar{z}))^T (\bar{\mu}_{I_+} - \mu_{I_+}^{k+1}) + o(\|z^k - \bar{z}\|).$$

For any $\zeta \in \ker G'_{I_+}(\bar{z})$, it holds that

$$\langle (G'_{I_+}(\bar{z}))^T (\bar{\mu}_{I_+} - \mu_{I_+}^{k+1}), \zeta \rangle = \langle \bar{\mu}_{I_+} - \mu_{I_+}^{k+1}, G'_{I_+}(\bar{z}) \zeta \rangle = 0,$$

which implies that

$$\Pi(G'_{I_+}(\bar{z}))^T (\bar{\mu}_{I_+} - \mu_{I_+}^{k+1}) = 0.$$

By the Lipschitz-continuity of the projection operator, we then have from (4.10) that

$$\Pi \left(\left(\frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+}) - H_k \right) (z^{k+1} - z^k) \right) = o(\|z^k - \bar{z}\|),$$

and (4.8) follows by noting that

$$\|z^{k+1} - z^k\| \geq \|z^k - \bar{z}\| - \|z^{k+1} - \bar{z}\| = \|z^k - \bar{z}\| - o(\|z^k - \bar{z}\|) \geq \|z^k - \bar{z}\|/2$$

for all k large enough.

We proceed to prove the second assertion. From (4.7), we have that

$$\begin{aligned} (4.11) \quad 0 &= G_{I_+}(z^k) + G'_{I_+}(z^k)(z^{k+1} - z^k) \\ &= G_{I_+}(\bar{z}) + G'_{I_+}(\bar{z})(z^k - \bar{z}) + G'_{I_+}(z^k)(z^{k+1} - z^k) + o(\|z^k - \bar{z}\|) \\ &= G'_{I_+}(\bar{z})(z^{k+1} - \bar{z}) + \eta^k, \end{aligned}$$

where taking into account that $\{z^k\} \rightarrow \bar{z}$,

$$\eta^k = (G'_{I_+}(z^k) - G'_{I_+}(\bar{z}))(z^{k+1} - z^k) + o(\|z^k - \bar{z}\|) = o(\|z^{k+1} - z^k\|) + o(\|z^k - \bar{z}\|).$$

By WLICQ, for every k there exists $v^k \in \mathbf{R}^p$ such that

$$(4.12) \quad G'_{I_+}(\bar{z})v^k = \eta^k, \quad v^k = o(\|z^{k+1} - z^k\|) + o(\|z^k - \bar{z}\|).$$

Denoting $\zeta^k = z^{k+1} - \bar{z} + v^k$, from (4.11) we then have that $G'_{I_+}(\bar{z})\zeta^k = 0$, i.e., $\zeta^k \in \ker G'_{I_+}(\bar{z})$. By (4.9), we have that

$$\begin{aligned} & \left\langle \left(\frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+}) - H_k \right) (z^{k+1} - z^k), \zeta^k \right\rangle \\ &= \left\langle \frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+})(z^{k+1} - \bar{z}), \zeta^k \right\rangle + \langle \bar{\mu}_{I_+} - \mu_{I_+}^{k+1}, G'_{I_+}(\bar{z})\zeta^k \rangle + o(\|\zeta^k\| \|z^k - \bar{z}\|). \end{aligned}$$

And using SSOSC, we further have that there exists $\gamma > 0$ such that

$$\begin{aligned} \gamma \|\zeta^k\|^2 &\leq \left| \left\langle \frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+})\zeta^k, \zeta^k \right\rangle \right| \\ &= \left| \left\langle \frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+})(z^{k+1} - \bar{z} + v^k), \zeta^k \right\rangle \right| \\ &= \left| \left\langle \left(\frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+}) - H_k \right) (z^{k+1} - z^k), \zeta^k \right\rangle + \left\langle \frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+})v^k, \zeta^k \right\rangle \right| \\ &\quad + o(\|\zeta^k\| \|z^k - \bar{z}\|) \\ &= \left| \left\langle \left(\frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+}) - H_k \right) (z^{k+1} - z^k), \zeta^k \right\rangle \right| \\ &\quad + o(\|\zeta^k\| \|z^{k+1} - z^k\|) + o(\|\zeta^k\| \|z^k - \bar{z}\|) \\ &= \left| \left\langle \Pi \left(\frac{\partial \phi}{\partial z}(\bar{z}, \bar{\mu}_{I_+}) - H_k \right) (z^{k+1} - z^k), \zeta^k \right\rangle \right| \\ &\quad + o(\|\zeta^k\| \|z^{k+1} - z^k\|) + o(\|\zeta^k\| \|z^k - \bar{z}\|) \\ &= o(\|\zeta^k\| \|z^{k+1} - z^k\|) + o(\|\zeta^k\| \|z^k - \bar{z}\|), \end{aligned}$$

where the third equality is by (4.12), and the last equality is by (4.8). Dividing both sides in the relation above by $\|\zeta^k\|$, we have that

$$\|\zeta^k\| = o(\|z^{k+1} - z^k\|) + o(\|z^k - \bar{z}\|).$$

Finally,

$$\|z^{k+1} - \bar{z}\| \leq \|\zeta^k\| + \|v^k\| = o(\|z^{k+1} - z^k\|) + o(\|z^k - \bar{z}\|) = o(\|z^{k+1} - \bar{z}\|) + o(\|z^k - \bar{z}\|),$$

from which it follows that $z^{k+1} - \bar{z} = o(\|z^k - \bar{z}\|)$. $\square$

5. Comparison with other Newton-type methods. In this section, we very briefly compare convergence conditions needed for the active-set Newton method described above with conditions required by other Newton-type methods for MCP.

We start with some comments on the special case of KKT. In that case, weak regularity and 2-regularity of Ψ_S are equivalent to semistability of the solution. In the case when this solution is a local minimizer in an optimization problem, this is further equivalent to the uniqueness of the multiplier and the standard second-order sufficient condition for optimality [3]. A detailed comparison with other constraint qualifications and regularity conditions for KKT systems is given in [15] (even though the weak regularity condition itself is different in [15], and the resulting class of active-set methods

is different, most of the comparison comments in [15] still apply here, due to the equivalence with semistability). Summarizing, conditions for the superlinear convergence of the proposed method are strictly weaker than those of SNM for KKT systems, and are the same as conditions required by the sequential quadratic programming method (SQP) in its basic form. Note that SQP subproblems are quadratic programs, while the subproblems of the method proposed here are just systems of linear equations. On the other hand, convergence conditions of SQP can be somewhat weakened, but at the expense of nontrivial modifications of the basic iteration, which come with a computational price [29, 11, 12]. We next focus on the more general case of MCP.

In the case of general MCP, convergence conditions for our method are, of course, again significantly weaker than conditions for SNM [20, 21, 26, 27] applied to Ψ_{FB} or Ψ_{NR} [1, 19, 9] (see also [8, Ch. 9]). Indeed, the latter need BD -regularity of the corresponding function at the solution, which certainly implies the corresponding error bound [25], and thus semistability, but not vice versa. We note, in passing, that BD -regularity of Ψ_{FB} and Ψ_{NR} are not related; i.e., none is weaker or stronger than the other. The fact that BD -regularity of Ψ_{NR} does not imply this property for Ψ_{FB} has been exhibited in [22, Example 2.1]. We have not been able to find in the literature an example showing that BD -regularity of Ψ_{FB} does not imply this property for Ψ_{NR} . So we provide such an example below.

Example 5.1. Let $n = 2$, $F(x) = (x_2, -x_1 + x_2)$, $l_i = 0$, $u_i = +\infty$, $i = 1, 2$. Then $\bar{x} = 0$ is a solution of this NCP, and it can be seen that all matrices in the B -differential of Ψ_{FB} at $\bar{x}$ are nonsingular, while the B -differential of Ψ_{NR} at $\bar{x}$ contains the singular matrix $\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$.

Most remarkably, convergence conditions for our method are also strictly weaker than those for the linearization (Joseph–Newton) method [17, 3], which consists of solving the following subproblems: Given the current iterate $x^k \in \mathbf{R}^n$,

$$\text{find } x^{k+1} \in B \text{ such that } \langle F(x^k) + F'(x^k)(x^{k+1} - x^k), y - x^{k+1} \rangle \geq 0 \quad \forall y \in B.$$

Those types of methods are sometimes referred to as point-based approximation methods [12]. In the case of optimization, they are closely related to SQP. Conditions for the local superlinear convergence of the linearization method are semistability and hemistability [3] of the solution. And in general [3, Remark 2.4], hemistability does not follow from semistability (it does in the case of KKT). Since our method requires something even less than semistability, we conclude that its local convergence properties are significantly stronger.

6. Local numerical experiments. In this section we report numerical experiments on some examples designed to highlight the case where various standard regularity conditions do not hold, and thus SNM-based methods have trouble or converge slowly. This is precisely the case where the switch to our local algorithm can be useful. We have also implemented a globalization of our local algorithm, in the spirit of hybrid versions of SNM/FB methods for complementarity problems based on the Ψ_{FB} reformulation (e.g., the “General Line Search Algorithm” of [22]). Details of this implementation and numerical results on the MCPLIB test problems collection (the newer version of [6]) can be found in the technical report [5]. Before describing our local experiments, we give some general conclusions for the globalized version of [5]. The option of switching to our active-set step never harms too much the global algorithm, though we certainly have to pay some extra price for computing this step at those iterations at which it is eventually rejected (there are safeguards and heuristics that can be used to avoid computing the active-set step too early; see [5]). We

TABLE 1
GNM/AS for Example 6.1.

k	0	1	2	3
$\ x^k - \bar{x}\ $	7.1e-01	1.3e-01	3.7e-03	9.9e-08
$\ \Psi_{FB}(x^k)\ $	8.2e-01	1.6e-02	1.4e-05	9.9e-15
$\frac{\ x^k - \bar{x}\ }{\ x^{k-1} - \bar{x}\ }$		1.8e-01	2.9e-02	2.7e-05

emphasize that the overall goal of the presented approach is not to improve the SNM (or any other algorithm) when it works efficiently but rather to safeguard *fast local convergence* in irregular cases for which other methods do not work well. The results of [5] indicate that this can be achieved without paying a high price over iterations of the globally convergent hybrid SNM/FB method.

We now describe some possible scenarios of the *local* behavior of SNM/FB and GNM/AS by applying them to some small test problems with various combinations of satisfied and violated regularity properties of the solution that have been discussed above. Constructing artificial examples allows us to obtain a rather complete selection of irregular MCPs with “various degrees of irregularity” and to make reliable conclusions about the reasons for the observed performance of the algorithms. This is not possible, for example, for MCPLIB problems, since precise regularity properties of solutions are typically not known. On the other hand, we cannot draw far-reaching conclusions based on small artificial examples. What follows is intended merely to illustrate the theoretical results and comparisons obtained above.

By GNM/AS we mean here the algorithm specified by (3.12), (3.13), with the index sets identified at the starting point (thus, the local properties of this algorithm are given by Theorem 3.6).

By SNM/FB we mean the iterative procedure

$$x^{k+1} = x^k - \Lambda_k^{-1} \Psi_{FB}(x^k), \quad \Lambda_k \in \partial_B \Psi_{FB}(x^k), \quad k = 0, 1, \dots,$$

where the element Λ_k of the B -differential of Ψ_{FB} at x^k is computed by the procedure suggested in [1] (with $z_i = 1 \ \forall i = 1, \dots, n$, in the notation of [1], and the “computer zero” parameter set to 10^{-10}).

The stopping criterion for both methods is

$$(6.1) \quad \|\Psi_{FB}(x^k)\| < 10^{-9}.$$

The first problem is a slight modification of Example 3.5.

Example 6.1. Let $n = 2$, $l_i = 0$, $u_i = +\infty$, $i = 1, 2$, and let $F(x) = ((x_1 - 1)^2, x_1 + x_2 + x_2^2 - 1)$. The point $\bar{x} = (1, 0)$ is the solution of this NCP. Semistability is violated here (and hence BD -regularity for Ψ_{NR} and Ψ_{FB} is violated), while 2-regularity of Ψ_S and weak regularity hold at $\bar{x}$. The starting point is $x^0 = (1.5, -0.5)$, with $\|x^0 - \bar{x}\| \approx 7.1e-01$, $\|\Psi_{FB}(x^0)\| \approx 8.2e-01$, $\det \Lambda_0 \approx 1.4e+00$.

SNM/FB converges in thirteen steps. At termination, $\|x^{13} - \bar{x}\| \approx 3.0e-05$, $\|\Psi_{FB}(x^{13})\| \approx 6.2e-10$, $\det \Lambda_{13} \approx -8.3e-05$ (note that the latter indicates degeneracy). The rate of convergence is linear, with the ratio approaching $1/2$.

The behavior of GNM/AS is reported in Table 1, and it clearly shows fast quadratic convergence.

The next problem is a modification of [15, Example 1].

TABLE 2
 GNM/AS for Example 6.2.

k	0	1	2	3	4	5	6	7
$\ x^k - \bar{x}\ $	2.2e+00	9.0e-01	3.2e-01	7.1e-02	5.0e-03	2.8e-05	9.1e-10	9.3e-19
$\ \Psi_{FB}(x^k)\ $	1.7e+01	4.1e+00	9.3e-01	1.6e-01	1.0e-02	5.7e-05	1.8e-09	1.8e-18
$\frac{\ x^k - \bar{x}\ }{\ x^{k-1} - \bar{x}\ }$		4.0e-01	3.5e-01	2.2e-01	7.1e-02	5.6e-03	3.2e-05	1.0e-09

 TABLE 3
 GNM/AS for Example 6.3.

k	0	1	2	3	4
$\ x^k - \bar{x}\ $	1.0e+00	6.0e-01	2.2e-01	2.7e-03	9.0e-13
$\ \Psi_{FB}(x^k)\ $	9.1e-01	2.2e-01	1.0e-03	2.0e-08	7.4e-37
$\frac{\ x^k - \bar{x}\ }{\ x^{k-1} - \bar{x}\ }$		6.0e-01	3.6e-01	1.3e-02	3.3e-10

Example 6.2. Let $p = m = 2$, $f(z) = (z_1 + z_2)^2/2 + (z_1 + z_2)^3/3$, $G(z) = (z_1, z_2)$, $z \in \mathbf{R}^2$, $\bar{z} = 0$, $\bar{\mu} = 0$. Semistability holds here, but for Ψ_{NR} (and hence for Ψ_{FB}), BD -regularity is violated. The starting point is $z^0 = (1, 2)$, $\mu^0 = (0.01, 0.01)$, with $\|x^0 - \bar{x}\| \approx 2.2e+00$, $\|\Psi_{FB}(x^0)\| \approx 1.7e+01$, $\det \Lambda_0 \approx 4.3e-04$.

The behavior of SNM/FB is as follows: $\det \Lambda_1 \approx 7.3e-10$, $\det \Lambda_2 \approx 0$, but the corresponding linear system is solvable, and the method manages to escape the “bad” region. Specifically, $\det \Lambda_3 \approx 4.3e-16$, while $\det \Lambda_4 \approx 2.5e-00$, and the algorithm converges in seven iterations. At the final step, $\|x^7 - \bar{x}\| \approx 1.0e-16$, $\|\Psi_{FB}(x^7)\| \approx 1.4e-16$, $\det \Lambda_7 \approx 2$. The rate of convergence is superlinear. The behavior of GNM/AS is reported in Table 2, and it also shows the superlinear rate.

Note that while SNM/FB and GNM/AS exhibit similar convergence for this problem, the performance of SNM/FB depends on the specific implementation. In particular, the solution which is produced for a given *degenerate* linear system clearly depends on the linear solver being used. The choice of this solution can affect the overall convergence. Also, in general (this is not the case for this example), when BD -regularity is violated different procedures to compute Λ_k could result in different linear systems some of which can be ill-conditioned close to the solution, preventing fast convergence of SNM/FB.

The next problem is [15, Example 4]. Combined with Example 6.5 below, it seems to indicate that weak regularity is somewhat more important for the success of GNM/AS than 2-regularity.

Example 6.3. Let $p = m = 1$, $f(z) = z^4/4$, $G(z) = z$, $z \in \mathbf{R}$, $\bar{z} = 0$, $\bar{\mu} = 0$. Weak regularity holds here but 2-regularity of Ψ_S does not, and thus semistability is violated (and hence BD -regularity for Ψ_{NR} and Ψ_{FB} is also violated). The starting point is $z^0 = 1$, $\mu^0 = 0.1$, with $\|x^0 - \bar{x}\| \approx 1.0e+00$, $\|\Psi_{FB}(x^0)\| \approx 9.1e-01$, $\det \Lambda_0 \approx 2.7e+00$.

SNM/FB converges in eighteen steps. At termination, $\|x^{18} - \bar{x}\| \approx 6.8e-04$, $\|\Psi_{FB}(x^{18})\| \approx 3.1e-10$, $\det \Lambda_{13} \approx 3.1e-06$. The rate of convergence is linear with ratio approaching $2/3$.

The behavior of GNM/AS is reported in Table 3, and it shows fast quadratic convergence.

The next example is borrowed from [22, Example 2.1].

Example 6.4. Let $n = 2$, $F(x) = (-x_1 + x_2, -x_2)$, $x \in \mathbf{R}^2$, $\bar{x} = 0$. BD -regularity holds for Ψ_{NR} (and hence semistability also holds) but not for Ψ_{FB} . The starting point is $x^0 = (2, 4)$, with $\|x^0 - \bar{x}\| \approx 4.5\text{e}+00$, $\|\Psi_{FB}(x^0)\| \approx 5.8\text{e}+00$, $\det \Lambda_0 = 0$.

Here, SNM/FB fails to make a step. At the same time, GNM/AS terminates after 1 step at the exact solution. The reason for this is that $A_0 = \{1, 2\}$. Thus, the iteration of GNM/AS reduces to identifying the index sets.

Note that the problem in Example 6.4 is actually a linear complementarity problem, that is, NCP with affine F . We point out that in the case of affine F , just one step of GNM/AS gives the exact solution, whenever the index sets are correctly identified. For example, this behavior is observed also for the problem `badfree` from the MCPLIB collection: once x^k is close to the solution $\bar{x} = (0, 0, 0.5, 0.5, 1)$, GNM/AS produces $x^{k+1} = \bar{x}$. At the same time, for x^k close to $\bar{x}$, a degenerate Λ_k is computed, and SNM/FB fails; see [5].

The next example is [15, Example 2] and it shows that both 2-regularity of Ψ_S and weak regularity are important for fast convergence of GNM/AS in general (recall, however, Example 6.3).

Example 6.5. Let $p = m = 2$, $f(z) = z_1^2/2 + z_2^3/3$, $G(z) = (z_1 - z_2^2/2, z_1 + z_2^2/2)$, $z \in \mathbf{R}^2$, $\bar{z} = 0$, $\bar{\mu} = 0$. Semistability is violated (and hence BD -regularity for Ψ_{NR} and Ψ_{FB} is violated). For Ψ_S , 2-regularity holds, but weak regularity does not. The starting point is $z^0 = (0.1, 0.1)$, $\mu^0 = (0.1, 0.1)$, with $\|x^0 - \bar{x}\| \approx 2.0\text{e}-01$, $\|\Psi_{FB}(x^0)\| \approx 1.3\text{e}-01$, $\det \Lambda_0 \approx 2.0\text{e}-01$.

Both SNM/FB and GNM/AS converge in 12 steps, and $\|x^{12} - \bar{x}\| \approx 2.4\text{e}-05$, $\|\Psi_{FB}(x^{12})\| \approx 8.4\text{e}-10$, $\det \Lambda_{12} \approx 2.9\text{e}-04$. The rate of convergence is linear with ratio $1/2$.

Acknowledgments. We thank the two anonymous referees for constructive suggestions. In particular, Proposition 4.1 and Example 4.2 are in response to a question posed by one of the referees.

REFERENCES

- [1] S. C. BILLUPS, *Algorithms for Complementarity Problems and Generalized Equations*, Tech. report 95-14 and Ph.D. thesis, Computer Sciences Department, University of Wisconsin, Madison, WI, 1995.
- [2] S. C. BILLUPS, *A homotopy-based algorithm for mixed complementarity problems*, SIAM J. Optim., 12 (2002), pp. 583–605.
- [3] J. F. BONNANS, *Local analysis of Newton-type methods for variational inequalities and nonlinear programming*, Appl. Math. Optim., 29 (1994), pp. 161–186.
- [4] H. DAN, N. YAMASHITA, AND M. FUKUSHIMA, *A superlinearly convergent algorithm for the monotone nonlinear complementarity problem without uniqueness and nondegeneracy conditions*, Math. Oper. Res., 27 (2002), pp. 743–754.
- [5] A. N. DARYINA, A. F. IZMAILOV, AND M. V. SOLODOV, *Numerical Results for a Globalized Active-Set Newton Method for Mixed Complementarity Problems*, IMPA Tech. report A282, Institute for Pure and Applied Mathematics, Rio de Janeiro, Brazil, 2004.
- [6] S. P. DIRKSE AND M. C. FERRIS, *MCPLIB: A collection of nonlinear mixed complementarity problems*, Optim. Methods Softw., 5 (1995), pp. 319–345.
- [7] F. FACCHINEI, A. FISCHER, AND C. KANZOW, *On the accurate identification of active constraints*, SIAM J. Optim., 9 (1998), pp. 14–32.
- [8] F. FACCHINEI AND J.-S. PANG, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Springer-Verlag, New York, 2003.
- [9] M. C. FERRIS, C. KANZOW, AND T. S. MUNSON, *Feasible descent algorithms for mixed complementarity problems*, Math. Program., 86 (1999), pp. 475–497.
- [10] M. C. FERRIS AND J. S. PANG, *Engineering and economic applications of complementarity problems*, SIAM Rev., 39 (1997), pp. 669–713.

- [11] A. FISCHER, *Modified Wilson's method for nonlinear programs with nonunique multipliers*, Math. Oper. Res., 24 (1999), pp. 699–727.
- [12] A. FISCHER, *Local behavior of an iterative framework for generalized equations with nonisolated solutions*, Math. Program., 94 (2002), pp. 91–124.
- [13] A. F. IZMAILOV AND M. V. SOLODOV, *Error bounds for 2-regular mappings with Lipschitzian derivatives and their applications*, Math. Program., 89 (2001), pp. 413–435.
- [14] A. F. IZMAILOV AND M. V. SOLODOV, *The theory of 2-regularity for mappings with Lipschitzian derivatives and its applications to optimality conditions*, Math. Oper. Res., 27 (2002), pp. 614–635.
- [15] A. F. IZMAILOV AND M. V. SOLODOV, *Karush-Kuhn-Tucker systems: Regularity conditions, error bounds and a class of Newton-type methods*, Math. Program., 95 (2003), pp. 631–650.
- [16] H. JIANG AND D. RALPH, *Global and local superlinear convergence analysis of Newton-type methods for semismooth equations with smooth least squares*, in Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, M. Fukushima and L. Qi, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999, pp. 181–210.
- [17] N. H. JOSEPHY, *Newton's Method for Generalized Equations and the PIES Energy Model*, Ph.D. thesis, Department of Industrial Engineering, University of Wisconsin, Madison, WI, 1979.
- [18] C. KANZOW, *Strictly feasible equation-based methods for mixed complementarity problems*, Numer. Math., 89 (2001), pp. 135–160.
- [19] C. KANZOW AND M. FUKUSHIMA, *Solving box constrained variational inequality problems by using the natural residual with D-gap function globalization*, Oper. Res. Lett., 23 (1998), pp. 45–51.
- [20] B. KUMMER, *Newton's method for nondifferentiable functions*, in Advances in Mathematical Optimization, J. Guddat, B. Bank, H. Hollatz, P. Kall, D. Klatte, B. Kummer, K. Lommalzsch, L. Tammer, M. Vlach, and K. Zimmerman, eds., Vol. 45, Akademie-Verlag, Berlin, 1988, pp. 114–125.
- [21] B. KUMMER, *Newton's method based on generalized derivatives for nonsmooth functions*, in Advances in Optimization, W. Oettli and D. Pallaschke, eds., Springer-Verlag, Berlin, 1992, pp. 171–194.
- [22] T. DE LUCA, F. FACCHINEI, AND C. KANZOW, *A theoretical and numerical comparison of some semismooth algorithms for complementarity problems*, Comput. Optim. Appl., 16 (2000), pp. 173–205.
- [23] Z.-Q. LUO AND J.-S. PANG, *Error bounds for analytic systems and their applications*, Math. Programming, 67 (1994), pp. 1–28.
- [24] J.-S. PANG, *Error bounds in mathematical programming*, Math. Programming, 79 (1997), pp. 299–332.
- [25] J.-S. PANG AND L. QI, *Nonsmooth equations: Motivation and algorithms*, SIAM J. Optim., 3 (1993), pp. 443–465.
- [26] L. QI, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res., 18 (1993), pp. 227–244.
- [27] L. QI AND J. SUN, *A nonsmooth version of Newton's method*, Math. Programming, 58 (1993), pp. 353–367.
- [28] P. TSENG, *Growth behavior of a class of merit functions for the nonlinear complementarity problem*, J. Optim. Theory Appl., 89 (1996), pp. 17–37.
- [29] S. J. WRIGHT, *Superlinear convergence of a stabilized SQP method to a degenerate solution*, Comput. Optim. Appl., 11 (1998), pp. 253–275.

A GENERIC ALGORITHM FOR SOLVING INCLUSIONS*

ADAM B. LEVY†

Abstract. We present a generic algorithm that solves inclusions by generating iterates from successively updated approximating inclusions. Our convergence analysis involves families of generalized rate functions that depend on the parameters in the approximating inclusions, which results in a very flexible format for studying convergence. The algorithm is called generic because it covers many existing methods for solving inclusions, including methods in numerical optimization. In addition to this, our generic algorithm exposes new directions for the development of numerical methods of solving inclusions. As one example, we present a novel generalization to composite optimization of the sequential quadratic programming method for nonlinear programs.

Key words. inclusions, numerical optimization, variational inequalities, variational analysis, sequential quadratic programming, composite optimization

AMS subject classifications. 65K10, 49J53, 65K05

DOI. 10.1137/040603292

1. Introduction. The basic inclusion problem is to solve for the variable x satisfying

$$(1) \quad f(x) \in F(x),$$

where $f : \mathcal{X} \rightarrow \mathcal{Y}$ is a single-valued mapping, $F : \mathcal{X} \rightrightarrows \mathcal{Y}$ is set-valued, $\mathcal{Y}$ is a normed linear space, and $\mathcal{X}$ is a Banach space (some results hold with $\mathcal{X}$ only a normed linear space). This inclusion is universal since the single-valued mapping can be taken as zero; however, it is convenient for the analysis to have both f and F in the expression. This model covers many problems, including systems of (smooth or nonsmooth) equations, optimality conditions, and variational inequalities.

The *generic algorithm* assumes a current iterate $x_k \in \mathcal{X}$ and parameter value u_k in some parameter space $\mathcal{U}$, from which the next iterate x_{k+1} is obtained by solving for $x_{k+1} = x'$ satisfying an approximating inclusion

$$(2) \quad a(u_k, x_k, x') \in A(u_k, x_k, x'),$$

where $a : \mathcal{U} \times \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{Y}$ is a single-valued mapping and $A : \mathcal{U} \times \mathcal{X} \times \mathcal{X} \rightrightarrows \mathcal{Y}$ is a set-valued mapping. The generic algorithm will only be successful in identifying a solution if the iterates x_k converge to some $\bar{x}$ satisfying the original inclusion (1). With this in mind, we will develop a convergence analysis that allows us to deduce when the iterates x_k conforming to the generic algorithm actually converge, and we will quantify the rate. In the presence of such convergence, the question of whether the limit point $\bar{x}$ satisfies the original inclusion (1) depends on the quality of the approximating inclusion (2). As one useful notion of quality, we introduce here the definition of an approximating inclusion (2) being a *valid approximation of (1) at $\bar{x}$ for parameter target $\bar{u}$* if it satisfies the following two criteria:

- (i) a is continuous at $(\bar{u}, \bar{x}, \bar{x})$ and $a(\bar{u}, \bar{x}, \bar{x}) = f(\bar{x})$;
- (ii) A is outer semicontinuous at $(\bar{u}, \bar{x}, \bar{x})$ and $A(\bar{u}, \bar{x}, \bar{x}) \subseteq F(\bar{x})$.

*Received by the editors January 16, 2004; accepted for publication (in revised form) June 23, 2004; published electronically December 30, 2004.

<http://www.siam.org/journals/siopt/15-2/60329.html>

†Department of Mathematics, Bowdoin College, Brunswick, ME 04011 (alevy@bowdoin.edu).

Recall from [12, Definition 5.4] that a set-valued mapping $A : \mathcal{X} \rightrightarrows \mathcal{Y}$ is outer semi-continuous at $\bar{x}$ if every outer limit point $\bar{y}$ is in the set $A(\bar{x})$. An outer limit point $\bar{y}$ is defined by the existence of sequences $x \rightarrow \bar{x}$ and $y \in A(x)$ with $y \rightarrow \bar{y}$. Weaker conditions than these would suffice to deduce that the limit point $\bar{x}$ is a solution to the original inclusion (1); however, our criteria are weak enough to include all of the specific examples of the generic algorithm with which we are concerned.

A compact form of the updating rule implicit in (2) is $x_{k+1} \in G(u_k, x_k)$ in terms of the *generic iteration mapping* $G : \mathcal{U} \times \mathcal{X} \rightrightarrows \mathcal{X}$ defined by

$$(3) \quad G(u, x) := (A(u, x, \cdot) - a(u, x, \cdot))^{-1}(0)$$

in terms of an inverse set-valued mapping (note that any set-valued mapping $M : \mathcal{X} \rightrightarrows \mathcal{Y}$ has an associated inverse mapping $M^{-1} : \mathcal{Y} \rightrightarrows \mathcal{X}$ defined by $x \in M^{-1}(y)$ if and only if $y \in M(x)$). After the new iterate x_{k+1} is chosen, a new parameter u_{k+1} is chosen from some *parameter set* $U \subseteq \mathcal{U}$. In some cases, the new parameter may depend explicitly on the current information, which we denote by $u_{k+1} \in \hat{U}(u_k, x_k, x_{k+1})$ in terms of a *parameter update mapping* $\hat{U} : \mathcal{U} \times \mathcal{X} \times \mathcal{X} \rightrightarrows U$.

There is a vast amount of literature on the particular problems covered by the inclusion (1), as well as on the solution methods associated with those particular problems. Examples include the proximal point algorithm [10] and its various generalizations (e.g., see [6]), as well as generalized Newton methods for inclusions [3], [1], [14]. Another approach to solving inclusions relies on the possibility of reformulating them as nonsmooth equations $f(x) = 0$ (which can be done with variational inequalities, for instance). Then, methods for solving nonsmooth equations can be applied to the reformulated inclusion (e.g., see [8]). In addition, many numerical methods for optimization can be thought of as inclusion-solving methods where the inclusions describe optimality conditions. These include discretization schemes in semi-infinite programming (e.g., [13]), trust region methods, interior point methods, and penalty/barrier methods in general. We call our algorithm *generic* because it covers all of these methods through various particular formulations of the mappings a and A in (2).

In this paper, we will focus our attention on detailing two major examples of the generic algorithm. The first of these is a novel generalization to composite optimization problems of the famous sequential quadratic programming (SQP) algorithm for nonlinear programs. The novelty of our generalized primal SQP method stems from the fact that it is not derived from a Newton-type method applied to primal-dual optimality conditions. Instead, our method is derived directly from primal optimality conditions, and the dual variables are treated as abstract parameters having a much less restricted connection to the problem than is traditional. The other major example we explore is a generalized quasi-Newton method developed in [3] (and earlier in [2]) for solving general inclusions. The approximating inclusion in this case uses the set-valued component of the original inclusion (1) as its own approximation, but uses a quasi-Newton approximation to the single-valued component. The resulting generalized quasi-Newton method covers many important cases of inclusions where the set-valued component has relatively nice structure (e.g., variational inequalities), but it does not cover our generalized primal SQP method. Moreover, the generalized quasi-Newton method from [3] is essentially constructed for the inclusion problem parameterized in one very simple way (though the analysis in [3] is adapted in one case to cover the proximal point algorithm where a parameter appears in a different way). Our generic algorithm is developed from the outset for inclusions having very

general parametric dependence. The examples we cover here do not begin to exhaust the possibilities for the generic algorithm, but should give the reader a sense of how the conditions underlying our analysis of the generic algorithm work out in two cases of interest.

To analyze the convergence of the generic algorithm, we first compare the differences $\|x_{k+1} - \bar{x}\|$ and $\|x_k - \bar{x}\|$ between some presumed target $\bar{x}$ and consecutive iterates. Traditionally, this type of analysis leads to results involving a “ Q -convergence rate” where the distances between the target and consecutive iterates are related by

$$(4) \quad \|x_{k+1} - \bar{x}\| \leq \epsilon_k \|x_k - \bar{x}\|^\gamma$$

in terms of some sequence of scalars $\{\epsilon_k\} \in \mathbb{R}^+$ and a rate constant $\gamma \geq 1$. For instance, there is a quadratic Q -convergence rate if $\gamma = 2$ and the sequence $\{\epsilon_k\}$ is bounded, a linear Q -convergence rate if $\gamma = 1$ and the sequence $\{\epsilon_k\}$ is bounded below 1, and a superlinear Q -convergence rate if $\gamma = 1$ and $\epsilon_k \rightarrow 0$. In the present paper, we obtain convergence results for the generic algorithm with a generalized Q -convergence rate where the function $(\cdot)^\gamma$ on the right side of the bound (4) is replaced by a general “convergence function”: A function $\psi : \mathbb{R} \rightarrow \mathbb{R}$ is called a *convergence function* if it is nondecreasing on $\mathbb{R}^+$ with $\psi(0) = 0$. Parameters like the ϵ_k in (4) are accommodated in our scheme by using parameterized families of convergence functions. In particular then, the bound (4) can be rewritten as

$$\|x_{k+1} - \bar{x}\| \leq \psi_{\epsilon_k}(\|x_k - \bar{x}\|)$$

for the family of convergence functions $\psi_\epsilon(r) := \epsilon r^\gamma$ parameterized by $\epsilon \in \mathbb{R}^+$. In our general analysis, it will be typical for the families of convergence functions $\psi_{(u,u')}$ to depend on parameter pairs (u, u') from the product set $U \times U$ for some parameter set $U \subseteq \mathcal{U}$. We will also use parameterized families of *bivariate convergence functions* $\phi_{(u,u')} : \mathbb{R}^2 \rightarrow \mathbb{R}$ which are defined by the following sublinearity relationship:

$$\phi_{(u,u')}(r, r') \leq \psi_{(u,u')}(r) + \alpha r' \text{ for all } (u, u') \in U \times U \text{ and } (r, r') \in \mathbb{R}^+ \times \mathbb{R}^+$$

in terms of a family of convergence functions $\psi_{(u,u')}$ and a fixed sublinearity constant $\alpha \in (0, \infty)$. One example of a bivariate convergence function that we will see in what follows is $\phi(r, r') := \alpha \sqrt{r^2 + (r')^2}$, which evidently satisfies the sublinearity relationship with convergence function $\psi(r) := \alpha r$. As is true in the traditional case, our Q -convergence results involve conditions on the problem data at the target $\bar{x}$.

To complement our Q -convergence analysis, we also analyze the convergence of the generic algorithm by comparing the differences $\|x_k - x_{k-1}\|$ and $\|x_{k+1} - x_k\|$ between consecutive iterates. This analysis ultimately leads to conditions ensuring the convergence of the sequence of iterates $\{x_k\}$ to a target $\bar{x}$. Moreover, we establish a generalized “ R -convergence rate” by describing a sequence of scalars $\{s_k\}$ converging to zero and satisfying

$$(5) \quad \|x_k - \bar{x}\| \leq s_k \text{ for all } k.$$

Of course, for any sequence $\{x_k\}$ converging to $\bar{x}$, the basic sequence $s_k := \|x_k - \bar{x}\|$ satisfies this bound; however, we are interested in more interesting bounding sequences which allow us to deduce something about the convergence rate without actually generating the entire sequence of iterates. For instance, the traditional notion of an R -convergence rate of $\gamma \geq 1$ entails the bound (5) holding with $s_k := c\xi^{\gamma^k}$ for

some constants $c \in \mathbb{R}^+$ and $\xi \in [0, 1)$. The sequence we describe for the generic algorithm generalizes the traditional notion, and as is typical in the traditional case, our R -convergence results rely on conditions involving only the first two iterates x_0 and x_1 .

In section 2, we present two examples which we will study throughout the paper to illustrate the general results. Section 3 contains our main convergence analysis results for the generic algorithm, and in section 4 we develop connections between generalized continuity properties and the “nonsingularity” conditions on which our convergence analysis depends. Finally, in section 5, we return to the two examples from section 2 and give detailed Q - and R -convergence results for both examples.

2. Two examples. In this section, we introduce two examples which we will explore throughout this paper. The first is a novel generalization of the classic SQP method for nonlinear programs and comes from a general composite minimization problem:

$$(6) \quad \min f_0(x) + \Theta(f_1(x), f_2(x), \dots, f_m(x)) \text{ over } x \in X$$

for some closed set $X \subseteq \mathbb{R}^n$, some $\mathcal{C}^2$ functions $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$, and some, possibly nonsmooth, lower semicontinuous extended real-valued function $\Theta : \mathbb{R}^m \rightarrow \mathbb{R} \cup \{\infty\}$. This model includes not only the standard nonlinear programming problem

$$(7) \quad \min f_0(x) \text{ over } x \text{ with } f_i(x) \leq 0 \text{ for } i \in \mathcal{I}, \text{ and } f_i(x) = 0 \text{ for } i \in \mathcal{E}$$

by defining

$$\Theta(y) = \begin{cases} 0 & \text{if } y_i \leq 0 \text{ for } i \in \mathcal{I}, \text{ and } y_i = 0 \text{ for } i \in \mathcal{E}, \\ \infty & \text{otherwise,} \end{cases}$$

and $X := \mathbb{R}^n$, but also covers many other optimization problems (for example, see [11], where Θ is assumed to be expressible in a particular way and the problem is called an extended nonlinear programming problem). If we define

$$g(x) := (f_1(x), f_2(x), \dots, f_m(x)),$$

then solutions to the composite optimization problem satisfy the stationary point inclusion

$$(8) \quad -\nabla f_0(x) \in \partial(\Theta \circ g)(x) + N_X(x),$$

where $\partial(\Theta \circ g)(x)$ denotes the set of subgradients at x associated with the composite function $\Theta \circ g$ and $N_X(x)$ denotes the cone of vectors normal to the set X at x (see [12] for a thorough discussion of subgradients and normal cones, including their role in the stationary point inclusion). This inclusion obviously fits our general model (1) with

$$\begin{aligned} f(x) &:= -\nabla f_0(x), \\ F(x) &:= \partial(\Theta \circ g)(x) + N_X(x). \end{aligned}$$

We can use the classical SQP algorithm for standard nonlinear programs as a model for the generic algorithm applied to the stationary point inclusion (8) associated with the general composite optimization problem (6). To follow this model, we

linearize the mappings $g(\cdot)$ and $-\nabla f_0(\cdot)$ at x , and add a (parameterized) second-order term for g . The result is the approximating inclusion

$$(9) \quad -\nabla f_0(x_k) - \nabla^2 f_0(x_k)[x' - x_k] \in \partial(\Theta \circ \text{lin}_{x_k} g)(x') + N_X(x') + \nabla^2 g_{u_k}(x_k)[x' - x_k],$$

where the linearization $\text{lin}_x g$ of g at x is defined by

$$\text{lin}_x g(x') := \nabla g(x)[x' - x] + g(x)$$

and $g_u : \mathbb{R}^n \rightarrow \mathbb{R}$ is some parameterized $\mathcal{C}^2$ replacement for g (in the case of classical sequential quadratic programming for standard nonlinear programs, $g_u(x) = u \cdot g(x)$ for the dual variable $u \in \mathbb{R}^m$). The approximating inclusion (9) fits the model of our generic algorithm (2) where the mappings a and A are defined by

$$(10) \quad \begin{aligned} a(u, x, x') &:= -\nabla f_0(x) - \nabla^2 f_0(x)[x' - x], \\ A(u, x, x') &:= \partial(\Theta \circ \text{lin}_x g)(x') + N_X(x') + \nabla^2 g_u(x)[x' - x]. \end{aligned}$$

The generalized primal SQP method. Given a current iterate x_k and parameter value u_k , choose the next iterate $x_{k+1} = x'$ by solving the approximating inclusion

$$-\nabla f_0(x_k) - \nabla^2 f_0(x_k)[x' - x_k] \in \partial(\Theta \circ \text{lin}_{x_k} g)(x') + N_X(x') + \nabla^2 g_{u_k}(x_k)[x' - x_k].$$

One important issue to consider is the implementability of the generalized primal SQP method. The approximating inclusion defining the generalized primal SQP method evidently encodes optimality conditions for the approximate minimization problem:

$$\min \nabla f_0(x_k)[x'] + \frac{1}{2}[x' - x_k]^T \nabla^2 (f_0(x_k) + g_{u_k}(x_k))[x' - x_k] + \Theta(\text{lin}_{x_k} g(x')) \text{ over } x' \in X,$$

so the generalized primal SQP method can be implemented by solving this minimization problem. In the case of a standard nonlinear program (7), this minimization problem is a quadratic program which is solved easily. In general, the structure of the function Θ and the set X dictate the nature of the approximation problem, and consequently affect the implementability of the generalized primal SQP method. In particular, this means that the abstract constraint set X in the original minimization problem (6) must represent relatively simple constraints (e.g., box constraints or any other constraints defined by linear functions).

Another issue we need to explore is the validity of the approximating inclusion defining the generalized primal SQP method, which we do in the next proposition.

PROPOSITION 2.0.1 (validity of the generalized primal SQP method). *Assume that the mapping $(u, x) \mapsto \nabla^2 g_u(x)$ is continuous at $(\bar{u}, \bar{x})$, the (set-valued) mapping $(x, x') \mapsto \partial(\Theta \circ \text{lin}_x g)(x') + N_X(x')$ is outer semicontinuous at $(\bar{x}, \bar{x})$, and the function Θ is convex and its effective domain (i.e., the points y with $\Theta(y) < \infty$) cannot be separated from the range of the linearization $\text{lin}_{\bar{x}} g$ of g at $\bar{x}$. It follows that the approximating inclusion (9) defining the generalized primal SQP method is a valid approximation of (8) at $\bar{x}$ for parameter target $\bar{u}$.*

Proof. According to the definitions (10) and under the differentiability assumption on f_0 , the first criterion for validity evidently holds in this case.

The outer semicontinuity at $(\bar{u}, \bar{x}, \bar{x})$ of the mapping A in this case follows from [12, Proposition 5.51], since we have assumed that the mapping $(u, x) \mapsto \nabla^2 g_u(x)$ is

continuous at $(\bar{u}, \bar{x})$ and the (set-valued) mapping $(x, x') \mapsto \partial(\Theta \circ \text{lin}_x g)(x') + N_X(x')$ is outer semicontinuous at $(\bar{x}, \bar{x})$. After evaluating $A(\bar{u}, \bar{x}, \bar{x})$, we see that the inclusion in the second criterion for validity in this case follows from the inclusion

$$(11) \quad \partial(\Theta \circ \text{lin}_{\bar{x}} g)(\bar{x}) \subseteq \partial(\Theta \circ g)(\bar{x}).$$

According to [12, Theorem 10.6] and since Θ is convex, the inclusion holds that

$$(12) \quad \nabla g(\bar{x})^* \partial \Theta(g(\bar{x})) \subseteq \partial(\Theta \circ g)(\bar{x})$$

as well as the first equality in

$$(13) \quad \partial(\Theta \circ \text{lin}_{\bar{x}} g)(\bar{x}) = \nabla g(\bar{x})^* \partial \Theta(\text{lin}_{\bar{x}} g(\bar{x})) = \nabla g(\bar{x})^* \partial \Theta(g(\bar{x}))$$

under the constraint qualification that the effective domain of Θ cannot be separated from the range of $\text{lin}_{\bar{x}} g$ (the second equality follows immediately from the definition of $\text{lin}_{\bar{x}} g$). Notice that, according to [12, Theorem 10.6], this same constraint qualification also implies that the inclusion in (12) is an identity. The desired inclusion (11) (which actually holds as an identity in this case) follows from the combination of the inclusion (12) and the identity (13). $\square$

Remark. We can apply [12, Propositions 5.51 and 6.6] to ensure that the mapping $(x, x') \mapsto \partial(\Theta \circ \text{lin}_x g)(x') + N_X(x')$ is outer semicontinuous at $(\bar{x}, \bar{x})$, as long as the mapping $(x, x') \mapsto \partial(\Theta \circ \text{lin}_x g)(x')$ is outer semicontinuous at $(\bar{x}, \bar{x})$ and the mapping $(x, x', y) \mapsto \partial(\Theta \circ \text{lin}_x g)(x') \cap (y - N_X(x'))$ is locally bounded.

Notice that for the generalized primal SQP method presented here, the parameter updating is not restricted to one particular scheme, as it is in the classical SQP method for standard nonlinear programs. The parameter update that best mimics the classical SQP method is $u' \in \widehat{U}(u, x, x')$ for

$$(14) \quad \widehat{U}(u, x, x') = \{u' \in \partial \Theta(\text{lin}_x g(x')) \mid -\nabla_x L(x, u') - \nabla_{xx}^2 L(x, u)[x' - x] \in N_X(x')\}$$

in terms of the generalized Lagrangian function $L : \mathbb{R}^n \times \mathcal{U} \rightarrow \mathbb{R}$ defined by $L(x, u) := f_0(x) + g_u(x)$. Under the assumptions of Proposition 2.0.1 and according to its proof, the set of subgradients $\partial(\Theta \circ g)(x)$ equals the set $\nabla g(x)^T \partial \Theta(g(x))$, in which case there is a useful primal-dual formulation of the problem where the primal variables (now labeled x_1) are paired with dual variables x_2 in a single variable $x = (x_1, x_2)$. Assuming the replacement function is the traditional one $g_u(x) := u \cdot g(x)$, we can recast the stationary point inclusion (8) as

$$(15) \quad -\nabla L(x) \in N_X(x_1) \times -(\partial \Theta)^{-1}(x_2).$$

Notice that the dual variable x_2 in this case plays the same role as the parameter u in the preceding formulation.

The inclusion (15) fits our general model (1) with

$$\begin{aligned} f(x) &:= -\nabla L(x), \\ F(x) &:= N_X(x_1) \times -(\partial \Theta)^{-1}(x_2). \end{aligned}$$

Under these circumstances, the parameter update mapping (14) and the approximating inclusion defined by (9) are together equivalent to the single inclusion

$$-\nabla L(x) - \nabla^2 L(x)[x' - x] \in N_X(x'_1) \times -(\partial \Theta)^{-1}(x'_2).$$

This inclusion fits our generic model (2) with

$$(16) \quad \begin{aligned} a(u, x, x') &:= -\nabla L(x) - \nabla^2 L(x)[x' - x], \\ A(u, x, x') &:= N_X(x'_1) \times -(\partial\Theta)^{-1}(x'_2). \end{aligned}$$

Note that there are no explicit parameters in this case since we have absorbed the dual variables into the variable x . We refer to this formulation as the generalized *primal-dual* SQP algorithm to distinguish it from the previous *primal* version, where the dual variables played the role of parameters.

Of course, according to the preceding discussion, the standard parameter update mapping (14) and the primal variable update resulting from (10) are together equivalent to the variable update resulting from (16) in the primal-dual formulation as long as the traditional replacement function $g_u(x) := u \cdot g(x)$ is used. However, by treating the dual variables as parameters in the primal formulation of this algorithm, we can use parameter updating rules and replacements g_u different from the traditional ones, including, for example, penalty terms like $u\|g(x)\|^2$ for a scalar $u \in \mathbb{R}^+$ in the case of equality constraints. Both of these novelties open up interesting possibilities for generalized primal SQP algorithms that depart from the traditional one.

Instead of directly discussing the validity of the generalized primal-dual SQP algorithm, we turn now to a generalization of it that was studied in [3]. Dontchev [3] studied a method for solving inclusions of the form

$$(17) \quad \bar{v} - f(x) \in F(x)$$

with $f \in \mathcal{C}^2$, which generalizes the primal-dual SQP method we have just described. Notice that Dontchev's inclusion (17) is covered by our general inclusion (1) if our $f(x)$ is identified with Dontchev's $\bar{v} - f(x)$.

The generalized quasi-Newton method. Given a current iterate x_k and parameter vector-matrix pair $u_k := (v_k, H_k)$, choose the next iterate $x_{k+1} = x'$ by solving the approximating inclusion

$$(18) \quad v_k - f(x_k) - H_k[x' - x_k] \in F(x').$$

This fits the model for our generic algorithm (2) with

$$(19) \quad \begin{aligned} a(u, x, x') &:= v - f(x) - H[x' - x], \\ A(u, x, x') &:= F(x'). \end{aligned}$$

The generalized quasi-Newton method restricts the set-valued components in the algorithm to depend only on x' , which certainly covers the primal-dual version of the generalized primal SQP method (16). This is apparent in our presentation from the dependence of the set-valued mapping A on x' only. However, the generalized quasi-Newton method does not cover the primal version of this algorithm (10), where the presence of the set-valued component $\partial(\Theta \circ \text{lin}_x g)(x')$ depending on both x and x' means that the set-valued mapping A has to depend on both of these variables too.

Validity in the case of the generalized quasi-Newton method is quite straightforward to establish. The single-valued mapping a is clearly continuous with $a(\bar{u}, \bar{x}, \bar{x}) = \bar{v} + f(\bar{x})$, and the set-valued mapping A satisfies $A(\bar{u}, \bar{x}, \bar{x}) = F(\bar{x})$, so both criteria for validity are satisfied if F is outer semicontinuous at $\bar{x}$ (which was assumed throughout [3]). In the special case (covered by the generalized quasi-Newton method) of the generalized primal-dual SQP method, the set-valued mapping F satisfies

$$(20) \quad F(x') := N_X(x'_1) \times -(\partial\Theta)^{-1}(x'_2).$$

Since the normal cone mapping $N_X(x'_1)$ is outer semicontinuous at $(\bar{x}_1)$ (cf. [12, Proposition 6.6]), the mapping F in this case is outer semicontinuous at $\bar{x}$ exactly when the partial derivative mapping $\partial\Theta(\cdot)$ is outer semicontinuous at $g(\bar{x}_1)$ (cf. [12, Theorem 5.7]). When the underlying problem is a standard nonlinear program (7), the graph of $\partial\Theta(\cdot)$ is the set of pairs (y, x'_2) satisfying

$$(21) \quad \begin{aligned} y_i &\leq 0 \text{ for all } i \in \mathcal{I}, \\ y_i &= 0 \text{ for all } i \in \mathcal{E}, \\ (x'_2)_i &\geq 0 \text{ for all } i \in \mathcal{I}, \\ (x'_2)_i \cdot y_i &= 0 \text{ for all } i \in \mathcal{I} \cup \mathcal{E}, \end{aligned}$$

in which case the partial derivative mapping $\partial\Theta(\cdot)$ is outer semicontinuous (globally). More generally, $\partial\Theta(\cdot)$ is outer semicontinuous whenever Θ is convex (cf. [9, Theorem 24.4]). Our discussion of the generalized quasi-Newton method leads directly to the following proposition.

PROPOSITION 2.0.2 (validity of the generalized quasi-Newton method). *The approximating inclusion (18) defining the generalized quasi-Newton method is a valid approximation of (17) at $\bar{x}$ for parameter target $\bar{u} := (\bar{v}, \bar{H})$ as long as the mapping F is outer semicontinuous at $\bar{x}$. Moreover, in the special case of the generalized primal-dual SQP method, the mapping F (20) is outer semicontinuous at $\bar{x}$ whenever the function Θ is convex.*

There are typically several different, but equivalent, versions of any approximating inclusion (2) that can be obtained by simple rearrangements of terms on the left and right sides. The next lemma describes changes that can be made without affecting the validity of the approximating inclusion.

LEMMA 2.1. *For a single-valued mapping $a_2 : \mathcal{U} \times \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{Y}$ that is continuous at $(\bar{u}, \bar{x}, \bar{x})$ with $a_2(\bar{u}, \bar{x}, \bar{x}) = 0$, the following are equivalent:*

- *The approximating inclusion (2) is solved by x' and moreover is a valid approximation of (1) at $\bar{x}$ for parameter target $\bar{u}$.*
- *The approximating inclusion*

$$(22) \quad (a + a_2)(u_k, x_k, x_{k+1}) \in (A + a_2)(u_k, x_k, x')$$

is solved by x' and moreover is a valid approximation of (1) at $\bar{x}$ for parameter target $\bar{u}$.

Proof. It is easy to see that any solution to (22) is a solution to (2), and vice versa. The first condition for the validity of (22) is equivalent to the first condition for the validity of (2) since a_2 is continuous at $(\bar{u}, \bar{x}, \bar{x})$ with $a_2(\bar{u}, \bar{x}, \bar{x}) = 0$. Likewise, the second condition for the validity of (22) is equivalent to the second condition for the validity of (2) since the sum of an outer semicontinuous set-valued mapping and an continuous single-valued mapping is outer semicontinuous. $\square$

Lemma 2.1 can be applied to both examples we consider in this paper, which we describe in the following pair of corollaries.

COROLLARY 2.0.1. *The generalized primal SQP method can be defined instead by the approximating inclusion*

$$(23) \quad -\nabla f_0(x_k) \in \partial(\Theta \circ \text{lin}_{x_k} g)(x') + N_X(x') + \nabla^2 L_{u_k}(x_k)[x' - x_k]$$

defined in terms of the parameterized family of Lagrangian functions $L_u(x) := f_0(x) + g_u(x)$.

Moreover, under the same assumptions as Proposition 2.0.1, the approximating inclusion (23) is a valid approximation of (8) at $\bar{x}$ for parameter target $\bar{u}$.

Proof. This follows immediately from Proposition 2.0.1 and Lemma 2.1 with

$$a_2(u, x, x') := \nabla^2 f_0(x)[x' - x]. \quad \square$$

COROLLARY 2.0.2. *The approximating inclusion (18) defining the generalized quasi-Newton method can be replaced by the approximating inclusion*

$$(24) \quad v_k - f(x_k) \in F(x') + H_k[x' - x_k],$$

which is a valid approximation of (17) at $\bar{x}$ for parameter target $\bar{u} := (\bar{v}, \bar{H})$ as long as the mapping F is outer semicontinuous at $\bar{x}$.

Proof. This follows immediately from Proposition 2.0.2 and Lemma 2.1 with

$$a_2(u, x, x') := H[x' - x]. \quad \square$$

3. Convergence analysis. To motivate the conditions appearing in our main results, we will now describe how the classical Newton's method is covered by the generic algorithm. Newton's method for solving a system of (unparameterized) $\mathcal{C}^1$ equations $f_1(x) = 0, f_2(x) = 0, \dots, f_n(x) = 0$ is one very basic particular example that fits the model of the generic algorithm. The system of equations can be represented as an inclusion of the form (1) by defining $F(x) = 0$ and

$$f(x) := (f_1(x), f_2(x), \dots, f_n(x)).$$

Newton's method gives new iterates x_{k+1} from current iterates x_k by solving the equation

$$f(x_k) + \nabla f(x_k)[x' - x_k] = 0$$

for $x_{k+1} = x'$. There are several different ways that this equation can fit the model of our generic algorithm, but one way that illustrates the basic conditions behind our general analysis has

$$a(u, x, x') := f(x) \quad \text{and} \quad A(u, x, x') := \nabla f(x)[x - x'].$$

The generic iteration mapping for Newton's method works out as

$$G(u, x) := (A(u, x, \cdot) - a(u, x, \cdot))^{-1}(0) = x + \nabla f(x)^{-1}[-f(x)],$$

which should be familiar to the reader.

3.1. Q-convergence. Traditionally, Newton's method is applied under the assumption of the "target nonsingularity condition" that $\nabla f(\bar{x})$ is invertible at the solution $\bar{x}$. Since the component functions f_i are continuously differentiable at $\bar{x}$, the invertibility of $\nabla f(\bar{x})$ implies that $\nabla f(x)$ is invertible for all x near $\bar{x}$ with $\|\nabla f(x)^{-1}\| \leq \alpha$ for some $\alpha \in (0, \infty)$. The generic iteration mapping in this case satisfies

$$x' \in G(u, x) \iff x' = x + \nabla f(x)^{-1}[-f(x)] \iff x' - \bar{x} = \nabla f(x)^{-1}[\nabla f(x)[x - \bar{x}] - f(x)].$$

It follows that the target nonsingularity condition for Newton's method implies the bound

$$(25) \quad \|x' - \bar{x}\| \leq \alpha \|f(x) - \nabla f(x)[x - \bar{x}]\|$$

for all x near $\bar{x}$. It is well known that in the presence of this nonsingularity, Newton's method converges at a quadratic Q -convergence rate when initiated close enough to the solution $\bar{x}$. The convergence analysis to prove this revolves around the fact that the term $\sigma(x, \bar{x}) := f(x) - \nabla f(x)[x - \bar{x}]$ appearing on the right side of (25) satisfies the approximation bound

$$(26) \quad \|\sigma(x, \bar{x})\| \leq \epsilon \|x - \bar{x}\|^2$$

for some fixed $\epsilon \geq 0$ and all x near $\bar{x}$, since the component functions f_i are continuously differentiable at $\bar{x}$. The combination of the approximation bound (26) and the target nonsingularity bound (25) gives the familiar statement of quadratic Q -convergence for Newton's method:

$$\|x' - \bar{x}\| \leq \alpha \epsilon \|x - \bar{x}\|^2.$$

Our main Q -convergence result for the generic algorithm relies on a pair of conditions that are essentially parameterized analogues of the target nonsingularity bound (25) and the approximation bound (26) for the classical Newton's method. We state the result in terms of a general set-valued mapping $X : \mathcal{U} \times \mathcal{X} \rightrightarrows \mathcal{X}$ and the closed ball $\mathbb{B}(\bar{x}; \epsilon_x) \subseteq \mathcal{X}$ defined for some $\epsilon_x \geq 0$ by

$$\mathbb{B}(\bar{x}; \epsilon_x) := \{x \in \mathcal{X} \mid \|x - \bar{x}\| \leq \epsilon_x\}.$$

Q -convergence condition for X . There exist scalars $\alpha \in (0, \infty)$ and $\epsilon_x \geq 0$, a parameter set $U \subseteq \mathcal{U}$, a parameter target $\bar{u} \in U$, and a single-valued mapping $\sigma : \mathcal{U} \times \mathcal{U} \times \mathcal{X} \times \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{Y}$ such that the following pair of conditions hold:

- (Target σ -approximation) There exists a family of bivariate convergence functions $\phi_{(u, u')}^1$ parameterized by $(u, u') \in U \times U$ and with sublinearity constant $1/(2\alpha)$, such that any trio $(u, x, x') \in U \times \mathbb{B}(\bar{x}; \epsilon_x) \times X(u, x)$ satisfies

$$\|\sigma(u, \bar{u}, x, x', \bar{x})\| \leq \phi_{(u, \bar{u})}^1(\|x - \bar{x}\|, \|x' - \bar{x}\|).$$

- (Target nonsingularity) There exists a family of bivariate convergence functions $\phi_{(u, u')}^2$ parameterized by $(u, u') \in U \times U$ and with sublinearity constant α , such that any trio $(u, x, x') \in U \times \mathbb{B}(\bar{x}; \epsilon_x) \times X(u, x)$ satisfies

$$\|x' - \bar{x}\| \leq \phi_{(u, \bar{u})}^2(\|x - \bar{x}\|, \|\sigma(u, \bar{u}, x, x', \bar{x})\|).$$

Our main Q -convergence result for the generic algorithm relies on the “target appropriateness” of the sequence of parameters when there is a parameter target $\bar{u} \in U$. In this case, we say the sequence of parameters $u_k \in U$ is *target appropriate* for $(\psi_{(u, \bar{u})}, r_0)$ with target $\bar{u}$ if the sequence of scalars $\{r_k\}$ defined by

$$r_0, r_k := \psi_{(u_{k-1}, \bar{u})}(r_{k-1}) \text{ for } k = 1, 2, \dots$$

converges to zero.

THEOREM 3.1 (Q -convergence of the generic algorithm). *Assume that the Q -convergence condition holds for the generic iteration mapping (3), and assume that the family of convergence functions $\psi_{(u, u')}$ defined by*

$$(27) \quad \psi_{(u, u')}(r) := 2\alpha \psi_{(u, u')}^1(r) + 2\psi_{(u, u')}^2(r)$$

together with the initial iterate x_0 satisfies

$$(28) \quad \psi_{(u, \bar{u})}(\|x_0 - \bar{x}\|) \leq \|x_0 - \bar{x}\| \leq \epsilon_x \text{ for all } u \in U.$$

It follows that any sequence initiated from x_0 and conforming to the generic algorithm $x_{k+1} \in G(u_k, x_k)$ with a sequence of parameters $\{u_k\} \subseteq U$ satisfies

$$(29) \quad \|x_{k+1} - \bar{x}\| \leq \psi_{(u_k, \bar{u})}(\|x_k - \bar{x}\|).$$

Under these circumstances, if the sequence of parameters $\{u_k\}$ is target appropriate for $(\psi_{(u, \bar{u})}, \|x_0 - \bar{x}\|)$ with target $\bar{u}$, then the sequence $\{x_k\}$ converges to $\bar{x}$.

If in addition $u_k \rightarrow \bar{u}$ and the approximating inclusion (2) is a valid approximation of (1) at $\bar{x}$ for parameter target $\bar{u}$, then $\bar{x}$ is a solution to the original inclusion (1).

Proof. According to the second inequality in the bound (28), we know that $x_0 \in \mathbb{B}(\bar{x}; \epsilon_x)$. Since $u_0 \in U$ and $x_1 \in G(u_0, x_0)$ we can apply the target nonsingularity bound with $(u, x, x') := (u_0, x_0, x_1)$ to deduce that

$$\begin{aligned} \|x_1 - \bar{x}\| &\leq \phi_{(u_0, \bar{u})}^2(\|x_0 - \bar{x}\|, \|\sigma(u_0, \bar{u}, x_0, x_1, \bar{x})\|) \\ &\leq \psi_{(u_0, \bar{u})}^2(\|x_0 - \bar{x}\|) + \alpha \|\sigma(u_0, \bar{u}, x_0, x_1, \bar{x})\|, \end{aligned}$$

where the second inequality comes from the sublinearity of ϕ^2 . Applying the target σ -approximation bound to this inequality, we get

$$\begin{aligned} \|x_1 - \bar{x}\| &\leq \psi_{(u_0, \bar{u})}^2(\|x_0 - \bar{x}\|) + \alpha \phi_{(u_0, \bar{u})}^1(\|x_0 - \bar{x}\|, \|x_1 - \bar{x}\|) \\ &\leq \psi_{(u_0, \bar{u})}^2(\|x_0 - \bar{x}\|) + \alpha \psi_{(u_0, \bar{u})}^1(\|x_0 - \bar{x}\|) + \frac{1}{2} \|x_1 - \bar{x}\|, \end{aligned}$$

where the second inequality comes from the sublinearity of ϕ^1 . After collecting the $\|x_1 - \bar{x}\|$ -terms and multiplying through by 2, we get

$$(30) \quad \|x_1 - \bar{x}\| \leq \psi_{(u_0, \bar{u})}(\|x_0 - \bar{x}\|) \leq \|x_0 - \bar{x}\| \leq \epsilon_x,$$

where the final two inequalities follow from the bound (28). We conclude that the bound (29) holds when $k = 0$, and that $x_1 \in \mathbb{B}(\bar{x}; \epsilon_x)$.

Since $(u_1, x_1) \in U \times \mathbb{B}(\bar{x}; \epsilon_x)$ and $x_2 \in G(u_1, x_1)$ we can apply the target nonsingularity and σ -approximation bounds as above, but now with $(u, x, x') := (u_1, x_1, x_2)$, to get

$$(31) \quad \|x_2 - \bar{x}\| \leq \psi_{(u_1, \bar{u})}(\|x_1 - \bar{x}\|),$$

which is just the bound (29) with $k = 1$. Moreover, from (30) we know that $\|x_1 - \bar{x}\| \leq \|x_0 - \bar{x}\|$, so the fact that the convergence function is nondecreasing on $\mathbb{R}^+$ leads from (31) to

$$\|x_2 - \bar{x}\| \leq \psi_{(u_1, \bar{u})}(\|x_0 - \bar{x}\|) \leq \|x_0 - \bar{x}\| \leq \epsilon_x,$$

where the final two inequalities follow from (28). If we repeat this process inductively for each $k = 2, 3, \dots$, we likewise get the bound (29) as well as the bound $\|x_{k+1} - \bar{x}\| \leq \|x_0 - \bar{x}\| \leq \epsilon_x$, which allows us to continue the induction.

If the sequence of parameters $\{u_k\}$ is target appropriate for $(\psi_{(u, \bar{u})}, \|x_0 - \bar{x}\|)$ with target $\bar{u}$, we know that the sequence of scalars

$$(32) \quad r_0 := \|x_0 - \bar{x}\|, r_k := \psi_{(u_{k-1}, \bar{u})}(r_{k-1}) \text{ for } k = 1, 2, \dots$$

converges to zero. From the bound (29) in the case when $k = 0$, we deduce that

$$\|x_1 - \bar{x}\| \leq \psi_{(u_0, \bar{u})}(\|x_0 - \bar{x}\|) = \psi_{(u_0, \bar{u})}(r_0) = r_1.$$

Likewise, from the bound (29) in the case when $k = 1$, we deduce that

$$\|x_2 - \bar{x}\| \leq \psi_{(u_1, \bar{u})}(\|x_1 - \bar{x}\|) \leq \psi_{(u_1, \bar{u})}(r_1) = r_2,$$

where the second inequality follows from the preceding bound and the fact that ψ is nondecreasing on $\mathbb{R}^+$. We can continue in this way inductively to deduce the general bound

$$\|x_k - \bar{x}\| \leq r_k \text{ for } k = 0, 1, 2, \dots$$

Since we know that the sequence of scalars $\{r_k\}$ converges to zero, we conclude that $x_k \rightarrow \bar{x}$.

If $u_k \rightarrow \bar{u}$ and the approximating inclusion (2) is a valid approximation of (1) at $\bar{x}$ for parameter target $\bar{u}$, then we know from the first criterion for validity that $a(u_k, x_k, x_{k+1}) \rightarrow f(\bar{x})$. Since $a(u_k, x_k, x_{k+1}) \in A(u_k, x_k, x_{k+1})$, we conclude that $f(\bar{x})$ is an outer limit point of A at $(\bar{u}, \bar{x}, \bar{x})$. The second criterion for validity then gives $f(\bar{x}) \in A(\bar{u}, \bar{x}, \bar{x}) \subseteq F(\bar{x})$. $\square$

We will find it useful in what follows to have the following version of Theorem 3.1, which replaces the bound (28) with the direct assumption that $\{x_k\} \subseteq \mathbb{B}(\bar{x}; \epsilon_x)$ by using the *target-localized generic iteration mapping* defined by

$$(33) \quad G_{\bar{x}\text{loc}}(u, x) := G(u, x) \cap \mathbb{B}(\bar{x}; \epsilon_x).$$

THEOREM 3.2 (*Q-convergence of the target-localized generic algorithm*). *If the Q-convergence condition holds for the target-localized generic iteration mapping (33), then any sequence initiated from x_0 and conforming to the target-localized generic algorithm $x_{k+1} \in G_{\bar{x}\text{loc}}(u_k, x_k)$ with a sequence of parameters $\{u_k\} \subseteq U$ satisfies (29) in terms of the convergence function $\psi_{(u, u')}$ defined by (27). Under these circumstances, if the sequence of parameters $\{u_k\}$ is target appropriate for $(\psi_{(u, \bar{u})}, \|x_0 - \bar{x}\|)$ with target $\bar{u}$, then the sequence $\{x_k\}$ converges to $\bar{x}$.*

If in addition $u_k \rightarrow \bar{u}$ and the approximating inclusion (2) is a valid approximation of (1) at $\bar{x}$ for parameter target $\bar{u}$, then $\bar{x}$ is a solution to the original inclusion (1).

Proof. This follows exactly the same pattern as the proof of Theorem 3.1, but now the inclusions $x_k \in \mathbb{B}(\bar{x}; \epsilon_x)$ come not from the bound (28) but directly from the assumptions. $\square$

Remark. Note that Theorems 3.1 and 3.2 both hold more generally when $\mathcal{X}$ is merely a normed linear space.

3.2. R-convergence. To motivate our main *R*-convergence results, we again turn to Newton's method, but now we are interested in how $\|x'' - x'\|$ compares to $\|x' - x\|$. For Newton's method, we know that $x'' \in G(u', x') = x' + \nabla f(x')^{-1}[-f(x')]$, so the first difference in the comparison of interest satisfies

$$(34) \quad \|x'' - x'\| = \|\nabla f(x')^{-1}[-f(x')]\| \leq \|\nabla f(x')^{-1}\| \| -f(x') \|,$$

where we have assumed the nonsingularity condition that $\nabla f(x')$ is invertible with $\|\nabla f(x')^{-1}\| \leq \alpha$ for all iterates x' . Combining this with the estimate (34) gives us a new version of the nonsingularity condition:

$$(35) \quad \|x'' - x'\| \leq \alpha \|\sigma(x')\|$$

in terms of the mapping $\sigma(x') := -f(x')$. To make the comparison between $\|x'' - x'\|$ and $\|x' - x\|$, then all we need is a connection between $\|\sigma(x')\|$ and $\|x' - x\|$. In this case, since $f(x) + \nabla f(x)[x' - x] = 0$, we know there exists an $\epsilon \geq 0$ such that $\sigma(x')$ satisfies the approximation bound

$$(36) \quad \|\sigma(x')\| = \|-f(x') + f(x) + \nabla f(x)[x' - x]\| \leq \epsilon \|x' - x\|^2$$

for all x' and x close enough to x_0 . Combining (36) with the nonsingularity condition (35) we already developed gives the comparison we seek:

$$(37) \quad \|x'' - x'\| \leq \alpha \epsilon \|x' - x\|^2.$$

From this bound, we can conclude that a sequence $\{x_k\}$ conforming to Newton's method converges to some $\bar{x}$ as long as the first iterate x_1 is close enough to the initial point x_0 . Moreover, we can apply the bound (37) inductively to deduce that

$$\|x_k - \bar{x}\| \leq c \xi^{2^k}$$

for some scalars $c \in \mathbb{R}^+$ and $\xi \in (0, 1)$, which means that $\{x_k\}$ converges to $\bar{x}$ with an R -convergence rate of 2. Notice that in contrast to the conditions underlying the Q -convergence analysis of Newton's method, the conditions for the R -convergence analysis do not explicitly involve any target solution $\bar{x}$.

The essential condition for our convergence analysis is the following generalization of the pair of conditions (36) and (35) for Newton's method, and stated in terms of a general set-valued mapping $X : \mathcal{U} \times \mathcal{X} \rightrightarrows \mathcal{X}$.

R-convergence condition for X . There exist scalars $\alpha \in (0, \infty)$ and $\epsilon_x \geq 0$, a parameter set $U \subseteq \mathcal{U}$, and a single-valued mapping $\sigma : \mathcal{U} \times \mathcal{U} \times \mathcal{X} \times \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{Y}$ such that the following pair of conditions is satisfied:

- (σ -approximation) There exists a family of bivariate convergence functions $\phi_{(u, u')}^1$ parameterized by $(u, u') \in U \times U$ (with sublinearity constant $1/(2\alpha)$) such that for any pairs $(u, x), (u', x') \in U \times \mathbb{B}(x_0; \epsilon_x)$ and any $x'' \in X(u', x')$ the bound is satisfied that

$$\|\sigma(u, u', x, x', x'')\| \leq \phi_{(u, u')}^1(\|x' - x\|, \|x'' - x'\|).$$

- (Nonsingularity) There exists a family of bivariate convergence functions $\phi_{(u, u')}^2$ parameterized by $(u, u') \in U \times U$ (with sublinearity constant α) such that for any pairs $(u, x), (u', x') \in U \times \mathbb{B}(x_0; \epsilon_x)$ with $x' \in X(u, x)$, there exists an $x'' \in X(u', x')$ which satisfies

$$\|x'' - x'\| \leq \phi_{(u, u')}^2(\|x' - x\|, \|\sigma(u, u', x, x', x'')\|).$$

Remark. The R -convergence condition can be viewed as a generalization of Robinson's "point-based approximations" from [8]. There, a point-based approximation $PBA(x, x')$ for $f(x')$ was a function satisfying two properties, the first of which is identical to our σ -approximation condition in the special case when

$$\begin{aligned} \sigma(x, x') &:= f(x') - PBA(x, x'), \\ \phi_{(u, u')}^1(r, r') &:= \frac{1}{2} k r^2. \end{aligned}$$

The second condition defining point-based approximations involves a kind of Lipschitz continuity for PBA and is related to our nonsingularity condition in the special case when $X(x) := PBA(x, \cdot)^{-1}(0)$.

Comparing the Q -convergence condition already presented to the R -convergence condition, we see that the latter involves a relationship between three consecutive iterates x , x' , and x'' , whereas the former involves a similar relationship between two consecutive iterates x and x' and the target solution $\bar{x}$. This is consistent with the fact that R -convergence is the type of convergence that naturally arises from comparisons of consecutive iterates without presuming a target $\bar{x}$. Moreover, if we compare target nonsingularity with the nonsingularity condition from the R -convergence condition, we see a fundamental difference between existence and universality. In particular, nonsingularity stipulates the *existence* of an x'' for which the subsequent inequality is satisfied, whereas target nonsingularity has the analogous inequality holding *for all* appropriate x' . This distinction will be reinforced in section 4, where we discuss how nonsingularity and target nonsingularity are related to different generalized continuity properties.

It is easy to see that the combination of the σ -approximation and nonsingularity bounds from the R -convergence condition yields the series of inequalities

$$\begin{aligned} \|x'' - x'\| &\leq \phi_{(u,u')}^2(\|x' - x\|, \|\sigma(u, u', x, x', x'')\|) \\ &\leq \psi_{(u,u')}^2(\|x' - x\|) + \alpha \|\sigma(u, u', x, x', x'')\| \\ &\leq \psi_{(u,u')}^2(\|x' - x\|) + \alpha \phi_{(u,u')}^1(\|x' - x\|, \|x'' - x\|) \\ &\leq \psi_{(u,u')}^2(\|x' - x\|) + \alpha \psi_{(u,u')}^1(\|x' - x\|) + \frac{1}{2} \|x'' - x'\|, \end{aligned}$$

where the second and final inequalities follow from the sublinearity of ϕ_2 and ϕ_1 , respectively. If we collect the $\|x'' - x'\|$ -terms and multiply by 2, we get the bound

$$(38) \quad \|x'' - x'\| \leq \psi_{(u,u')}(\|x' - x\|)$$

in terms of the convergence function

$$(39) \quad \psi_{(u,u')}(r) := 2\alpha \psi_{(u,u')}^1(r) + 2\psi_{(u,u')}^2(r)$$

as long as all the variables involved are the same in both conditions. This is the general analogue of the comparison (37) for Newton's method, and just as in that case, it is the key to establishing that the iterates from the generic algorithm converge to a limit point $\bar{x}$.

We will consider first the generic algorithm and then two augmentations of it which are distinguished by different ways of accommodating the local nature of the R -convergence condition. The convergence theorems for all three algorithms depend on the notion of an "appropriate" sequence of parameters. Given a family of convergence functions $\psi_{(u,u')}$ parameterized by $(u, u') \in U \times U$ and a scalar r_0 , we say that a sequence of parameters $u_k \in U$ is *appropriate for* $(\psi_{(u,u')}, r_0)$ if the sequence $\{r_k\}$ defined by

$$r_0, r_k := \psi_{(u_{k-1}, u_k)}(r_{k-1}) \text{ for } k = 1, 2, \dots$$

converges to zero. Families of convergence functions with appropriate parameterizations can be viewed as parametric versions of the "convergence rates" used in [7] to establish R -convergence of induction processes in a general setting.

THEOREM 3.3 (R -convergence of the generic algorithm). *If the R -convergence condition holds for the generic iteration mapping (3), then there exists a sequence $\{x_k\}$ initiated from x_0 and conforming to the generic algorithm $x_{k+1} \in G(u_k, x_k)$*

that converges to a point $\bar{x} \in \mathbb{B}(x_0; \epsilon_x)$ as long as $x_1 \in G(u_0, x_0) \cap \mathbb{B}(x_0; \epsilon_x)$ and the corresponding sequence of parameters $\{u_k\} \subseteq U$ is appropriate for $(\psi_{(u, u')}, \|x_1 - x_0\|)$ with $\psi_{(u, u')}$ defined by (39) and with

$$(40) \quad \sum_{i=0}^{\infty} r_i \leq \epsilon_x$$

for the sequence of scalars $\{r_k\}$ defined by

$$(41) \quad r_0 := \|x_1 - x_0\|, r_k := \psi_{(u_{k-1}, u_k)}(r_{k-1}) \text{ for } k = 1, 2, \dots$$

Moreover, the R -convergence rate of $\{x_k\}$ to $\bar{x}$ is governed by

$$(42) \quad \|x_k - \bar{x}\| \leq s_k := \sum_{i=k}^{\infty} r_i.$$

If nonsingularity instead holds for all $x'' \in G(u', x')$, then the same statements are true for every sequence $\{x_k\}$ initiated from x_0 and conforming to the generic algorithm.

In either case, if the sequence of parameters $\{u_k\}$ converges to a parameter target $\bar{u}$ and the approximating inclusion (2) is a valid approximation of (1) at $\bar{x}$ for parameter target $\bar{u}$, then $\bar{x}$ is a solution to the original inclusion (1).

Proof. Starting with the pairs $(u, x) := (u_0, x_0)$ and $(u', x') := (u_1, x_1)$, we apply the R -convergence condition for G to obtain the next element of the sequence $x_2 \in G(u_1, x_1)$ which satisfies

$$(43) \quad \|x_2 - x_1\| \leq \psi_{(u_0, u_1)}(\|x_1 - x_0\|) = r_1.$$

In order to apply the R -convergence condition inductively, we need to ensure that $x_2 \in \mathbb{B}(x_0; \epsilon_x)$. To get this, we note that the bound (43) gives us

$$(44) \quad \|x_2 - x_0\| \leq \|x_2 - x_1\| + \|x_1 - x_0\| \leq r_1 + r_0 \leq \epsilon_x,$$

where the last inequality follows from the assumption (40). We conclude that x_2 is indeed in $\mathbb{B}(x_0; \epsilon_x)$ and we again apply the R -convergence condition for G , but this time with $(u, x) := (u_1, x_1)$ and $(u', x') := (u_2, x_2)$. In this way, we obtain the next element of the sequence $x_3 \in G(u_2, x_2)$ which satisfies

$$\|x_3 - x_2\| \leq \psi_{(u_1, u_2)}(\|x_2 - x_1\|) \leq \psi_{(u_1, u_2)}(r_1) = r_2,$$

where the second inequality follows from (43) and the fact that the convergence function is nondecreasing on $\mathbb{R}^+$. We use this inequality together with (44) and (40) to get the bound

$$\|x_3 - x_0\| \leq \|x_3 - x_2\| + \|x_2 - x_0\| \leq r_2 + r_1 + r_0 \leq \epsilon_x,$$

from which we deduce that $x_3 \in \mathbb{B}(x_0; \epsilon_x)$. By continuing this process, we show that for each current iterate $x_k \in \mathbb{B}(x_0; \epsilon_x)$, there exists a new iterate $x_{k+1} \in G(u_k, x_k)$ satisfying

$$(45) \quad \|x_{k+1} - x_k\| \leq \psi_{(u_{k-1}, u_k)}(\|x_k - x_{k-1}\|) \leq \psi_{(u_{k-1}, u_k)}(r_{k-1}) = r_k$$

as well as

$$(46) \quad \|x_{k+1} - x_0\| \leq \|x_{k+1} - x_k\| + \|x_k - x_{k-1}\| + \cdots + \|x_1 - x_0\| \leq \sum_{i=0}^k r_i \leq \epsilon_x,$$

so that the new iterate x_{k+1} is also in $\mathbb{B}(x_0; \epsilon_x)$. Since we know that $r_k \rightarrow 0$, it follows from (45) that the differences $\|x_{k+1} - x_k\|$ also converge to zero. As $\mathcal{X}$ is a Banach space, we conclude that the sequence $\{x_k\} \subseteq \mathbb{B}(x_0; \epsilon_x)$ converges to a limit point $\bar{x} \in \mathbb{B}(x_0; \epsilon_x)$ as claimed.

The bound (42) follows from the combination of the bound (45) and the fact that

$$\begin{aligned} \|x_k - \bar{x}\| &\leq \|x_k - x_{k+1}\| + \|x_{k+1} - x_{k+2}\| + \cdots \\ &= \|x_{k+1} - x_k\| + \|x_{k+2} - x_{k+1}\| + \cdots \end{aligned}$$

If nonsingularity instead holds for all $x'' \in G(u', x')$, then the bounds (45) and (46) hold for *any* new iterate $x_{k+1} \in G(u_k, x_k)$ since the bound (38) holds for any $x'' \in G(u', x')$. Everything else follows from this in the same way as above.

Finally, if $u_k \rightarrow \bar{u}$ and the approximating inclusion (2) is a valid approximation of (1) at $\bar{x}$ for parameter target $\bar{u}$, then we know from the first criterion for validity that $a(u_k, x_k, x_{k+1}) \rightarrow f(\bar{x})$. Since $a(u_k, x_k, x_{k+1}) \in A(u_k, x_k, x_{k+1})$, we conclude that $f(\bar{x})$ is an outer limit point of A at $(\bar{u}, \bar{x}, \bar{x})$. The second criterion for validity then gives $f(\bar{x}) \in A(\bar{u}, \bar{x}, \bar{x}) \subseteq F(\bar{x})$. $\square$

The first augmentation of the generic algorithm (3) on which we focus our attention is the localized generic algorithm

$$(47) \quad x' \in G_{\text{loc}}(u, x) := G(u, x) \cap \mathbb{B}(x_0; \epsilon_x)$$

which assimilates the local nature of the convergence and nonsingularity conditions by directly building them into the algorithm.

THEOREM 3.4 (*R-convergence of the localized generic algorithm*). *If the R-convergence condition holds for the localized generic iteration mapping (47), then there exists a sequence $\{x_k\}$ initiated from x_0 and conforming to the localized generic algorithm $x_{k+1} \in G_{\text{loc}}(u_k, x_k)$ that converges to a point $\bar{x} \in \mathbb{B}(x_0; \epsilon_x)$ as long as the corresponding sequence of parameters $\{u_k\} \subseteq U$ is appropriate for $(\psi_{(u, u')}, \|x_1 - x_0\|)$ with $\psi_{(u, u')}$ defined by (39).*

Moreover, if the sequence of scalars $\{r_k\}$ defined by (41) is summable, the sequence $\{x_k\}$ converges to $\bar{x}$ at an R-convergence rate governed by (42).

If nonsingularity instead holds for all $x'' \in G_{\text{loc}}(u', x')$, then the same statements are true for every sequence $\{x_k\}$ initiated from x_0 and conforming to the localized generic algorithm.

In either case, if the sequence of parameters $\{u_k\}$ converges to a parameter target $\bar{u}$ and the approximating inclusion (2) is a valid approximation of (1) at $\bar{x}$ for parameter target $\bar{u}$, then $\bar{x}$ is a solution to the original inclusion (1).

Proof. The argument here is very similar to that for Theorem 3.3; however, now the new iterates $x_{k+1} \in G_{\text{loc}}(u_k, x_k)$ at each step are automatically in the ball $\mathbb{B}(x_0; \epsilon_x)$, so we do not need the condition (40). To conclude the statement about an R-convergence rate in this case, we need the summability of the sequence of scalars $\{r_k\}$ defined by (41) (which, in particular, follows from the stronger assumption (40) in Theorem 3.3). $\square$

Notice that the generic algorithm (localized or not) might present multiple choices for the next iterate at any given step since the generic iteration mapping G (3) can

be set-valued in general. One practical way of dealing with this situation is to use instead the *conservative generic algorithm* defined by

$$(48) \quad x_{k+1} \in G_0(u_k, x_k) := \operatorname{argmin}_{x' \in G(u_k, x_k)} \|x' - x_0\|.$$

This is labeled “conservative” because at every step it stays as close as possible to the original iterate x_0 . As we will show in the next result, the conservative generic algorithm ensures that all iterates are close enough to x_0 as long as there exists at least one point in each relevant image of the localized generic iteration mapping.

THEOREM 3.5 (*R-convergence of the conservative generic algorithm*). *Assume that the R-convergence condition holds for the localized generic iteration mapping (47), with nonsingularity instead holding for all $x'' \in G_{\text{loc}}(u', x')$. If $G_{\text{loc}}(u, x) \neq \emptyset$ whenever $(u, x) \in U \times \mathbb{B}(x_0; \epsilon_x)$, then every sequence $\{x_k\}$ initiated from x_0 and conforming to the conservative generic algorithm $x_{k+1} \in G_0(u_k, x_k)$ converges to a point $\bar{x} \in \mathbb{B}(x_0; \epsilon_x)$ as long as the corresponding sequence of parameters $\{u_k\} \subseteq U$ is appropriate for $(\psi_{(u, u')}, \|x_1 - x_0\|)$ with $\psi_{(u, u')}$ defined by (39). Moreover, if the sequence of scalars $\{r_k\}$ (41) is summable, then the sequence $\{x_k\}$ converges to $\bar{x}$ at an R-convergence rate governed by (42).*

Finally, if the sequence of parameters $\{u_k\}$ converges to a parameter target $\bar{u}$ and the approximating inclusion (2) is a valid approximation of (1) at $\bar{x}$ for parameter target $\bar{u}$, then $\bar{x}$ is a solution to the original inclusion (1).

Proof. This result follows from an argument similar to that in the proof of Theorem 3.3. In this case, however, we know that $x_{k+1} \in G_0(u_k, x_k) \subseteq G(u_k, x_k)$, and the inclusion $x_{k+1} \in \mathbb{B}(x_0; \epsilon_x)$ follows from the nonemptiness of $G_{\text{loc}}(u_k, x_k)$, since the conservative algorithm chooses the element of $G(u_k, x_k)$ closest to x_0 . $\square$

Remark. Note that from the beginning here we assume that nonsingularity holds for all $x'' \in G_{\text{loc}}(u', x')$, which is different from Theorem 3.4, where we first demand only the existence of one $x'' \in G_{\text{loc}}(u', x')$ for which the nonsingularity condition holds. The reason for this difference is that the conservative generic algorithm cannot be guaranteed to choose any particular element $x'' \in G_{\text{loc}}(u', x')$, so all such elements must satisfy the nonsingularity condition.

4. Nonsingularity from generalized continuity. The target nonsingularity condition in our analysis of Q -convergence in Theorems 3.1 and 3.2 and the nonsingularity condition in our R -convergence analysis in Theorems 3.4, 3.3, and 3.5 are both related to generalized continuity properties. In particular, we will see that target nonsingularity is related to a generalized calmness property and nonsingularity to a generalized Lipschitz continuity property. Recall that calmness is a kind of restricted Lipschitz continuity where one point of comparison is fixed as the base point.

4.1. Target nonsingularity from ϕ -calmness. We begin this section by discussing how target nonsingularity is related to a generalized continuity property for the generic iteration mapping (3). The continuity property we use is a parametric version of “selection calmness” for set-valued mappings (also sometimes called “local upper Lipschitz continuity” and “upper Lipschitz continuity at a point”). Extending the definition from [5] to normed linear spaces $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$, we say that a set-valued mapping $Z : \mathcal{X} \times \mathcal{Y} \rightrightarrows \mathcal{Z}$ has *calm selections near* $((\bar{x}, \bar{y}), \bar{z})$ if there exist scalars $\alpha \in (0, \infty)$, $\epsilon_x \geq 0$, $\epsilon_y \geq 0$, and $\epsilon_z \geq 0$ such that

$$Z(x, y) \cap \mathbb{B}(\bar{z}; \epsilon_z) \subseteq \mathbb{B}(\bar{z}; \alpha \|(x, y) - (\bar{x}, \bar{y})\|) \quad \text{for all } (x, y) \in \mathbb{B}(\bar{x}; \epsilon_x) \times \mathbb{B}(\bar{y}; \epsilon_y).$$

The following proposition shows that selection calmness of G is equivalent to the target nonsingularity bound holding in the Q -convergence condition for the target-localized generic iteration mapping with a special family of bivariate convergence functions $\phi_{(u,u')}$ and function σ .

PROPOSITION 4.1.1. *The following are equivalent:*

- G has calm selections near $((\bar{u}, \bar{x}), \bar{x})$.
- There are scalars $\alpha \in (0, \infty)$, $\epsilon_u \geq 0$, and $\epsilon_x \geq 0$ such that any trio $(u, x, x') \in \mathbb{B}(\bar{u}; \epsilon_u) \times \mathbb{B}(\bar{x}; \epsilon_x) \times G_{\bar{x}\text{loc}}(u, x)$ satisfies the target nonsingularity bound in the Q -convergence condition for $G_{\bar{x}\text{loc}}$ with $\sigma(u, \bar{u}, x, x', \bar{x}) := (u, x) - (\bar{u}, \bar{x})$ and $\phi_{(u,u')}^2(r, r') := \alpha r'$.

Proof. This follows since the selection calmness of G is equivalent to the existence of scalars $\alpha \in (0, \infty)$, $\epsilon_u \geq 0$, and $\epsilon_x \geq 0$ such that any trio $(u, x, x') \in \mathbb{B}(\bar{u}; \epsilon_u) \times \mathbb{B}(\bar{x}; \epsilon_x) \times G_{\bar{x}\text{loc}}(u, x)$ satisfies

$$\|x' - \bar{x}\| \leq \alpha \|(u, x) - (\bar{u}, \bar{x})\| = \phi_{(u, \bar{u})}^2(\|x - \bar{x}\|, \|\sigma(u, \bar{u}, x, x', \bar{x})\|). \quad \square$$

In order to connect the target nonsingularity bound with general ϕ^2 and σ , we need to extend the definition of selection the calmness by considering parameterized families of set-valued mappings and replacing the term on the right side of the calmness bound with a bivariate convergence function. A family of set-valued mappings $Z_u : \mathcal{X} \times \mathcal{Y} \rightrightarrows \mathcal{Z}$ parameterized by $u \in \mathcal{U}$ is said to have ϕ -calm selections near $((\bar{x}, \bar{y}), \bar{z})$ for parameter target $\bar{u}$ if there is a parameter set $U \subseteq \mathcal{U}$ containing $\bar{u}$, and scalars $\epsilon_x \geq 0$, $\epsilon_y \geq 0$, and $\epsilon_z \geq 0$ such that the family of bivariate convergence functions $\phi_{(u,u')}$ parameterized by $(u, u') \in U \times U$ satisfies

$$Z_u(x, y) \cap \mathbb{B}(\bar{z}; \epsilon_z) \subseteq \mathbb{B}(\bar{z}; \phi_{(u, \bar{u})}(\|x - \bar{x}\|, \|y - \bar{y}\|))$$

for all $(u, x, y) \in U \times \mathbb{B}(\bar{x}; \epsilon_x) \times \mathbb{B}(\bar{y}; \epsilon_y)$.

For set-valued mappings $Z : \mathbb{R}^{n+m} \rightrightarrows \mathcal{Z}$, the earlier notion of calm selections is covered by selection ϕ -calmness when the (unparameterized) bivariate convergence function is $\phi(r, r') := \alpha \sqrt{r^2 + (r')^2}$.

The next result shows a relationship between selection ϕ -calmness and target nonsingularity of the target-localized generic iteration mapping (33) restricted as follows:

$$(49) \quad G_{\text{res}}(u, x) := \{x' \in G_{\bar{x}\text{loc}}(u, x) \mid \|\sigma(u, \bar{u}, x, x', \bar{x})\| \leq \epsilon_y\}.$$

PROPOSITION 4.1.2. *The following are equivalent:*

- The family of partial inverse mappings $X_u : \mathcal{X} \times \mathcal{Y} \rightrightarrows \mathcal{X}$ defined by

$$(50) \quad X_u(x, y) := G(u, x) \cap \sigma(u, \bar{u}, x, \cdot, \bar{x})^{-1}(y)$$

has ϕ^2 -calm selections near $((\bar{x}, 0), \bar{x})$ for parameter target $\bar{u}$.

- There exist scalars $\epsilon_x \geq 0$ and $\epsilon_y \geq 0$, a parameter set $U \subseteq \mathcal{U}$, and a parameter target $\bar{u} \in U$ such that any pair $(u, x) \in U \times \mathbb{B}(\bar{x}; \epsilon_x)$ with $x' \in G_{\text{res}}(u, x)$ satisfies the target nonsingularity bound in the Q -convergence condition for G_{res} with σ and ϕ^2 .

Proof. This follows from the definition of ϕ^2 -calmness since $x' \in X_u(x, y)$ if and only if $x' \in G(u, x)$ and $y := \sigma(u, \bar{u}, x, x', \bar{x})$. $\square$

Remark. Recall from the statement of Theorem 3.2 that what we would really like from Proposition 4.1.2 is the target nonsingularity bound holding in the Q -convergence condition for the target-localized generic iteration mapping $G_{\bar{x}\text{loc}}$ itself, and not just for its restricted version G_{res} given by (49). However, in practice, it turns out that target nonsingularity for G_{res} as provided by Proposition 4.1.2 is often enough to make exactly the same conclusions as in Theorem 3.2, since the restriction of $\|\sigma(u, \bar{u}, x, x', \bar{x})\| \leq \epsilon_y$ can easily be satisfied as long as $\sigma(u, \bar{u}, x, x', \bar{x}) \rightarrow 0$ (which, in particular, is the case in both of the examples we cover).

4.2. Nonsingularity from (ϕ, σ) -continuity. Nonsingularity is related to an extension of a generalized Lipschitz continuity property for set-valued mappings called “Aubin continuity” (also sometimes called “pseudo-Lipschitz continuity” and the “Aubin property”; see [12, Definition 9.26]). A set-valued mapping $Z : \mathcal{U} \times \mathcal{X} \rightrightarrows \mathcal{Z}$ is said to be *Aubin continuous near* $((u_0, x_0), z_0)$ if there is a closed neighborhood $U \subseteq \mathcal{U}$ of u_0 and scalars $\alpha \in (0, \infty)$, $\epsilon_x \geq 0$, and $\epsilon_z \geq 0$ such that

$$Z(u, x) \cap \mathbb{B}(z_0; \epsilon_z) \subseteq Z(u', x') + \alpha \|(u, x) - (u', x')\| \mathbb{B}(0; 1)$$

for all pairs (u, x) and (u', x') in $U \times \mathbb{B}(x_0; \epsilon_x)$. The property we use to study nonsingularity is an extension of Aubin continuity where the norm on the right side of the Aubin continuity bound is replaced by a more general term involving ϕ and σ . A set-valued mapping $Z : \mathcal{U} \times \mathcal{X} \rightrightarrows \mathcal{Z}$ is said to be (ϕ, σ) -continuous near $((u_0, x_0), z_0)$ if there is a closed neighborhood $U \subseteq \mathcal{U}$ of u_0 as well as scalars $\epsilon_x \geq 0$ and $\epsilon_z \geq 0$, such that the family of bivariate convergence functions $\phi_{(u, u')}$ parameterized by $(u, u') \in U \times U$ and the single-valued mapping $\sigma : \mathcal{U} \times \mathcal{U} \times \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{Y}$ satisfy

$$(51) \quad Z(u, x) \cap \mathbb{B}(z_0; \epsilon_z) \subseteq Z(u', x') + \phi_{(u, u')}(\|x' - x\|, \|\sigma(u, u', x, x')\|) \mathbb{B}(0; 1)$$

for all pairs (u, x) and (u', x') in $U \times \mathbb{B}(x_0; \epsilon_x)$. When we use the particular mapping $\sigma(u, u', x, x') := (u, x) - (u', x')$ and the bivariate convergence functions $\phi_{(u, u')}(r, r') := \alpha r'$, (ϕ, σ) -continuity is the same as Aubin continuity. The following proposition is immediate from the definition of (ϕ, σ) -continuity.

PROPOSITION 4.2.1. *If the set-valued mapping $X : \mathcal{U} \times \mathcal{X} \rightrightarrows \mathcal{X}$ is (ϕ^2, σ) -continuous near $((u_0, x_0), x_0)$, then the nonsingularity bound in the R -convergence condition for X holds with ϕ^2 and σ .*

Remark. Notice that (ϕ, σ) -continuity does not have x'' appearing in the σ -term, so neither does the nonsingularity bound given by Proposition 4.2.1.

Notice too that the (ϕ^2, σ) -continuity stipulated in Proposition 4.2.1 is a stronger property in general than nonsingularity. This follows because (ϕ^2, σ) -continuity stipulates that there exists an $x'' \in X(u', x')$ for each $z \in X(u, x)$ and not just for the particular element $z = x'$ matching the second argument generating the image set $X(u', x')$.

The next corollary follows immediately from Proposition 4.2.1 and our observation about (ϕ, σ) -continuity generalizing Aubin continuity.

COROLLARY 4.2.1. *If the set-valued mapping $X : \mathcal{U} \times \mathcal{X} \rightrightarrows \mathcal{X}$ is Aubin continuous near (u_0, x_0) for x_0 , then the nonsingularity bound holds in the R -convergence condition for X with $\sigma(u, u', x, x') := (u, x) - (u', x')$ and $\phi_{(u, u')}^2(r, r') := \alpha r'$.*

5. Examples revisited. The results presented in Theorems 3.1–3.5 can be applied to analyze the convergence of any algorithm that fits the generic model (1).

In this section, we will explore one way in which these general results can be applied to the two examples already introduced. The reader should note, however, that the general results developed in this paper can be applied in many different ways to analyze these same two examples or to analyze any particular algorithm for that matter. In particular, different choices of “bridge” mappings σ lead to different convergence functions in the approximation conditions as well as in the nonsingularity conditions. Thus, the mappings σ chosen for each example below illustrate only one possibility for the convergence analysis.

5.1. Q -convergence.

COROLLARY 5.1.1 (Q -convergence of the generalized primal SQP method). *Consider the family of partial inverse mappings X_u (50), where G encodes the generalized primal SQP method and where σ is defined by*

$$\begin{aligned} \sigma(u, \bar{u}, x, x', \bar{x}) := & -\nabla f_0(x) - \nabla^2 L_u(x)[x' - x] + \nabla f_0(\bar{x}) \\ & + \nabla^2 L_{\bar{u}}(x)[x' - \bar{x}] + \nabla g_{\bar{u}}(\bar{x}) - \nabla g_{\bar{u}}(x) \end{aligned}$$

in terms of the parameterized family of Lagrangian functions $L_u(x) := f_0(x) + g_u(x)$. Assume that X_u has ϕ^2 -calm selections near $((\bar{x}, 0), \bar{x})$ for parameter target $\bar{u}$, that the mappings $(u, x) \mapsto \nabla g_u(x)$ and $(u, x) \mapsto \nabla^2 g_u(x)$ are uniformly continuous on $U \times \mathbb{B}(\bar{x}; \epsilon_x)$, and that there exists a family of convergence functions $\psi_{(u, u')}$ parameterized by $(u, u') \in U \times U$ such that the bound holds:

$$(52) \quad \|\nabla g_u(x) - \nabla g_u(\bar{x}) + \nabla g_{\bar{u}}(\bar{x}) - \nabla g_{\bar{u}}(x)\| \leq \psi_{(u, \bar{u})}^3(\|x - \bar{x}\|)$$

for all $u \in U$ and $x \in \mathbb{B}(\bar{x}; \epsilon_x)$.

It follows that there exists a scalar $\epsilon \geq 0$ such that every sequence $\{x_k\} \subseteq \mathbb{B}(\bar{x}; \epsilon_x)$ initiated at x_0 and conforming to the generalized primal SQP method with a sequence of parameters $\{u_k\} \subseteq U$ satisfies

$$\|x_{k+1} - \bar{x}\| \leq \psi_{(u_k, \bar{u})}(\|x_k - \bar{x}\|)$$

for the family of convergence functions $\psi_{(u, u')}$ defined by

$$\psi_{(u, u')}(r) := 2\alpha(\epsilon r^2 + \psi_{(u, u')}^3(r)) + 2\psi_{(u, u')}^2(r)$$

(in terms of the sublinearity constant $\alpha \in (0, \infty)$ for ϕ^2). If the sequence of parameters $\{u_k\}$ is target appropriate for $(\psi_{(u, u')}, \|x_0 - \bar{x}\|)$ with target $\bar{u}$, then the sequence $\{x_k\}$ converges to $\bar{x}$. If, in addition, the conditions of Proposition 2.0.1 hold and $u_k \rightarrow \bar{u}$, then $\bar{x}$ is a solution to the original inclusion (8).

Proof. We will apply Proposition 4.1.2, Theorem 3.2, and Proposition 2.0.1 to get this result, noting that the definition of σ ensures that, by shrinking U and $\mathbb{B}(\bar{x}; \epsilon_x)$ if necessary, we have $x' \in G_{\text{res}}(u, x)$ whenever $(u, x, x') \in U \times \mathbb{B}(\bar{x}; \epsilon_x) \times G_{\bar{x}\text{loc}}(u, x)$. It follows that all we need to establish is the target σ -approximation bound in the Q -convergence condition with $\phi_{(u, u')}^1$ defined by

$$(53) \quad \phi_{(u, u')}^1(r, r') := \epsilon r^2 + \psi_{(u, u')}^3(r) + \frac{1}{2\alpha} r'.$$

To do this, we consider the series of inequalities

$$\begin{aligned}
\|\sigma(u, \bar{u}, x, x', \bar{x})\| &= \left\| \begin{aligned} &-\nabla f_0(x) - \nabla^2 L_u(x)[x' - x] + \nabla f_0(\bar{x}) \\ &+ \nabla^2 L_{\bar{u}}(x)[x' - \bar{x}] + \nabla g_{\bar{u}}(\bar{x}) - \nabla g_{\bar{u}}(x) \\ &+ \nabla^2 L_u(x)[\bar{x} - \bar{x}] + \nabla g_u(x) - \nabla g_u(x) + \nabla g_u(\bar{x}) - \nabla g_u(\bar{x}) \end{aligned} \right\| \\
&= \left\| \begin{aligned} &-\nabla L_u(x) + \nabla L_u(\bar{x}) - \nabla^2 L_u(x)[\bar{x} - x] \\ &+ \nabla^2 L_{\bar{u}}(x)[x' - \bar{x}] - \nabla^2 L_u(x)[x' - \bar{x}] \\ &+ \nabla g_u(x) - \nabla g_u(\bar{x}) + \nabla g_{\bar{u}}(\bar{x}) - \nabla g_{\bar{u}}(x) \end{aligned} \right\| \\
&\leq \left\| -\nabla L_u(x) + \nabla L_u(\bar{x}) - \nabla^2 L_u(x)[\bar{x} - x] \right\| \\
&\quad + \left\| \nabla^2 L_{\bar{u}}(x)[x' - \bar{x}] - \nabla^2 L_u(x)[x' - \bar{x}] \right\| \\
&\quad + \left\| \nabla g_u(x) - \nabla g_u(\bar{x}) + \nabla g_{\bar{u}}(\bar{x}) - \nabla g_{\bar{u}}(x) \right\|.
\end{aligned}$$

By the differentiability and continuity assumptions on f_0 and the g_u , we know that there exists an $\epsilon \geq 0$ such that the first of the final trio of elements is bounded above by $\epsilon\|x - \bar{x}\|^2$ for all (u, x) close enough to $(\bar{u}, \bar{x})$, and the second element in the trio is bounded above by $\|x' - \bar{x}\|/(2\alpha)$ as long as (u, x) is close enough to $(\bar{u}, \bar{x})$. According to the bound (52), the target σ -approximation bound holds in the Q -convergence condition with $\phi_{(u, u')}^1$ (53). $\square$

Remark. Notice that if we define the family of mappings $\zeta_{(u, \bar{u})} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ by

$$\zeta_{(u, \bar{u})}(x) := \nabla g_u(x) - \nabla g_{\bar{u}}(x),$$

then the bound (52) can be rewritten as

$$\|\zeta_{(u, \bar{u})}(x) - \zeta_{(u, \bar{u})}(\bar{x})\| \leq \psi_{(u, \bar{u})}^3(\|x - \bar{x}\|) \text{ for all } u \in U \text{ and } x \in \mathbb{B}(\bar{x}; \epsilon_x).$$

Consequently, the convergence function $\psi_{(u, \bar{u})}^3$ acts as a Lipschitz bounding term on $\zeta_{(u, \bar{u})}$. In the traditional case when $g_u(x) := u \cdot g(x)$, the mapping $\zeta_{(u, \bar{u})}$ simplifies to $\nabla g(x)[u - \bar{u}]$, and $\psi_{(u, u')}^3$ can be chosen as

$$\psi_{(u, u')}^3(r) = c \|u - u'\| r,$$

where $c \geq 0$ is a Lipschitz constant for $\nabla g(\cdot)$ near $\bar{x}$.

COROLLARY 5.1.2 (Q -convergence of the generalized quasi-Newton method). *Consider the family of partial inverse mappings X_u (50), where G encodes the generalized quasi-Newton method and where σ is defined by*

$$\sigma(u, \bar{u}, x, x', \bar{x}) := v - \bar{v} - f(x) + f(\bar{x}) - (H - \nabla f(\bar{x}))[x' - x]$$

in terms of parameter pairs $u := (v, H)$ and a parameter target $\bar{u} := (\bar{v}, \nabla f(\bar{x}))$. Assume that X_u has ϕ^2 -calm selections near $((\bar{x}, 0), \bar{x})$ for parameter target $\bar{u}$.

It follows that there exists a scalar $\epsilon \geq 0$ such that every sequence $\{x_k\} \subseteq \mathbb{B}(\bar{x}; \epsilon_x)$ initiated at x_0 and conforming to the generalized quasi-Newton method with a sequence of parameters $\{u_k\} \subseteq U$ satisfies

$$\|x_{k+1} - \bar{x}\| \leq \psi_{(u_k, \bar{u})}(\|x_k - \bar{x}\|)$$

in terms of the family of convergence functions $\psi_{(u, u')}$ defined by

$$\psi_{(u, u')}(r) := 2\alpha(\|v - v'\| + \epsilon r^2 + \|H - H'\| r) + 2\psi_{(u, u')}^2(r)$$

(where $\alpha \in (0, \infty)$ is the sublinearity constant for ϕ^2). If the sequence of parameters $\{u_k\}$ is target appropriate for $(\psi_{(u,u')}, \|x_0 - \bar{x}\|)$ with target $\bar{u}$, then the sequence $\{x_k\}$ converges to $\bar{x}$.

If in addition $u_k \rightarrow \bar{u}$ and the mapping F is outer semicontinuous at $\bar{x}$, then $\bar{x}$ is a solution to the original inclusion (17).

Proof. Again, by the definition of σ we know that, by shrinking U and $\mathbb{B}(\bar{x}; \epsilon_x)$ if necessary, we have $x' \in G_{\text{res}}(u, x)$ whenever $(u, x, x') \in U \times \mathbb{B}(\bar{x}; \epsilon_x) \times G_{\bar{x}\text{loc}}(u, x)$. Thus the result follows from Proposition 4.1.2, Theorem 3.2, and Proposition 2.0.2 as soon as we establish the target σ -approximation bound in the Q -convergence condition with $\phi_{(u,u')}^1$ defined by

$$(54) \quad \phi_{(u,u')}^1(r, r') := \|v - v'\| + \epsilon r^2 + \|H - H'\| r + \frac{1}{2\alpha} r'.$$

To do this, we consider the series of bounds

$$(55) \quad \begin{aligned} \|\sigma(u, \bar{u}, x, x', \bar{x})\| &\leq \left\| \begin{aligned} &v - \bar{v} + f(\bar{x}) - f(x) + \nabla f(\bar{x})[x - \bar{x}] \\ &+ (H - \nabla f(\bar{x}))[x - \bar{x}] + (\nabla f(\bar{x}) - H)[x' - \bar{x}] \end{aligned} \right\| \\ &\leq \|v - \bar{v}\| + \|f(\bar{x}) - f(x) + \nabla f(\bar{x})[x - \bar{x}]\| \\ &\quad + \|H - \nabla f(\bar{x})\| \|x - \bar{x}\| + \|H - \nabla f(\bar{x})\| \|x' - \bar{x}\|. \end{aligned}$$

Under the differentiability assumptions on f , we know that there is some $\epsilon \geq 0$ such that for all x close to $\bar{x}$, the bound holds that

$$\|f(\bar{x}) - f(x) + \nabla f(\bar{x})[x - \bar{x}]\| \leq \epsilon \|x - \bar{x}\|^2,$$

and, by shrinking the parameter set U if necessary, we know that $\|H - \nabla f(\bar{x})\| \leq \frac{1}{2\alpha}$, so the result follows. $\square$

Remark. In [3, Theorem 3.1], Dontchev gave a result with essentially the same conclusions as Corollary 5.1.2, but he assumed the selection calmness near $(\bar{v}, \bar{x})$ of the mapping

$$L(y) := (f(\bar{x}) + \nabla f(\bar{x})[\cdot - \bar{x}] + F(\cdot))^{-1}(y)$$

instead of our more general selection ϕ^2 -calmness assumption. For the special case in Corollary 5.1.2 when the bivariate convergence functions are given by

$$(56) \quad \phi_{(u,u')}^2(r, r') := \alpha r',$$

the selection ϕ^2 -calmness of X_u near $((\bar{x}, 0), \bar{x})$ for parameter target $\bar{u}$ follows from Dontchev's assumption. This follows since X_u and σ as in Corollary 5.1.2 satisfy

$$x' \in X_u(x, y) \iff x' \in L(y + \bar{v}) \cap B(\bar{x}; \epsilon_x) \text{ with } y = \sigma(u, \bar{u}, x, x', \bar{x})$$

and $\sigma(u, \bar{u}, x, x', \bar{x})$ can be assumed to be as close to zero as necessary by shrinking U and ϵ_x . Of course, Corollary 5.1.2 allows other possibilities for selection ϕ^2 -calmness than the one used in [3, Theorem 3.1]; however, even in this special case, Dontchev's assumption is stronger than the selection ϕ^2 -calmness of X_u , since the selection calmness bound for L has to hold for all y near 0, and not just for $y = \sigma(u, \bar{u}, x, x', \bar{x})$.

As we have already noted, the generalized quasi-Newton method covers the generalized primal-dual SQP method where variables are paired with dual variables satisfying the standard update. In the classical case when this method is applied to a standard nonlinear program (7), the selection calmness of L (and hence the selection ϕ^2 -calmness of X_u in Corollary 5.1.2) is assured under the combination of the strict Mangasarian–Fromovitz constraint qualification and the second-order sufficient optimality condition (cf. [4]).

5.2. R -convergence.

COROLLARY 5.2.1 (R -convergence of the generalized primal SQP method). *Consider the special case of the generic iteration mapping G encoding the generalized primal SQP method, and consider the mapping σ defined by*

$$\sigma(u, u', x, x') := \nabla^2 L_u(x)[x' - x] + \nabla L_{u'}(x) - \nabla L_{u'}(x')$$

in terms of the parameterized family of Lagrangian functions $L_u(x) := f_0(x) + g_u(x)$. If there exists a closed neighborhood $U \subseteq \mathcal{U}$ of u_0 and a scalar $\epsilon_x \geq 0$ such that the mappings $(u, x) \mapsto \nabla g_u(x)$ and $(u, x) \mapsto \nabla^2 g_u(x)$ are uniformly continuous on $U \times \mathbb{B}(x_0; \epsilon_x)$ and there exists a family of convergence functions $\psi_{(u, u')}$ parameterized by $(u, u') \in U \times U$ such that the following bound holds:

$$(57) \quad \|\nabla g_{u'}(x) - \nabla g_u(x) + \nabla g_u(x') - \nabla g_{u'}(x')\| \leq \psi_{(u, u')}^3(\|x' - x\|)$$

for all $u \in U$ and $x \in \mathbb{B}(x_0; \epsilon_x)$;

then there is a scalar $\epsilon \geq 0$ such that the σ -approximation bound holds in the R -convergence condition for G with the family of (bivariate) convergence functions defined as follows:

$$\phi_{(u, u')}^1(r, r') = \psi_{(u, u')}^1(r) := \epsilon r^2 + \psi_{(u, u')}^3(r).$$

If, in addition, the localized generic iteration mapping in this case is (ϕ^2, σ) -continuous near $((u_0, x_0), x_0)$, then there is a sequence $\{x_k\} \subseteq \mathbb{B}(x_0; \epsilon_x)$ initiated at x_0 and conforming to the generalized primal SQP method that converges to a point $\bar{x} \in \mathbb{B}(x_0; \epsilon_x)$ as long as the corresponding sequence of parameters $\{u_k\} \subseteq U$ is appropriate for $(\psi_{(u, u')}, \|x_1 - x_0\|)$ with $\psi_{(u, u')}$ defined by

$$\psi_{(u, u')}(r) := 2\alpha(\epsilon r^2 + \psi_{(u, u')}^3(r)) + 2\psi_{(u, u')}^2(r),$$

where $\alpha \in (0, \infty)$ is the sublinearity constant for $\phi_{(u, u')}^2$.

Moreover, if the sequence of scalars $\{r_k\}$ defined by (41) is summable, the sequence $\{x_k\}$ converges to $\bar{x}$ at an R -convergence rate governed by (42).

Finally, if the sequence of parameters $\{u_k\}$ converges to a parameter target $\bar{u}$ and the conditions of Proposition 2.0.1 hold, then $\bar{x}$ is a solution to the original inclusion (8).

Proof. From the definition of σ , it follows that

$$\begin{aligned} \|\sigma(u, u', x, x', x'')\| &= \|\nabla^2 L_u(x)[x' - x] + \nabla L_{u'}(x) - \nabla L_{u'}(x')\| \\ &= \left\| \begin{array}{l} \nabla^2 L_u(x)[x' - x] + (\nabla L_u(x) - \nabla L_u(x')) \\ -(\nabla L_u(x) - \nabla L_u(x')) + \nabla L_{u'}(x) - \nabla L_{u'}(x') \end{array} \right\| \\ &\leq \|\nabla^2 L_u(x)[x' - x] + \nabla L_u(x) - \nabla L_u(x')\| \\ &\quad + \|\nabla L_{u'}(x) - \nabla L_u(x) + \nabla L_u(x') - \nabla L_{u'}(x')\| \\ &= \|\nabla^2 L_u(x)[x' - x] + \nabla L_u(x) - \nabla L_u(x')\| \\ &\quad + \|\nabla g_{u'}(x) - \nabla g_u(x) + \nabla g_u(x') - \nabla g_{u'}(x')\|. \end{aligned}$$

The first term in the final equality is bounded above by $\epsilon\|x' - x\|^2$ for some $\epsilon \geq 0$ by the uniform continuity assumption on $(u, x) \mapsto \nabla g_u(x)$ and $(u, x) \mapsto \nabla^2 g_u(x)$. Thus, the σ -approximation bound follows as claimed from (57). Proposition 4.2.1, Theorem 3.4, and Proposition 2.0.1 then give the result. $\square$

Remark. Notice that the bound (57) is closely related to the bound (52) appearing in Corollary 5.1.1, and it can be interpreted in the same way as a Lipschitz bounding term.

Note also that the (ϕ^2, σ) -continuity assumption is merely one way to ensure that the nonsingularity bound holds in the R -convergence condition for G_{loc} and that Theorems 3.3 or 3.5 could instead be applied with the σ -approximation bound conclusion from Corollary 5.2.1 to obtain complementary results for the generalized primal SQP method.

COROLLARY 5.2.2 (R -convergence of the generalized quasi-Newton method, part I). *Consider the special case of the generic iteration mapping G encoding the generalized quasi-Newton method, and consider the mapping σ defined by*

$$\sigma(u, u', x, x') := v' - v - f(x') + f(x) + H[x' - x]$$

in terms of parameter-pairs $u := (v, H)$. If there exists a closed neighborhood $U \subseteq \mathcal{U}$ of u_0 , a scalar $\epsilon_x \geq 0$, and a family of convergence functions $\psi_{(u, u')}^3$ parameterized by $(u, u') \in U \times U$ such that the following bound holds:

$$(58) \quad \|f(x) - f(x') + H[x' - x]\| \leq \psi_{(u, u')}^3(\|x' - x\|) \\ \text{for all } u \in U \text{ and } (x, x') \in \mathbb{B}(x_0; \epsilon_x) \times \mathbb{B}(x_0; \epsilon_x);$$

then the σ -approximation bound holds in the R -convergence condition for G with the family of (bivariate) convergence functions defined:

$$\phi_{(u, u')}^1(r, r') = \psi_{(u, u')}^1(r) := \|v' - v\| + \psi_{(u, u')}^3(r).$$

If, in addition, the localized generic iteration mapping in this case is (ϕ^2, σ) -continuous near $((u_0, x_0), x_0)$, then there is a sequence $\{x_k\} \subseteq \mathbb{B}(x_0; \epsilon_x)$ initiated at x_0 and conforming to the generalized quasi-Newton method that converges to a point $\bar{x} \in \mathbb{B}(x_0; \epsilon_x)$ as long as the corresponding sequence of parameters $\{u_k\} \subseteq U$ is appropriate for $(\psi_{(u, u')}, \|x_1 - x_0\|)$ with $\psi_{(u, u')}$ defined by

$$\psi_{(u, u')}(r) := 2\alpha(\|v' - v\| + \psi_{(u, u')}^3(r)) + 2\psi_{(u, u')}^2(r),$$

where $\alpha \in (0, \infty)$ is the sublinearity constant for $\phi_{(u, u')}^2$.

Moreover, if the sequence of scalars $\{r_k\}$ defined by (41) is summable, the sequence $\{x_k\}$ converges to $\bar{x}$ at an R -convergence rate governed by (42).

Finally, if the sequence of parameters $\{u_k\}$ converges to a parameter target $\bar{u}$ and the mapping F is outer semicontinuous at $\bar{x}$, then $\bar{x}$ is a solution to the original inclusion (17).

Proof. This follows from Proposition 4.2.1, Theorem 3.4, and Proposition 2.0.2 in tandem with the bounds

$$\begin{aligned} \|\sigma(u, u', x, x')\| &= \|v' - v - f(x') + f(x) + H[x' - x]\| \\ &\leq \|v' - v\| + \|-f(x') + f(x) + H[x' - x]\| \\ &\leq \|v' - v\| + \psi_{(u, u')}^3(\|x' - x\|). \quad \square \end{aligned}$$

Remark. Notice that one form that the convergence functions $\psi_{(u, u')}^3$ in (58) might take is

$$(59) \quad \psi_{(u, u')}^3(r) := \epsilon r^\gamma$$

for some fixed scalars $\epsilon \geq 0$ and $\gamma \geq 1$. For instance, if $H := \nabla f(x)$, then $\gamma = 2$ and ϵ is a Lipschitz constant for $\nabla f(\cdot)$ on $\mathbb{B}(x_0; \epsilon_x)$. Smaller $\gamma \geq 1$ result when H merely approximates $\nabla f(x)$.

As we mentioned at the beginning of this section, different choices of bridge mappings σ lead to different convergence analyses, which we will illustrate now in the following alternative to Corollary 5.2.2.

COROLLARY 5.2.3 (*R*-convergence of the generalized quasi-Newton method, part II). *Consider the special case of the generic iteration mapping G encoding the generalized quasi-Newton method, and consider the mapping σ defined by*

$$\sigma(u, u', x, x', x'') := v' - v - f(x') + f(x) + H[x' - x] - (H' - H_0)[x'' - x']$$

in terms of the fixed matrix H_0 . If the set-valued mapping $(H_0 + F)^{-1} : \mathcal{Y} \rightrightarrows \mathcal{X}$ is Aubin continuous near (y_0, x_0) with $y_0 := v_0 - f(x_0) + H_0[x_0]$, and there exists a closed neighborhood $U \subseteq \mathcal{U}$ of (v_0, H_0) and a family of convergence functions $\psi_{(u, u')}$ parameterized by $(u, u') \in U \times U$ such that the bound (58) holds, then there is a scalar $\epsilon_x \geq 0$ and a sequence $\{x_k\} \subseteq \mathbb{B}(x_0; \epsilon_x)$ initiated at x_0 and conforming to the generalized quasi-Newton method that converges to a point $\bar{x} \in \mathbb{B}(x_0; \epsilon_x)$ as long as the corresponding sequence of parameters $\{u_k\} \subseteq U$ is appropriate for $(\psi_{(u, u')}, \|x_1 - x_0\|)$ with $\psi_{(u, u')}$ defined by

$$\psi_{(u, u')}(r) := 2\alpha (\|v' - v\| + \psi_{(u, u')}^3(r)),$$

where $\alpha \in (0, \infty)$ is the Aubin continuity constant for $(H_0 + F)^{-1}$.

*Moreover, if the sequence of scalars $\{r_k\}$ defined by (41) is summable, the sequence $\{x_k\}$ converges to $\bar{x}$ at an *R*-convergence rate governed by (42).*

Finally, if the sequence of parameters $\{u_k\}$ converges to a parameter target $\bar{u}$ and the mapping F is outer semicontinuous at $\bar{x}$, then $\bar{x}$ is a solution to the original inclusion (17).

Proof. To show this, we consider the bounds

(60)

$$\begin{aligned} \|\sigma(u, u', x, x', x'')\| &= \|v' - v - f(x') + f(x) + H[x' - x] - (H' - H_0)[x'' - x']\| \\ &\leq \|v' - v\| + \|-f(x') + f(x) + H[x' - x]\| + \|H' - H_0\| \|x'' - x'\| \\ &\leq \|v' - v\| + \psi_{(u, u')}^3(\|x' - x\|) + \frac{1}{2\alpha} \|x'' - x'\|, \end{aligned}$$

where the last inequality follows from (58) and the fact that we can shrink U if necessary so that $\|H' - H_0\| \leq \frac{1}{2\alpha}$. It follows from (60) that the σ -approximation bound holds in the *R*-convergence condition for G_{loc} with the bivariate convergence function

$$\phi_{(u, u')}^1(r, r') := \|v' - v\| + \psi_{(u, u')}^3(r) + \frac{1}{2\alpha} r'.$$

The Aubin continuity assumption yields scalars $\epsilon_y \geq 0$ and $\epsilon_x \geq 0$ such that

$$\begin{aligned} (H_0 + F)^{-1}(y') \cap \mathbb{B}(x_0; \epsilon_x) &\subseteq (H_0 + F)^{-1}(y'') + \alpha \|y'' - y'\| \mathbb{B}(0; 1) \\ &\text{for all } y', y'' \in \mathbb{B}(y_0; \epsilon_y). \end{aligned}$$

The nonsingularity bound in the *R*-convergence condition for G_{loc} with $\phi_{(u, u')}^2(r, r') := \alpha r'$ then follows from the pair of equivalences

$$\begin{aligned} x' \in G(u, x) &\iff x' \in (H_0 + F)^{-1}(y') \text{ with } y' := v - f(x) - H[x' - x] + H_0[x'], \\ x'' \in G(u', x') &\iff x'' \in (H_0 + F)^{-1}(y'') \text{ with } y'' := v' - f(x') - H'[x'' - x'] + H_0[x''], \end{aligned}$$

and the fact that $\sigma(u, u', x, x', x'') = y'' - y'$ in this case. The result then follows from Theorem 3.4 and Proposition 2.0.2. $\square$

To illustrate the details of the R -convergence rate that results from Corollary 5.2.3, we will consider the special case of $\psi_{(u,u')}^1$ which arises when we use the trivial parameter update $v' = v$ and the convergence functions $\psi_{(u,u')}^3$ from (59). It follows that the sequence of scalars $\{r_k\}$ from (41) in this case is defined simply by $r_{k+1} = 2\alpha\epsilon r_k^\gamma$. We can apply this inductively to get the expression

$$r_k = (2\alpha\epsilon)^{\frac{1}{1-\gamma}} \left(\frac{r_0}{(2\alpha\epsilon)^{\gamma-1}} \right)^{\gamma^k}.$$

Thus the infinite series $\sum_{i=0}^{\infty} r_i$ is geometric, and as long as $r_0 < (2\alpha\epsilon)^{\gamma-1}$, we know that the residual sums $s_k := \sum_{i=k}^{\infty} r_i$ from (42) in this case satisfy

$$s_k = c\xi^{\gamma^k} \text{ for } c := (2\alpha\epsilon)^{\frac{(\gamma-1)^2-1}{\gamma-1}} \text{ and } \xi := \frac{r_0}{(2\alpha\epsilon)^{\gamma-1}}.$$

Therefore, (42) in this case gives

$$\|x_k - \bar{x}\| \leq c\xi^{\gamma^k} \text{ as long as } r_0 := \|x_1 - x_0\| < (\alpha\epsilon)^{\gamma-1},$$

which means an R -convergence rate of γ .

Acknowledgment. This paper was improved significantly by the feedback from an anonymous referee.

REFERENCES

- [1] D. AZÉ AND C. C. CHOU, *On a Newton type iterative method for solving inclusions*, Math. Oper. Res., 20 (1995), pp. 790–800.
- [2] A. L. DONTCHEV, *Uniform convergence of the Newton method for Aubin continuous maps. Well-posedness and stability of variational problems*, Serdica Math. J., 22 (1996), pp. 283–296.
- [3] A. L. DONTCHEV, *Lipschitzian stability of Newton's method for variational inclusions*, in System Modeling and Optimization, Kluwer Academic Publishers, Boston, 2000, pp. 119–147.
- [4] A. L. DONTCHEV AND R. T. ROCKAFELLAR, *Characterizations of Lipschitzian stability in nonlinear programming*, in Mathematical Programming with Data Perturbations, Lecture Notes in Pure and Appl. Math. 195, Dekker, New York, 1998, pp. 65–82.
- [5] A. B. LEVY, *Solution sensitivity from general principles*, SIAM J. Control Optim., 40 (2001), pp. 1–38.
- [6] T. PENNANEN, *Local convergence of the proximal point algorithm and multiplier methods without monotonicity*, Math. Oper. Res., 27 (2002), pp. 170–191.
- [7] F.-A. POTRA AND V. PTÁK, *Nondiscrete Induction and Iterative Processes*, Pitman, Boston, 1984.
- [8] S. M. ROBINSON, *Newton's method for a class of nonsmooth functions*, Set-Valued Anal., 2 (1994), pp. 291–305.
- [9] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [10] R. T. ROCKAFELLAR, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
- [11] R. T. ROCKAFELLAR, *Extended nonlinear programming*, in Nonlinear Optimization and Related Topics, Appl. Optim. 36, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000, pp. 381–391.
- [12] R. T. ROCKAFELLAR AND R. J.-B. WETS, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
- [13] J. L. ZHOU AND A. L. TITS, *An SQP algorithm for finely discretized continuous minimax problems and other minimax problems with many objective functions*, SIAM J. Optim., 6 (1996), pp. 461–487.
- [14] H. XU, *Set-valued approximations and Newton's methods*, Math. Program., 84 (1999), pp. 401–420.

WEAK SHARP MINIMA IN MULTICRITERIA LINEAR PROGRAMMING*

SIEN DENG[†] AND X. Q. YANG[‡]

Abstract. In this short note, we study the existence of weak sharp minima in multicriteria linear programming. It is shown that weak sharp minimality holds for certain residual functions and gap functions.

Key words. multicriteria linear optimization, weak sharp minima, natural residual function, gap function

AMS subject classifications. 90C29, 90C31, 90C05

DOI. 10.1137/S1052623403434401

1. Introduction. Consider the following multicriteria linear programming (MCLP) problem:

$$\begin{aligned} (\mathcal{P}) \quad & \min \quad Cx \\ & \text{s.t.} \quad x \in X, \end{aligned}$$

where $Cx = (c_1^T x, \dots, c_m^T x)^T$, and $X \subset \mathbb{R}^n$ is a nonempty polyhedral convex set.

Let $W = \mathbb{R}^m \setminus (-\text{int } \mathbb{R}_+^m)$, where int denotes interior. A vector $\bar{x} \in X$ is a weakly efficient solution of problem $(\mathcal{P})$ if and only if

$$Cx - C\bar{x} \in W \quad \forall x \in X.$$

Denote by E_w the set of all weakly efficient solutions to problem $(\mathcal{P})$.

Weak sharp minima play important roles in mathematical programming. They have been well studied for scalar minimization problems. See [2, 3, 9] and references therein. In this short note, we study weak sharp minima for MCLP problems. In scalar convex optimization, as is well known, weak sharp minimality holds for linear programming, certain quadratic programming, and linear complementarity problems [2]; it is shown that convexity and polyhedrality of solution sets are very important for the existence of weak sharp minima. Unlike the scalar case, for MCLP problems, E_w is not convex in general. However, E_w is a finite union of polyhedral convex sets. By examining such structures carefully, we are able to show that weak sharp minimality holds for certain natural residual functions associated with the underlying MCLP problems. Specifically, we obtain weak sharp minima of the solution set for the natural residual function $\text{dist}(Cx \mid CE_w)$ and for the associated gap function, respectively. To the best of our knowledge, these results are new and should be useful in sensitivity analysis and in designing algorithms for solving multicriteria optimization problems.

The notation that we employ is for the most part the same as that in [7, 8]. A partial list is provided below for the reader's convenience.

*Received by the editors September 7, 2003; accepted for publication (in revised form) April 14, 2004; published electronically December 30, 2004. This research was partially supported by the Research Grant Council of Hong Kong (B-Q475).

<http://www.siam.org/journals/siopt/15-2/43440.html>

[†]Department of Mathematical Sciences, Northern Illinois University, DeKalb, IL 60115 (deng@math.niu.edu).

[‡]Department of Applied Mathematics, The Hong Kong Polytechnic University, Kowloon, Hong Kong (mayangxq@polyu.edu.hk).

We denote the set $\{1, 2, \dots, m\}$ by $[1, m]$. For any $J \subset [1, m]$, we define C_J to be the matrix obtained from C by removing rows whose indexes are in $[1, m] \setminus J$. For any given $J \subset [1, m]$, the subproblem $\min C_J x$ s.t. $x \in X$, and its solution set, are denoted by $(\mathcal{P}(\mathcal{J}))$ and $E_w(J)$, respectively.

2. Weak sharp minima. In this section, we discuss two existence results of weak sharp minima. We begin by reviewing a basic result on structures of solution sets for MCLP that was given in [1]. See also [6, Thm. 3.3, p. 96].

THEOREM 2.1 (see [1]). *Let Λ be the canonical simplex in $\mathbb{R}^m$ [8, p. 318]. Then there are finitely many vectors $\lambda(1), \dots, \lambda(r)$ of Λ such that $E_w = \cup_{k=1}^r S_{\lambda(k)}$, where $S_{\lambda(k)} = \arg \min_{x \in X} \lambda(k)^T Cx$.*

2.1. Weak sharp minima for $\text{dist}(Cx \mid CE_w)$. For $y \in \mathbb{R}^n$, define

$$\text{dist}(Cy \mid CE_w) = \inf_{x \in E_w} \max_{i \in [1, m]} |c_i y - c_i x|.$$

It is easy to verify that $\text{dist}(Cx \mid CE_w) = 0$ and $x \in X$ if and only if $x \in E_w$. So $\text{dist}(Cx \mid CE_w)$ serves as a natural residual function for problem $(\mathcal{P})$. When $m = 1$ (the scalar linear programming case), $\text{dist}(Cx \mid CE_w) = Cx - f_{\min}$ for $x \in X$, where $f_{\min}$ is the optimal value of $(\mathcal{P})$. We say that E_w is a set of weak sharp minima for the function $\text{dist}(Cx \mid CE_w)$ if there is some positive constant τ such that

$$\text{dist}(x \mid E_w) \leq \tau \text{dist}(Cx \mid CE_w) \quad \forall x \in X,$$

where $\text{dist}(x \mid E_w) = \inf_{z \in E_w} \max_{i \in [1, m]} |z_i - x_i|$.

The first main result of this note follows.

THEOREM 2.2. *Suppose that E_w is nonempty. Then E_w is a set of weak sharp minima for $\text{dist}(Cx \mid CE_w)$.*

Proof. For any $y \notin E_w$, there is some $\bar{x} \in E_w$ such that $\text{dist}(Cy \mid CE_w) = \max_{i \in [1, m]} |c_i y - c_i \bar{x}|$. By Theorem 2.1, there is some $\lambda(j) \in \{\lambda(1), \dots, \lambda(r)\}$ such that

$$\bar{x} \in \arg \min_{x \in X} \lambda(j)^T Cx.$$

Hence, by Hoffman's lemma [5], we have

$$\begin{aligned} \text{dist}(y \mid E_w) &\leq \text{dist}(y \mid S_{\lambda(j)}) \\ &\leq \tau(j) \lambda(j)^T (Cy - C\bar{x}) \\ &\leq \tau \|\lambda(j)\|_1 \|Cy - C\bar{x}\|_\infty \\ &\leq \tau \text{dist}(Cy \mid CE_w), \end{aligned}$$

where $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are l_1 and l_∞ norms, respectively, and $\tau = \max_{i=1}^r \tau(i)$. The last inequality follows from $\|\lambda(j)\|_1 = 1$. $\square$

2.2. Weak sharp minima for the gap function. For any $y \in X$, define the gap function as follows:

$$g(y) = \max_{x \in X} \left\{ \min_{i \in [1, m]} (c_i y - c_i x) \right\}.$$

The gap function was first introduced in [4]. The gap function g_J for the subproblem $(\mathcal{P}(\mathcal{J}))$ is defined accordingly. It is easy to see that $g(y) \geq 0$ for any $y \in X$, and

$g(y) = 0$ if and only if $y \in E_w$. This gap function can also be expressed as follows:

$$\begin{aligned} g(y) &= \max_{x \in X} \left\{ \min_{i \in [1, m]} (c_i y - c_i x) \right\} \\ &= \max_{x \in X} \left\{ - \max_{i \in [1, m]} (c_i x - c_i y) \right\} \\ &= - \min_{x \in X} \left\{ \max_{i \in [1, m]} (c_i x - c_i y) \right\}. \end{aligned}$$

The following proposition gives some basic properties of gap functions.

PROPOSITION 2.3. *Suppose that E_w is nonempty. Then the following is true for subproblems $(\mathcal{P}(\mathcal{J}))$ and gap functions:*

(a) g is a finite concave function.

(b) $\min_{J \subset [1, m]} g_J(y) = g(y)$ for all $y \in X$, and the following inequality holds:

$$(1) \quad g(y) \geq \min_{J \subset [1, m]} \text{dist}(C_J y \mid C_J(E_w(J))) \quad \forall y \in X.$$

Proof. (a) Since $c_i x - c_i y$ is jointly convex in x and y , so is $\max_{i \in [1, m]} (c_i x - c_i y)$. This, along with the nonemptiness of E_w , implies that

$$\min_{x \in X} \max_{i \in [1, m]} (c_i x - c_i y)$$

is a finite convex function in y . So g is a finite concave function.

(b) For any $J \subset [1, m]$, by definition, we always have $g_J(y) \geq g(y)$ for all $y \in X$. On the other hand, for any given $y \in X$, let $f(x) = \max_{i \in [1, m]} (c_i x - c_i y)$. Then f is bounded below on X since $E_w \neq \emptyset$. But f is piecewise linear. So $\arg \min_{x \in X} f(x)$ is nonempty. Suppose that $\hat{x} \in \arg \min_{x \in X} f(x)$. Then

$$0 \in \partial f(\hat{x}) + N_X(\hat{x}),$$

where $\partial f(\hat{x}) = \text{co}_{i \in I} \{c_i\}$ and $I = \{i \in [1, m] \mid c_i \hat{x} - c_i y = f(\hat{x})\}$. It follows that there is some $\lambda \in \Lambda(I)$ (the canonical simplex in $\mathbb{R}^{|I|}$) such that $\hat{x} \in \arg \min_{x \in X} \lambda^T C_I x$. This implies that $\hat{x} \in E_w(I) \subset E_w$, which, in turn, implies that

$$g(y) = - \min_{x \in E_w} \left\{ \max_{i \in [1, m]} (c_i x - c_i y) \right\}.$$

Let $f_I(x) = \max_{i \in I} (c_i x - c_i y)$. Then

$$0 \in \partial f_I(\hat{x}) + N_X(\hat{x}).$$

So $g(y) = g_I(y)$. This shows that

$$g(y) = \min_{J \subset [1, m]} g_J(y) \quad \forall y \in X.$$

With this given $\hat{x}$ and the choice of I , we have $g(y) = c_i y - c_i \hat{x}$ for any $i \in I$. Thus, for any $\lambda \in \Lambda(I)$,

$$g(y) = \lambda^T (C_I y - C_I \hat{x}) = \|C_I y - C_I \hat{x}\|_\infty.$$

Inequality (1) follows from

$$\|C_I y - C_I \hat{x}\|_\infty \geq \text{dist}(C_I y \mid C_I(E_w(I))) \geq \min_{J \subset [1, m]} \text{dist}(C_J y \mid C_J(E_w(J))). \quad \square$$

We say that E_w is a set of weak sharp minima for the gap function g if there is some positive constant γ such that

$$\text{dist}(x \mid E_w) \leq \gamma g(x) \quad \forall x \in X.$$

To prove the second main result of this note, we need to use the following result on structures of solution sets for multicriteria convex programming problems.

PROPOSITION 2.4. *For any $J \subset [1, m]$, let $E_w(J)$ be the set of weakly efficient solutions to minimizing $C_J x$ s.t. $x \in X$. Then*

$$\cup_{J \subset [1, m]} E_w(J) = E_w.$$

The second main result now follows.

THEOREM 2.5. *Suppose that E_w is nonempty. Then E_w is a set of weak sharp minima for the gap function g .*

Proof. For any given $I \subset [1, m]$ with $E_w(I)$ nonempty, by Theorem 2.2 there is a $\tau(I) > 0$ such that

$$\text{dist}(y \mid E_w(I)) \leq \tau(I) \text{dist}(C_I y \mid C_I E_w(I)) \quad \forall y \in X.$$

By Proposition 2.4, we have $E_w = \cup_{I \subset [1, m]} E_w(I)$, and it follows that for any $y \in X$,

$$\text{dist}(y \mid E_w) \leq \min_{I \subset [1, m]} \text{dist}(y \mid E_w(I)).$$

So

$$\begin{aligned} \text{dist}(y \mid E_w) &\leq \min_{I \subset [1, m]} \text{dist}(y \mid E_w(I)) \\ &\leq \min_{I \subset [1, m]} (\tau(I) \text{dist}(C_I y \mid C_I E_w(I))) \\ &\leq \left(\max_{I \subset [1, m]} \tau(I) \right) \left(\min_{I \subset [1, m]} \text{dist}(C_I y \mid C_I E_w(I)) \right) \\ &\leq \left(\max_{I \subset [1, m]} \tau(I) \right) g(y) \quad (\text{by (1) in Proposition 2.3}). \end{aligned}$$

Setting $\hat{\tau} = \max_{I \subset [1, m]} \tau(I)$ yields the desired result. $\square$

Note. It is easy to see that $\text{dist}(C y \mid C E_w) \geq \min_{I \subset [1, m]} \text{dist}(C_I y \mid C_I E_w(I))$. However, we don't know how $g(y)$ and $\text{dist}(C y \mid C E_w)$ are related to each other.

We conclude this section with the following example, which illustrates that the technique of parametric linear programming will not yield the sharp results in Theorems 2.2 and 2.5.

Consider the following MCLP:

$$\begin{aligned} (\mathcal{Q}) \quad & \min \quad Cx \\ & \text{s.t.} \quad Ax \leq 0, \end{aligned}$$

where $C = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ and $A = \begin{pmatrix} -1 & 1 \\ -1 & -1 \end{pmatrix}$. Its associated scalar linear programming problems are to minimize $c_\mu^T x$ s.t. $Ax \leq 0$, where

$$c_\mu = \mu \begin{pmatrix} 1 \\ 1 \end{pmatrix} + (1 - \mu) \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 \\ 2\mu - 1 \end{pmatrix},$$

and $0 \leq \mu \leq 1$. Then

$$S_\mu = \begin{cases} \mathbb{R}_+(1, 1) & \text{if } \mu = 0, \\ (0, 0) & \text{if } 0 < \mu < 1, \\ \mathbb{R}_+(1, -1) & \text{if } \mu = 1. \end{cases}$$

For $0 < \mu < 1$, consider the distance between the vector $\bar{x} = (a, a)^T$, where $a > 0$ and $S_\mu = (0, 0)$. We have $\text{dist}(\bar{x} \mid S_\mu) = a$, and $c_\mu^T \bar{x} = 2\mu a$. So

$$\tau(\mu) \geq \frac{\text{dist}(\bar{x} \mid S_\mu)}{c_\mu^T \bar{x}} = (2\mu)^{-1} \rightarrow \infty \quad \text{as } \mu \downarrow 0.$$

Acknowledgments. We thank the referees for their helpful comments.

REFERENCES

- [1] K. J. ARROW, E. W. BARANKIN, AND D. BLACKWELL, *Admissible points of convex sets*, in Contributions to the Theory of Games, Vol. 2, H. W. Kuhn and A. W. Tucker, eds., Princeton University Press, Princeton, NJ, 1953, pp. 87–91.
- [2] J. V. BURKE AND M. C. FERRIS, *Weak sharp minima in mathematical programming*, SIAM J. Control Optim., 31 (1993), pp. 1340–1359.
- [3] J. V. BURKE AND S. DENG, *Weak sharp minima revisited. I. Basic theory*, Control Cybernet., 31 (2002), pp. 439–469.
- [4] C. J. GOH AND X. Q. YANG, *On the solution of vector variational inequality*, in Proceedings of the Fourth International Conference on Optimization Techniques and Applications, L. Caccetta et al., eds., Curtin University of Technology, Perth, Australia, 1998, pp. 1158–1164.
- [5] A. J. HOFFMAN, *On approximate solutions of systems of linear inequalities*, J. Research Nat. Bur. Standards, 49 (1952), pp. 263–265.
- [6] D. T. LUC, *Theory of Vector Optimization*, Springer-Verlag, Berlin, 1989.
- [7] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [8] R. T. ROCKAFELLAR AND R. J.-B. WETS, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
- [9] M. STUDNIARSKI AND D. E. WARD, *Weak sharp minima: Characterizations and sufficient conditions*, SIAM J. Control Optim., 38 (1999), pp. 219–236.

ERROR ESTIMATES FOR APPROXIMATE OPTIMIZATION BY THE EXTENDED RITZ METHOD*

VĚRA KŮRKOVÁ[†] AND MARCELLO SANGUINETI[‡]

Abstract. An alternative to the classical Ritz method for approximate optimization is investigated. In the extended Ritz method, sets of admissible solutions are approximated by their intersections with sets of linear combinations of all n -tuples of functions from a given basis. This alternative scheme, called variable-basis approximation, includes functions computable by trigonometric polynomials with free frequencies, free-node splines, neural networks, and other nonlinear approximating families. Estimates of rates of approximate optimization by the extended Ritz method are derived. Upper bounds on rates of convergence of suboptimal solutions to the optimal one are expressed in terms of the degree n of variable-basis functions, the modulus of continuity of the functional to be minimized, the modulus of Tikhonov well-posedness of the problem, and certain norms tailored to the type of basis. The results are applied to convex best approximation and to kernel methods in machine learning.

Key words. functional optimization, rates of convergence of suboptimal solutions, (extended) Ritz method, curse of dimensionality, convex best approximation problems, learning from data by kernel methods

AMS subject classifications. 49M15, 90B99, 90C90, 41A25, 68T05, 41A50

DOI. 10.1137/S1052623403426507

1. Introduction. In many high-dimensional optimization problems (routing in communications networks, stochastic optimal control, management of water resources, large-scale traffic networks [13, 24, 46, 81], etc.), optimal solutions cannot be found analytically or, even when they can be found, they may not be computable efficiently by numerical methods. However, in some cases, optimal solutions can be approximated by suboptimal ones. In the classical Ritz method [37, Chapter 8], such an approximation is accomplished by a sequence of solutions over intersections of the original admissible set with a nested family of linear subspaces of increasing dimensionality.

Although linear approximation methods have many convenient properties, their practical applications are often limited by the “curse of dimensionality” [14], i.e., an exponential growth, as a function of the number of variables, of the dimension a linear subspace would need to achieve a desired accuracy of approximation of the optimal solution. Experimental results indicate that the Ritz method is often unable to deal efficiently with high-dimensional optimization tasks [81]. Theoretical results estimating rates of convergence of the Ritz method for the case of admissible solutions dependent on only one variable were derived in [6, 19, 27, 36, 41, 73], but we have not found in the literature any explicit estimate for the multivariable case.

Since the late 1980s, neural networks became a successful alternative to linear

*Received by the editors April 20, 2003; accepted for publication (in revised form) March 11, 2004; published electronically February 19, 2005. Collaboration between the authors was supported by the Scientific Agreement Italy-Czech Republic, Area MC 6, Project 22.

<http://www.siam.org/journals/siopt/15-2/42650.html>

[†]Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07, Prague 8 Czech Republic (vera@cs.cas.cz). This author was partially supported by GA ČR grant 201/02/0428.

[‡]Department of Communications, Computer, and System Sciences (DIST), University of Genoa, Via Opera Pia 13, 16145 Genova, Italy (marcello@dist.unige.it). This author was partially supported by a PRIN grant from the Italian Ministry of University and Research (Project “New Techniques for the Identification and Adaptive Control of Industrial Systems”).

methods for approximate solutions of high-dimensional optimization problems (see, e.g., [18, 23, 47, 61, 62, 74]). Also, a new branch of nonlinear approximation theory investigating approximation capabilities of neural networks was developed [11, 12, 21, 28, 38, 45, 50, 51, 52, 53, 54, 55, 56, 57, 58]. In a series of papers [3, 5, 8, 9, 10, 64, 65, 66, 80, 81], a new method of approximate optimization was developed, called in [81] the *extended Ritz method*. In these papers, approximate solutions were used that were obtained over restrictions of sets of admissible solutions to linear combinations of all n -tuples of functions with varying “free” parameters, instead of linear combinations of first n functions from a basis with fixed ordering as in the classical Ritz method. In the extended Ritz method, a nested family of linear subspaces of increasing dimensionality, which in the Ritz method approximates the set of admissible solutions, is replaced by a nested family of nonlinear approximating sets called *variable-basis functions*. The variable-basis approximation scheme includes a variety of nonlinear approximators such as free-nodes splines [31, Chapter 13], polynomials with free frequencies and phases [32], radial-basis-function networks with variable variances and centers [38], and feedforward neural networks [48, 56].

For bases formed by functions computable by neural-network units or, more generally, for bases consisting of functions parameterized by vectors from finite-dimensional Euclidean spaces, the extended Ritz method reduces the original optimization task to the problem of finding optimal values of finitely many parameters. This is a nonlinear programming problem, for which various algorithms are available [1, 4, 16, 39, 40, 76, 79].

The extended Ritz method with such bases was successfully tested on a variety of problems with admissible solutions dependent on a large number of variables: stochastic optimal control [64, 65, 66, 80] and optimal estimation of state variables [3] in nonlinear dynamic systems with a large number of state variables, team optimal control [8], optimal control of freeway traffic [81], routing in large-scale communication networks [9, 10], optimal fault diagnosis [5], etc. In these applications, admissible sets of variable-basis functions were used, for which the degree n necessary to guarantee a fixed approximation accuracy grows only polynomially with the number of variables of admissible solutions. Numerical comparisons with the classical Ritz method showing advantages of the extended Ritz method were made in [81].

Motivated by these experimental results, we investigate the extended Ritz method theoretically. We derive upper bounds on the speed of convergence of suboptimal solutions over nested families of variable-basis functions of increasing degree n to the optimal solution over the whole admissible set. The upper bounds depend on the degree of the variable-basis functions, a norm tailored to the type of the basis, the modulus of continuity of the functional to be minimized, and the modulus of Tichonov well-posedness of the problem. As our bounds are not merely asymptotic, they enable one to estimate the quality of suboptimal solutions achievable over admissible sets for any degree n (in particular, for n small enough to allow an efficient implementation of such suboptimal solutions).

By inspection of the derived estimates, we obtain some insights into optimization problems for which the extended Ritz method performs well. The critical term in our bounds is of the form $1/\sqrt{n}$ multiplied by a certain norm of the optimal solution. Such a norm is tailored to the basis used in the extended Ritz method. To keep this norm small with increasing number of variables of admissible solutions, one has to increase a certain type of regularity related to smoothness [12, 21, 50, 54].

We illustrate our results on two examples. In the first one, we apply them to the problem of convex best approximation, and in the second one, to learning from data

modeled as minimization of a regularized empirical error functional over a reproducing kernel Hilbert space.

The paper is organized as follows. Section 2 introduces basic concepts and some results from optimization theory, which are used throughout the paper. Section 3 describes the variable-basis approximation scheme and the extended Ritz method. Section 4 contains our main results on rates of convergence of the extended Ritz method, and section 5 contains their interpretation in the special case of convex problems. Sections 6 and 7 apply the derived estimates to convex best approximation and to kernel methods in machine learning, respectively. Section 8 contains a brief discussion. For the reader's convenience, we include an appendix containing some tools from nonlinear approximation theory that are used in the paper.

2. Preliminaries. By a normed linear space $(X, \|\cdot\|)$, we mean a *real normed linear space*. We write only X when it is clear which norm is used. $\mathcal{R}$ denotes the set of real numbers and $\mathcal{R}_+$ the set of positive reals. For a positive integer d , $\Omega \subseteq \mathcal{R}^d$, and $p \in [1, \infty)$, $(\mathcal{L}_p(\Omega), \|\cdot\|_p)$ denotes the space of measurable, real-valued functions on Ω such that $\int_{\Omega} |f(x)|^p dx < \infty$ endowed with the $\mathcal{L}_p$ -norm.

A ball and a sphere of radius r centered at $h \in X$ are denoted by $B_r(h, \|\cdot\|) = \{f \in X : \|f - h\| \leq r\}$ and $S_r(h, \|\cdot\|) = \{f \in X : \|f - h\| = r\}$, respectively. We write shortly $B_r(\|\cdot\|) = B_r(0, \|\cdot\|)$ and merely $B_r(h) = B_r(h, \|\cdot\|)$, $B_r = B_r(\|\cdot\|)$ when it is clear which norm is used; similarly for spheres.

A Banach space X is called *uniformly convex* if for any $\varepsilon \in (0, 2]$, there exists $\delta > 0$ such that if $\|f\| = \|g\| = 1$ and $\|(f + g)/2\| > 1 - \delta$, then $\|f - g\| < \varepsilon$ (i.e., whenever the midpoint of the line segment joining two points on the unit sphere approaches the sphere, then the endpoints of the segment must approach each other).

Sequences (of elements of linear spaces or sets) are denoted by $\{x_n\}$ instead of $\{x_n : n \in \mathcal{N}_+\}$, where $\mathcal{N}_+$ is the set of positive integers.

A functional $\Phi : X \rightarrow (-\infty, +\infty]$ is called *proper* if it is not identically equal to $+\infty$. The set $\text{dom } \Phi = \{f \in X : \Phi(f) < +\infty\}$ is called the *domain* of Φ .

Φ is *continuous* at $f \in \text{dom } \Phi$ if for all $\varepsilon > 0$, there exists $\eta > 0$ such that for every $g \in \text{dom } \Phi$, $\|f - g\| < \eta$ implies $|\Phi(f) - \Phi(g)| < \varepsilon$. The function $\alpha_f : \mathcal{R}_+ \rightarrow \mathcal{R}_+$ defined as $\alpha_f(t) = \sup\{|\Phi(f) - \Phi(g)| : f, g \in \text{dom } \Phi, \|f - g\| \leq t\}$ is the *modulus of continuity* of Φ at f . We write merely α instead of α_f when f is clear from the context. Φ is *Lipschitz continuous* on M with a Lipschitz constant c if for all $f, g \in M$, $|\Phi(f) - \Phi(g)| \leq c\|f - g\|$.

A functional Φ is *convex* on a convex set $M \subseteq X$ if for all $h, g \in M$ and all $\lambda \in [0, 1]$, $\Phi(\lambda h + (1 - \lambda)g) \leq \lambda\Phi(h) + (1 - \lambda)\Phi(g)$. Φ is *uniformly convex* on a convex set $M \subseteq X$ if there exists a nonnegative function $\delta : \mathcal{R}_+ \rightarrow \mathcal{R}_+$ such that $\delta(0) = 0$, for all $t > 0$ it satisfies $\delta(t) > 0$, and for all $h, g \in M$ and all $\lambda \in [0, 1]$ it satisfies $\Phi(\lambda h + (1 - \lambda)g) \leq \lambda\Phi(h) + (1 - \lambda)\Phi(g) - \lambda(1 - \lambda)\delta(\|h - g\|)$. Any such function δ is called a *modulus of convexity* of Φ (see, e.g., [59]).¹

Using standard notation [34], we denote by

$$(M, \Phi)$$

the problem of minimization of a functional Φ over a subset M of X . M is called a set of *admissible solutions* or *admissible set*. When both M and Φ are convex, (M, Φ) is called a *convex optimization problem*.

¹The terminology is not unified: some authors use the term “strictly uniformly convex” instead of “uniformly convex,” and they reserve the latter term for the case where $\delta : [0, +\infty) \rightarrow [0, +\infty)$ merely satisfies $\delta(0) = 0$ and for some $t_0 > 0$, $\delta(t_0) > 0$ (see, e.g., [78], [33, p. 10]).

A sequence $\{g_n\}$ of elements of M is called Φ -*minimizing over* M if $\lim_{n \rightarrow \infty} \Phi(g_n) = \inf_{g \in M} \Phi(g)$. By the definition of infimum, for any problem (M, Φ) where M is non-empty, there always exists a minimizing sequence. We denote by $\operatorname{argmin}(M, \Phi) = \{g^o \in M : \Phi(g^o) = \inf_{g \in M} \Phi(g)\}$ the set of *minimum points* of the problem (M, Φ) , and for $\varepsilon > 0$, we denote by $\operatorname{argmin}_\varepsilon(M, \Phi) = \{g^\varepsilon \in M : \Phi(g^\varepsilon) < \inf_{g \in M} \Phi(g) + \varepsilon\}$ the set of its ε -*near minimum points*.

The following proposition summarizes elementary properties of uniformly convex functionals.

PROPOSITION 2.1. *Let $(X, \|\cdot\|)$ be a normed linear space, $M \subseteq X$ be convex, and Φ be a uniformly convex functional on M with a modulus of convexity δ . Then the following hold:*

- (i) *if Ψ is convex on M , then $\Phi + \Psi$ is uniformly convex on M with a modulus of convexity δ ;*
- (ii) *if $\Phi : X \rightarrow \mathcal{R}$, then for every $f \in X$ the translated functional $\Phi(\cdot - f)$ is uniformly convex on $M - f$ with a modulus of convexity δ ;*
- (iii) *if $g^o \in \operatorname{argmin}(M, \Phi)$, then for every $g \in M$, $\delta(\|g - g^o\|) \leq \Phi(g) - \Phi(g^o)$;*
- (iv) *if $(X, \|\cdot\|)$ is a Hilbert space, then the functional $\|\cdot\|^2 : X \rightarrow \mathcal{R}$ is uniformly convex with a modulus of convexity $\delta(t) = t^2$.*

Proof. (i) and (ii) follow directly from the definitions.

(iii) By the definition of uniformly convex functional, for every $\lambda \in [0, 1]$ we have $\lambda(1-\lambda)\delta(\|g - g^o\|) \leq \lambda\Phi(g) + (1-\lambda)\Phi(g^o) - \Phi(\lambda g + (1-\lambda)g^o)$. As $\Phi(g^o) \leq \Phi(\lambda g + (1-\lambda)g^o)$, we get $\lambda(1-\lambda)\delta(\|g - g^o\|) \leq \lambda\Phi(g) + (1-\lambda)\Phi(g^o) - \Phi(g^o) = \lambda(\Phi(g) - \Phi(g^o))$. Hence $(1-\lambda)\delta(\|g - g^o\|) \leq \Phi(g) - \Phi(g^o)$. Taking the infimum over λ , we obtain $\delta(\|g - g^o\|) \leq \Phi(g) - \Phi(g^o)$.

(iv) It is easy to check that for every $h, g \in X$ and $\lambda \in [0, 1]$, we have $\|\lambda h + (1-\lambda)g\|^2 \leq \lambda\|h\|^2 + (1-\lambda)\|g\|^2 - \lambda(1-\lambda)\|h - g\|^2$. $\square$

The problem (M, Φ) is *Tikhonov well-posed* if it has a unique minimum to which every minimizing sequence converges [34, p. 1]. The *modulus of Tikhonov well-posedness* of (M, Φ) at a minimum point g^o is a function $\xi_{g^o} : \mathcal{R}_+ \rightarrow \mathcal{R}_+$ such that for every $t \in \mathcal{R}_+$, $\xi_{g^o}(t) = \inf_{g \in M \cap S_t(g^o)} \Phi(g) - \Phi(g^o)$.

The *linear span* of M is $\operatorname{span} M = \{\sum_{i=1}^n w_i g_i : w_i \in \mathcal{R}, g_i \in M, n \in \mathcal{N}_+\}$. The *convex hull* of M is $\operatorname{conv} M = \{\sum_{i=1}^n w_i g_i : w_i \in [0, 1], \sum_{i=1}^n w_i = 1, g_i \in M, n \in \mathcal{N}_+\}$. The *topological interior* of M is $\operatorname{int} M = \{g \in M : (\exists \varepsilon > 0) (B_\varepsilon(g) \subset M)\}$ and its *closure* is $\operatorname{cl} M = \{f \in X : (\forall \varepsilon > 0) (B_\varepsilon(f) \cap M) \neq \emptyset\}$. If $\operatorname{cl} M = Y$ for $Y \subseteq X$, then M is said to be *dense* in Y . The *diameter* of M is defined as $\operatorname{diam} M = \sup\{\|f - g\| : f, g \in M\}$.

For a subset M of a normed linear space, its *affine hull* is defined as $\operatorname{aff} M = \{\sum_{i=1}^n w_i g_i : w_i \in \mathcal{R}, \sum_{i=1}^n w_i = 1, g_i \in M, n \in \mathcal{N}_+\}$. An element $g \in X$ is called a *relative interior point* of $M \subseteq X$ if it is an interior point of M in the topological sense with respect to the topology induced on $\operatorname{aff} M$. The set of all relative interior points of M is called the *relative interior* of M and denoted by $\operatorname{ri} M$. Thus, $\operatorname{ri} M = \{g \in \operatorname{aff} M : \exists \varepsilon > 0, B_\varepsilon(g) \cap \operatorname{aff} M \subseteq M\}$. Note that $\operatorname{ri} M$ is the interior of M as a subset of its affine hull, instead of as a subset of the whole space X .

The *Minkowski functional* of $M \subseteq X$ is the functional $p_M : X \rightarrow [0, +\infty]$ defined for every $f \in X$ as

$$p_M(f) = \inf\{\lambda \in \mathcal{R}_+ : f/\lambda \in M\}.$$

M is called *absorbing* if $\operatorname{dom} p_M = X$. For every M , p_M is positively homogeneous, and if M is convex, then p_M is convex, too. The following proposition states elemen-

tary properties of Minkowski functionals of convex sets containing zero, which will be used in our proofs.

PROPOSITION 2.2. *Let $(X, \|\cdot\|)$ be a normed linear space, M be a subset of X containing 0, and $r_0 = \sup\{r > 0 : B_r(\|\cdot\|) \subseteq M\}$. Then the following hold:*

- (i) *if M is convex, then $M \subseteq \{f \in X : p_M(f) \leq 1\}$;*
- (ii) *if M is convex, then $\{f \in X : p_M(f) < 1\} \subseteq M$;*
- (iii) *if M is closed and convex, then $M = \{f \in X : p_M(f) \leq 1\}$;*
- (iv) *if $0 \in \text{int } M$, then $\text{dom } p_M = X$;*
- (v) *if $0 \in \text{int } M$ and $r_0 < \infty$, then for every $f \in \text{dom } p_M$, $p_M(f) \leq \|f\|/r_0$;*
- (vi) *if M is convex and $0 \in \text{int } M$, then p_M is Lipschitz on X with the Lipschitz constant $c = 1/r_0$ if $r_0 < \infty$ and $c = 0$ if $r_0 = \infty$.*

Proof. (i) By the definition of p_M , $f \in M$ implies $p_M(f) \leq 1$, and so $M \subseteq \{f \in X : p_M(f) \leq 1\}$.

(ii) Let $f \in X$ be such that $p_M(f) < 1$. By the definition of p_M , there exists $\lambda < 1$ such that $f/\lambda \in M$. As M is convex and $0 \in M$, $f = \lambda(f/\lambda) + (1-\lambda)0 \in M$.

(iii) By (i) and (ii), it is sufficient to check that for every $f \in X$ with $p_M(f) = 1$, $f \in M$. By the definition of p_M , there exists a sequence $\{\lambda_i\}$ such that $\lim_{i \rightarrow \infty} \lambda_i = 1$ and for every i , $f/\lambda_i \in M$. As M is closed and $f = \lim_{i \rightarrow \infty} (f/\lambda_i)$, we have $f \in M$.

(iv) As $0 \in \text{int } M$, there exists $r > 0$ such that $B_r(\|\cdot\|) \subseteq M$. So for every $f \in B_r(\|\cdot\|)$, $p_M(f) \leq 1$. Let $g \in X$. Then, $p_M(g) = p_M(r\|g\|(g/r\|g\|))$ and by the positive homogeneity of p_M , $p_M(g) = (\|g\|/r)p_M(r(g/\|g\|))$. As $\|r(g/\|g\|)\| = r$, we have $r(g/\|g\|) \in B_r(0)$ and so $p_M(g) = (\|g\|/r)p_M(r(g/\|g\|)) \leq \|g\|/r < \infty$.

(v) By the definition of p_M , for every $r > 0$ such that $B_r(\|\cdot\|) \subseteq M$ and every $f \in \text{dom } p_M$ we have $p_M(f) \leq \|f\|/r$. Hence, by the definition of r_0 , $p_M(f) \leq \|f\|/r_0$.

(vi) When M is convex, p_M is also convex. By the convexity and positive homogeneity of p_M , we have $(1/2)p_M(f) = p_M((1/2)f) = p_M((1/2)g + (1/2)(f-g)) \leq (1/2)p_M(g) + (1/2)p_M(f-g)$. Thus, $p_M(f) - p_M(g) \leq p_M(f-g) \leq \|f-g\|/r_0$. By exchanging the roles of f and g , we obtain the inequality $-\|f-g\| \leq p_M(f) - p_M(g)$. Hence $|p_M(f) - p_M(g)| \leq \|f-g\|/r_0$. $\square$

3. Variable-basis approximation and the extended Ritz method. The classical *Ritz method* [37, p. 192] for approximate optimization replaces the problem (M, Φ) with a sequence of problems

$$\{(M \cap X_n, \Phi)\},$$

where for each n , X_n is an n -dimensional subspace of X . Under suitable conditions on Φ , M , and $\{X_n\}$, for every n there exists a minimum point g_n of the approximate problem $(M \cap X_n, \Phi)$, the sequence $\{g_n\}$ converges to some $g^o \in M$, and $\lim_{n \rightarrow \infty} \Phi(g_n) = \Phi(g^o)$.

Typically, the subspaces X_n are generated by the first n elements of a subset of X with a fixed linear ordering. So this approximation scheme can be called *fixed-basis approximation* in contrast to *variable-basis approximation*, which uses nonlinear approximating sets formed by linear combinations of at most n elements of a given subset G of X . Such sets are denoted by $\text{span}_n G = \{\sum_{i=1}^n w_i g_i : w_i \in \mathcal{R}, g_i \in G\}$. We call n the *degree of the variable-basis functions* in $\text{span}_n G$. The variable-basis approximation scheme includes free-node splines [31, Chapter 13], polynomials with free frequencies and phases [32], radial-basis-function networks with variable variances and centers [38], feedforward neural networks [48, 56], and so on.

In an alternative to the classical Ritz method, the problem (M, Φ) is approximated by a sequence of problems

$$\{M \cap \text{span}_n G, \Phi\}.$$

For G formed by parameterized families of the form $G = \{g_a : a \in A, g_a \in G\}$ with $A \subseteq \mathcal{R}^p$, this method was applied to a variety of optimization tasks considered in a series of papers [3, 5, 8, 9, 10, 64, 65, 66, 80] and in [81], where it was called the *extended Ritz method*. Here we use this term even more generally for an approximate optimization by the sequence of problems $\{M \cap \text{span}_n G\}$, where G is any subset of X .

Sets $\text{span}_n G$ are not convex and so, when the classical Ritz method is replaced with the extended one, minimum points over approximate admissible sets might not exist. However, the requirement of achieving a minimum point can be relaxed to a merely ε_n -near minimum, for which we shall formulate our estimates.

Typically, a basis is formed by functions parameterized by vectors from a finite-dimensional Euclidean space. For such bases, minimization over $M \cap \text{span}_n G$ reduces to a finite-dimensional nonlinear programming problem. Such a problem can be solved by algorithms based on gradient descent with stochastic perturbations [17, pp. 38–40, 103–104], genetic algorithms [39], simulated annealing [1], global stochastic optimization based on Monte Carlo [79] or quasi-Monte Carlo [77, Chapter 4] methods, etc. When a basis is formed by functions computable by neural-network units, various learning algorithms can be applied (see, e.g., [4, 16, 40, 76] and the references therein). In [5, 9, 10, 65, 66, 81], applications of some of these algorithms to the extended Ritz method are described and illustrated by numerical results showing the algorithms' effectiveness in a variety of cases.

A sequence of ε_n -near minimum points of Φ over $M \cap \text{span}_n G$ might converge to a minimum point of Φ over the whole M much faster than minima over $M \cap X_n$ in the classical Ritz method. Indeed, the union of subspaces spanned by all n -tuples of elements of a set G is “much larger” than a single n -dimensional subspace generated by the first n elements of G , and so the functional to be minimized might achieve over such unions of subspaces values that are closer to the infimum over the whole M .

To estimate rates of convergence of approximate solutions that can be obtained by the extended Ritz method, we take advantage of a result from nonlinear approximation theory by Maurey (reported in [68, Lemma 2, p.V.2]), Jones [45, p. 611], and Barron [12, p. 934, Lemma 1]. Here we use a reformulation of this result in terms of a norm tailored to a given basis G . Such a norm, called *G-variation* and denoted by $\|\cdot\|_G$, was introduced in [51] as an extension of the concept of variation with respect to half-spaces [11]. For a subset G of a normed linear space $(X, \|\cdot\|)$, *G-variation* is defined as the Minkowski functional of the set $\text{cl conv}(G \cup -G)$:

$$\|f\|_G = \inf \{c > 0 : c^{-1}f \in \text{cl conv}(G \cup -G)\}.$$

So *G-variation* of f measures how much the set G should be dilated to contain f in the closure of its symmetric convex hull. *G-variation* is a norm on the subspace $\{f \in X : \|f\|_G < \infty\} \subseteq X$ and

$$(3.1) \quad \|\cdot\| \leq s_G \|\cdot\|_G.$$

Indeed, if for $b > 0$, $f/b \in \text{cl conv}(G \cup -G)$, then $f/b = \lim_{\varepsilon \rightarrow 0} h_\varepsilon$, where $h_\varepsilon \in \text{conv}(G \cup -G)$, and so $\|h_\varepsilon\| \leq s_G$. Thus, $\|f\| \leq s_G b$. Hence, by the definition of $\|f\|_G$ we have $\|f\| \leq s_G \|f\|_G$.

When G is an orthonormal basis of a separable Hilbert space, G -variation is equal to the l_1 -norm with respect to G , which is defined for every $f \in X$ as $\|f\|_{1,G} = \sum_{g \in G} |f \cdot g|$ [55, 58]. Besides being a generalization of the notion of l_1 -norm, G -variation is also a generalization of the concept of total variation studied in integration theory.

The next theorem is a reformulation in terms of G -variation of the estimates derived for Hilbert spaces by Maurey, Jones, and Barron and of an extension of these estimates to $\mathcal{L}_p$ -spaces, $p \in (1, \infty)$, derived by Darken et al. [28, Theorem 5]. For the proof, see the appendix.

THEOREM 3.1. *Let $(X, \|\cdot\|)$ be a normed linear space, G be its bounded subset, and $s_G = \sup_{g \in G} \|g\|$. For every $f \in X$ and every positive integer n , the following estimates hold:*

(i) *if $(X, \|\cdot\|)$ is a Hilbert space, then*

$$(3.2) \quad \|f - \text{span}_n G\| \leq \sqrt{\frac{(s_G \|f\|_G)^2 - \|f\|^2}{n}};$$

(ii) *if $(X, \|\cdot\|) = (\mathcal{L}_p(\Omega), \|\cdot\|_p)$, where $p \in (1, \infty)$ and $\Omega \subseteq \mathcal{R}^d$ is open, then*

$$(3.3) \quad \|f - \text{span}_n G\| \leq \frac{2^{1/\bar{p}} s_G \|f\|_G}{n^{1/\bar{q}}},$$

where $q = p/(p-1)$, $\bar{p} = \min(p, q)$, and $\bar{q} = \max(p, q)$.

In contrast to some estimates of rates of linear (i.e., fixed-basis) approximation in spaces X of d -variable functions [67, pp. 232–233], where the denominator is of the form $n^{c/d}$ for some $c > 0$, in the bounds from Theorem 3.1, the denominator is $n^{1/2}$, independently of the number d of variables. However, for both fixed-basis and variable-basis approximation the numerators depend on d (see the discussion in section 8).

4. Rates of approximate optimization over variable-basis functions.

In this section, we investigate approximate solutions by $\{(M \cap \text{span}_n G, \Phi)\}$ of a problem (M, Φ) that has a minimum point. The existence of such a point is guaranteed for various convex problems in reflexive Banach spaces [26, 35, 59, 70]. Many problems that do not have minimum points can be transformed into problems with minimum points by regularization [34, p. 29]. So the following results apply to a wide class of regularized problems.

Let g° be a minimum point of the problem (M, Φ) to which the extended Ritz method based on an approximation of M by sets $M \cap \text{span}_n G$ is applied. As the existence of minimum points of approximating problems $(M \cap \text{span}_n G, \Phi)$ is not guaranteed, we consider ε_n -near minimum points. To estimate the speed of convergence of these ε_n -near minimum points to the minimum point g° of Φ over the whole M , we take advantage of Theorem 3.1. As this theorem estimates the distance of g° from $\text{span}_n G$ but not from $M \cap \text{span}_n G$, we construct an auxiliary sequence of elements of $M \cap \text{span}_n G$ using the following technical lemma. It extends [75, Lemma 3], proven there for finite-dimensional subspaces of a linear space, to subsets satisfying a kind of restricted homogeneity condition. Our lemma applies to a closed convex admissible set M containing zero. In the case when zero is in the interior of M , it gives an estimate in terms of the Lipschitz constant of the Minkowski functional of M . When M is a ball $B_r(\|\cdot\|)$, such a Lipschitz constant is equal to $1/r$.

LEMMA 4.1. *Let A and M be subsets of a normed linear space $(X, \|\cdot\|)$, M be closed and convex, $0 \in M$, and $\lambda A \subseteq A$ for all $\lambda \in [0, 1]$. Then for every $g \in M$ and every $f \in A$ with $p_M(f) < +\infty$, there exists $h \in M \cap A$ such that*

- (i) $\|h - g\| \leq \|f - g\| + \|g\| |p_M(f) - p_M(g)|$;
- (ii) if $0 \in \text{int } M$, then $\|h - g\| \leq (1 + c \|g\|) \|f - g\|$, where c is the Lipschitz constant of p_M on X .

Proof. (See Figure 4.1.) (i) When $f \in A \cap M$, the estimate holds trivially with $h = f$. If $f \in A - \text{cl } M$, then $f \neq 0$ and so we can set $h = \frac{p_M(g)}{p_M(f)} f$. Hence $p_M(h) = p_M(g) \leq 1$, and by Proposition 2.2(ii), $h \in M$. As $f \notin M$, by Proposition 2.2(iii) we have $p_M(f) > 1$. Thus $h = \frac{p_M(g)}{p_M(f)} f$ with $\frac{p_M(g)}{p_M(f)} < 1$ and $f \in A$, which implies $h \in A$. Hence $h \in A \cap M$ and $\|h - g\| = \|\frac{p_M(g)}{p_M(f)} f - g\| = \|\frac{p_M(g)}{p_M(f)}(f - g) - (1 - \frac{p_M(g)}{p_M(f)})g\| \leq \|\frac{p_M(g)}{p_M(f)}\| \|f - g\| + \|1 - \frac{p_M(g)}{p_M(f)}\| \|g\| < \|f - g\| + \|\frac{p_M(f) - p_M(g)}{p_M(f)}\| \|g\| < \|f - g\| + |p_M(f) - p_M(g)| \|g\|$.

(ii) If $0 \in \text{int } M$, then by Proposition 2.2(vi) p_M is Lipschitz continuous on X . Denoting by c its Lipschitz constant, we have $|p_M(f) - p_M(g)| \leq c \|f - g\|$. So $\|h - g\| \leq \|f - g\| + \|g\| |p_M(f) - p_M(g)|$ implies $\|h - g\| \leq (1 + c \|g\|) \|f - g\|$. $\square$

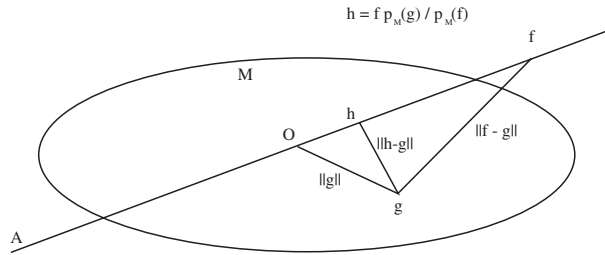


FIG. 4.1. The construction used in the proof of Lemma 4.1.

As we shall employ Lemma 4.1(ii) in the proof of the next theorem estimating rates of approximate optimization by the extended Ritz method, we need to assume that $0 \in \text{int } M$. Although this condition is restrictive, it still allows important applications. For example, when M is the whole ambient space X , one can apply the next theorem to Tikhonov's regularization [15, pp. 68–78] (see the application in section 7), and when M is a ball of some radius r in the norm $\|\cdot\|$, one can apply it to Ivanov's regularization [15, pp. 68–78]. Also, the case where M is a subspace of a Hilbert space X can be treated using the next theorem by replacing the ambient space X with M (since M , as a closed subspace of X , is a Hilbert space).

THEOREM 4.2. *Let $(X, \|\cdot\|)$ be a Hilbert space, M and G be its subsets, G be bounded, $s_G = \sup_{g \in G} \|g\|$, M be closed and convex, and $0 \in \text{int } M$. Let $\Phi : X \rightarrow (-\infty, +\infty]$ be a proper functional, $g^o \in \text{argmin}(M, \Phi)$, Φ be continuous at g^o with a modulus of continuity α , $\{\varepsilon_n\}$ be a sequence of positive reals, and $\{g_n\}$ be such that $g_n \in \text{argmin}_{\varepsilon_n}(M \cap \text{span}_n G, \Phi)$. Then p_M is Lipschitz on X , and if c is its Lipschitz constant, then the following estimates hold for every positive integer n :*

(i)

$$\inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^o) \leq \alpha \left((1 + c \|g^o\|) \sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}} \right);$$

- (ii) if $\|g^o\|_G < \infty$ and $\lim_{n \rightarrow \infty} \varepsilon_n = 0$, then $\{g_n\}$ is a Φ -minimizing sequence over M and

$$\Phi(g_n) - \Phi(g^o) \leq \alpha \left((1 + c\|g^o\|) \sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}} \right) + \varepsilon_n;$$

- (iii) if ξ is the modulus of Tikhonov well-posedness of (M, Φ) at g^o , then

$$\xi(\|g_n - g^o\|) \leq \alpha \left((1 + c\|g^o\|) \sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}} \right) + \varepsilon_n;$$

- (iv) if Φ is uniformly convex on M with a modulus of convexity δ , then

$$\delta(\|g_n - g^o\|) \leq \alpha \left((1 + c\|g^o\|) \sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}} \right) + \varepsilon_n.$$

The proof of Theorem 4.2 is based on Theorem 3.1(i) combined with the following idea: Lemma 4.1(ii) allows us to construct an auxiliary sequence $h_n^\varepsilon \in M \cap \text{span}_n G$ satisfying $\|g^o - h_n^\varepsilon\| \leq C\|g^o - \text{span}_n G\| + \varepsilon$ for a constant $C = 1 + c\|g^o\|$ dependent only on $\|g^o\|$ and on the Lipschitz constant c of p_M .

Proof. As $0 \in \text{int } M$, by Proposition 2.2(iv) and (v), $\text{dom } p_M = X$ and p_M is Lipschitz on X .

(i) For every n and every $\varepsilon > 0$, choose an ε -near best approximation f_n^ε of g^o from $\text{span}_n G$, i.e., $\|g^o - f_n^\varepsilon\| < \|g^o - \text{span}_n G\| + \varepsilon$. As M is closed, convex, $0 \in \text{int } M$, and $f_n^\varepsilon \in \text{dom } p_M = X$, by applying Lemma 4.1(ii) with $f = f_n^\varepsilon$, $g = g^o$, and $A = \text{span}_n G$, we obtain that there exists $h_n^\varepsilon \in M \cap \text{span}_n G$ satisfying

$$(4.1) \quad \|h_n^\varepsilon - g^o\| \leq (1 + c\|g^o\|) \|f_n^\varepsilon - g^o\| \leq (1 + c\|g^o\|)(\|g^o - \text{span}_n G\| + \varepsilon).$$

As $h_n^\varepsilon \in M \cap \text{span}_n G$, we have $\inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^o) \leq \Phi(h_n^\varepsilon) - \Phi(g^o)$. Estimating the right-hand side of this inequality in terms of the modulus of continuity α of Φ at g^o , we obtain $\inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^o) \leq \alpha(\|h_n^\varepsilon - g^o\|)$. Combining this estimate with inequality (4.1), we get

$$\inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^o) \leq \alpha((1 + c\|g^o\|)\|g^o - \text{span}_n G\| + \varepsilon).$$

By Theorem 3.1(i) we have

$$(4.2) \quad \inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^o) \leq \alpha \left((1 + c\|g^o\|) \sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}} + \varepsilon \right),$$

and by infimizing (4.2) over ε , we obtain

$$\inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^o) \leq \alpha \left((1 + c\|g^o\|) \sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}} \right),$$

which completes the proof of (i).

(ii) By the definition of ε_n -minimum point, $\Phi(g_n) - \Phi(g^o) \leq \inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^o) + \varepsilon_n$. So by item (i) we have

$$(4.3) \quad \Phi(g_n) - \Phi(g^o) \leq \alpha \left((1 + c\|g^o\|) \sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}} \right) + \varepsilon_n.$$

If $\lim_{n \rightarrow \infty} \varepsilon_n = 0$ and $\|g^o\|_G$ is finite, then the right-hand side of (4.2) converges to zero and so $\{g_n\}$ is Φ -minimizing.

(iii) By the definitions of ε_n -argmin and of the modulus of Tikhonov well-posedness of (M, Φ) at g^o and by item (i), we have

$$\begin{aligned} \xi(\|g_n - g^o\|) &= \inf_{g \in M \cap S_{\|g_n - g^o\|}(g^o)} \Phi(g) - \Phi(g^o) \leq \Phi(g_n) - \Phi(g^o) < \inf_{g \in M \cap \text{span}_n G} \Phi(g) \\ &\quad - \Phi(g^o) + \varepsilon_n \leq \alpha \left((1 + c\|g^o\|) \sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}} \right) + \varepsilon_n. \end{aligned}$$

(iv) By the definition of ε_n -argmin, Proposition 2.1(iii), and item (i), we have

$$\begin{aligned} \delta(\|g_n - g^o\|) &\leq \Phi(g_n) - \Phi(g^o) < \inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^o) + \varepsilon_n \\ &\leq \alpha \left((1 + c\|g^o\|) \sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}} \right) + \varepsilon_n. \quad \square \end{aligned}$$

Theorem 4.2 shows that for $\|g^o\|_G$ finite, the ε_n -near minimum points $\{g_n\}$ form a Φ -minimizing sequence over M and the speed of convergence of $\{\Phi(g_n)\}$ to $\Phi(g^o)$ is bounded from above by

$$\alpha \left((1 + c\|g^o\|) \sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}} \right) + \varepsilon_n.$$

When minimization is performed over the whole space X , the Lipschitz constant of the Minkowski functional $p_M = p_X$ is equal to zero; thus, Theorem 4.2 gives an upper bound

$$\alpha \left(\sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}} \right) + \varepsilon_n,$$

which depends on the modulus of continuity α of Φ , on the G -variation norm of g^o , and on the ambient space norm of g^o .

When the admissible set is a ball $B_r(\|\cdot\|)$, the Lipschitz constant is $1/r$ and we get an upper bound

$$\alpha \left(\left(1 + \frac{\|g^o\|}{r} \right) \sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}} \right) + \varepsilon_n.$$

As the estimates derived from Theorem 4.2 are not merely asymptotic, they can be applied to any degree n of variable-basis functions. Moreover, for admissible sets of d -variable functions the estimates hold for any number d of variables. Inspection of the upper bounds from Theorem 4.2 allows one to describe problems for which the rates of approximate optimization do not exhibit the curse of dimensionality (i.e., the degree n of variable-basis functions required for a satisfactory approximate optimization does not grow exponentially with the number d of variables of admissible solutions). A sufficient property of such problems is that the G -variation of their minimum point g^o does not depend exponentially on the number d of variables. Examples of classes of functions with small variations with respect to some bases used in neurocomputing were given in [12, 58] (see also the discussion in section 8).

The next theorem is an extension of Theorem 4.2 to $\mathcal{L}_p$ -spaces with $p \in (1, \infty)$. Its proof proceeds similarly as the proof of Theorem 4.2, but instead of the upper bound (i) from Theorem 3.1, it uses (ii) from the same theorem. The same remarks about the assumption $0 \in \text{int } M$ and the replacement of X with M as those preceding Theorem 4.2 apply here, as any closed subspace of a reflexive Banach space is a reflexive Banach space [22, Proposition III.17].

THEOREM 4.3. *Let $\Omega \subseteq \mathcal{R}^d$ be open, M and G be subsets of $(\mathcal{L}_p(\Omega), \|\cdot\|_p)$, $p \in (1, \infty)$, G be bounded, $s_G = \sup_{g \in G} \|g\|$, M be closed and convex, $0 \in \text{int } M$, $q = p/(p-1)$, $\bar{p} = \min(p, q)$, and $\bar{q} = \max(p, q)$. Let $\Phi : X \rightarrow (-\infty, +\infty]$ be a proper functional, $g^o \in \text{argmin}(M, \Phi)$, Φ be continuous at g^o with a modulus of continuity α , and $\{\varepsilon_n\}$ be a sequence of positive reals such that $g_n \in \text{argmin}_{\varepsilon_n}(M \cap \text{span}_n G, \Phi)$. Then p_M is Lipschitz on X , and if c is its Lipschitz constant, then the following estimates hold for every positive integer n :*

(i)

$$\inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^o) \leq \alpha \left((1 + c\|g^o\|) \frac{2^{1/\bar{p}} s_G \|g^o\|_G}{n^{1/\bar{q}}} \right);$$

(ii) *if $\|g^o\|_G < \infty$ and $\lim_{n \rightarrow \infty} \varepsilon_n = 0$, then $\{g_n\}$ is a Φ -minimizing sequence over M and*

$$\Phi(g_n) - \Phi(g^o) \leq \alpha \left((1 + c\|g^o\|) \frac{2^{1/\bar{p}} s_G \|g^o\|_G}{n^{1/\bar{q}}} \right) + \varepsilon_n;$$

(iii) *if ξ is the modulus of Tikhonov well-posedness of (M, Φ) at g^o , then*

$$\xi(\|g_n - g^o\|) \leq \alpha \left((1 + c\|g^o\|) \frac{2^{1/\bar{p}} s_G \|g^o\|_G}{n^{1/\bar{q}}} \right) + \varepsilon_n;$$

(iv) *if Φ is uniformly convex on M with a modulus of convexity δ , then*

$$\delta(\|g_n - g^o\|) \leq \alpha \left((1 + c\|g^o\|) \frac{2^{1/\bar{p}} s_G \|g^o\|_G}{n^{1/\bar{q}}} \right) + \varepsilon_n.$$

In the calculus of variations, the notion of a *direct method* [37, p. 192] is used to refer to a method for solving an optimization problem (M, Φ) by obtaining its minimum point g^o as a limit of a Φ -minimizing sequence $\{g_n\} \subseteq M$ satisfying $\lim_{n \rightarrow \infty} \Phi(g_n) = \Phi(g^o)$. Using this notion, we can rephrase our results as conditions on (M, Φ) under which the extended Ritz method has some of the properties of a direct method. By Theorems 4.2 and 4.3 for $\|g^o\|_G$ finite, any sequence $\{g_n\}$ of ε_n -minimum points of $(M \cap \text{span}_n G, \Phi)$ is Φ -minimizing and $\Phi(g^o) = \lim_{n \rightarrow \infty} \Phi(g_n) = \Phi(\lim_{n \rightarrow \infty} g_n)$. The convergence of $\{g_n\}$ to g^o is not always guaranteed (it depends on the behavior of the modulus of Tikhonov well-posedness of (M, Φ) at g^o). However, when applied to convex best approximation problems (see section 6) and to learning from data by kernel methods (see section 7), the extended Ritz method is a direct method.

5. Asymptotic estimates for convex problems. For convex problems with the functional to be minimized bounded in a neighborhood of a minimum point, under an additional assumption of density of $M \cap \text{span } G$ in M , the upper bounds from Theorem 4.2 can be simplified. But the simplified bounds are only asymptotic as their derivation takes advantage of the local behavior of the functional in a neighborhood

of a minimum point. For $f, g : \mathcal{N}_+ \rightarrow \mathcal{N}_+$ we write $g(n) \leq \mathcal{O}(f(n))$ when there exists $a > 0$ such that for all but finitely many $n \in \mathcal{N}_+$, $g(n) \leq a f(n)$.

THEOREM 5.1. *Let $(X, \|\cdot\|)$ be a Hilbert space, M and G be its subsets, G be bounded, $s_G = \sup_{g \in G} \|g\|$, M be closed and convex, $0 \in \text{int } M$, and $M \cap \text{span } G$ dense in M . Let $\Phi : X \rightarrow (-\infty, +\infty]$ be a proper convex functional, $g^o \in \text{argmin}(M, \Phi)$ be such that Φ is bounded in its neighborhood, $\{\varepsilon_n\}$ be a sequence of positive reals such that $\varepsilon_n \leq \mathcal{O}(1/\sqrt{n})$ and $g_n \in \text{argmin}_{\varepsilon_n}(M \cap \text{span}_n G, \Phi)$. Then the following estimates hold for every positive integer n :*

(i)

$$\inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^o) \leq \mathcal{O} \left(\sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}} \right);$$

(ii) if $\|g^o\|_G < \infty$, then $\{g_n\}$ is a Φ -minimizing sequence over M and

$$\Phi(g_n) - \Phi(g^o) \leq \mathcal{O} \left(\sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}} \right);$$

(iii) if ξ is the modulus of Tikhonov well-posedness of (M, Φ) at g^o , then

$$\xi(\|g_n - g^o\|) \leq \mathcal{O} \left(\sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}} \right);$$

(iv) if Φ is uniformly convex on M with a modulus of convexity δ , then

$$\delta(\|g_n - g^o\|) \leq \mathcal{O} \left(\sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}} \right).$$

Proof. (i) Let $\nu > 0$ be such that Φ is bounded on $B_\nu(g^o, \|\cdot\|)$. As $B_\nu(g^o, \|\cdot\|) \subseteq \text{dom } \Phi$, we have $g^o \in \text{int dom } \Phi$. Since Φ is a proper convex functional bounded on $B_\nu(g^o, \|\cdot\|)$, Φ is locally Lipschitz on $B_\nu(g^o, \|\cdot\|)$ [35, Corollary 2.4, p. 12]. Let $\eta \leq \nu$ be such that Φ is Lipschitz continuous with the constant c_1 on $B_\eta(g^o, \|\cdot\|)$.

As $M \cap \text{span } G$ is dense in M , $\lim_{n \rightarrow \infty} \|g^o - \text{span}_n G\| = 0$, and so there exist $\varepsilon_0 > 0$ and $n_0 \in \mathcal{N}_+$ such that $\|g^o - \text{span}_{n_0} G\| + \varepsilon_0 \leq \frac{\eta}{1+c\|g^o\|}$. For every $n \geq n_0$ and $\varepsilon \leq \varepsilon_0$, choose $f_n^\varepsilon \in \text{span}_n G$ such that $\|g^o - f_n^\varepsilon\| \leq \|g^o - \text{span}_n G\| + \varepsilon$.

As M is closed, convex, $0 \in \text{int } M$, and $\text{dom } p_M = X$, we can apply Lemma 4.1(ii) with $f = f_n^\varepsilon$, $g = g^o$, and $A = \text{span}_n G$ to obtain $h_n^\varepsilon \in M \cap \text{span}_n G$ satisfying

$$(5.1) \quad \|h_n^\varepsilon - g^o\| \leq (1 + c\|g^o\|) \|f_n^\varepsilon - g^o\| < \eta.$$

So h_n^ε is in the ball $B_\eta(g^o, \|\cdot\|)$, on which Φ is Lipschitz continuous with the constant c_1 . So we have

$$(5.2) \quad \inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^o) \leq \Phi(h_n^\varepsilon) - \Phi(g^o) \leq c_1 \|h_n^\varepsilon - g^o\|.$$

From (5.1) and (5.2) we obtain

$$(5.3) \quad \inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^o) \leq C \|f_n^\varepsilon - g^o\|,$$

where $C = c_1 (1 + c \|g^o\|)$. By Theorem 3.1(i) we get

$$(5.4) \quad \|g^o - f_n^\varepsilon\| \leq \|g^o - \text{span}_n G\| + \varepsilon \leq \sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}} + \varepsilon.$$

Infimizing over ε , from (5.3) and (5.4) we obtain for all $n \geq n_0$

$$\inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^o) \leq C \sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}}.$$

(ii) As $g_n \in \text{argmin}_{\varepsilon_n} (M \cap \text{span}_n G)$, we have $\Phi(g_n) < \inf_{g \in M \cap \text{span}_n G} \Phi(g) + \varepsilon_n$. Combining this inequality with the one from item (i) and with $\varepsilon_n \leq \mathcal{O}(1/\sqrt{n})$, we obtain

$$\Phi(g_n) - \Phi(g^o) \leq \mathcal{O} \left(\sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}} \right).$$

(iii) By the definitions of ε_n -argmin and of the modulus of Tikhonov well-posedness of (M, Φ) at g^o and by item (i), for every $n \geq n_0$ we have

$$\begin{aligned} \xi(\|g_n - g^o\|) &= \inf_{g \in M \cap S_{\|g_n - g^o\|}(g^o)} \Phi(g) - \Phi(g^o) \\ &\leq \Phi(g_n) - \Phi(g^o) < \inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^o) + \varepsilon_n \\ &\leq C \sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}} + \varepsilon_n. \end{aligned}$$

As $\varepsilon_n \leq \mathcal{O}(1/\sqrt{n})$, we obtain

$$\xi(\|g_n - g^o\|) \leq \mathcal{O} \left(\sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}} \right).$$

(iv) By the definition of ε_n -argmin, item (i), and Proposition 2.1(iii), for all $n \geq n_0$ we get

$$\begin{aligned} \delta(\|g_n - g^o\|) &\leq \Phi(g_n) - \Phi(g^o) < \inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^o) + \varepsilon_n \\ &\leq \mathcal{O} \left(\sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}} \right) + \varepsilon_n. \end{aligned}$$

As $\varepsilon_n \leq \mathcal{O}(1/\sqrt{n})$, we obtain

$$\delta(\|g_n - g^o\|) \leq \mathcal{O} \left(\sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}} \right). \quad \square$$

Inspection of the proof of Theorem 5.1 shows that there exists $n_0 \in \mathcal{N}_+$ such that the expression

$$\mathcal{O} \left(\sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}} \right)$$

can be written for $n \geq n_o$ as

$$C \sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}},$$

where $C = c_1(1 + c\|g^o\|)$, c is the Lipschitz constant of p_M , and c_1 is the Lipschitz constant of Φ in a neighborhood of g^o . The proof also shows that for any sequences $\{\varepsilon_n\}$ and $\{g_n\}$ such that $g_n \in \operatorname{argmin}_{\varepsilon_n}(M, \Phi)$, the statements of Theorem 5.1(ii), (iii), and (iv) hold with the bounds replaced with

$$\mathcal{O}\left(\sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}}\right) + \varepsilon_n.$$

Applying Theorem 3.1(ii) instead of Theorem 3.1(i) and following steps analogous to those in the proof of Theorem 5.1, one can obtain for $\mathcal{L}_p$ -spaces estimates similar to those stated in Theorem 5.1 for Hilbert spaces (the condition $\varepsilon_n \leq \mathcal{O}(1/\sqrt{n})$ has to be replaced with $\varepsilon_n \leq \mathcal{O}(n^{1/\bar{q}})$, where $q = p/(p-1)$ and $\bar{q} = \max(p, q)$).

6. Application to convex best approximation problems. The simplest example illustrating the estimates derived in section 4 is an application of the extended Ritz method to convex best approximation problems.

For any $f \in X$, let e_f denote the functional defined as the distance from f ; i.e., $e_f(g) = \|g - f\|$ for any $g \in X$.

When M is a closed convex subset of X , (M, e_f) is called a *convex best approximation problem* [34, p. 40]. We recall that M is a *Chebyshev set* if each $f \in X$ has a unique best approximation in M (i.e., there exists a unique $g^o \in M$ such that $\|f - g^o\| = \|f - M\|$) [29, p. 21].

In [43], the classical Ritz method was used to solve approximately the problem (M, e_f) with M a closed separable subspace of X , but rates of convergence were not estimated. For X finite-dimensional, other methods of approximate solution of the best approximation problem have also been studied, and, for some of them, estimates of rates of convergence have been derived (e.g., [44, pp. 118–122]). For X infinite-dimensional, a method of approximation of best approximation for which estimates of rates of convergence are available is Dijkstra's algorithm [29, p. 207] applied to a special class of admissible sets M of the form $\bigcup_{i=1}^r M_i$, where M_i are closed affine sets and r is finite [29, p. 201].

Here, taking advantage of the upper bounds from section 4, we estimate rates of convergence of approximate solutions of the problem (M, e_f) , where M is closed and convex, that are obtained by the extended Ritz method. Applying Theorem 4.2 to the best approximation problems (M, e_f) and (M, e_f^2) , we derive the following upper bounds.

THEOREM 6.1. *Let $(X, \|\cdot\|)$ be a Hilbert space, M and G be its subsets, G be bounded, $s_G = \sup_{g \in G} \|g\|$, M be closed and convex, $0 \in \operatorname{int} M$, and $f \in X$. Then p_M is Lipschitz on X , and if c is its Lipschitz constant, then there exists a unique minimum point g^o of (M, e_f) , and the following estimates hold for every positive integer n :*

(i)

$$\begin{aligned} \inf_{g \in M \cap \operatorname{span}_n G} e_f(g) - e_f(g^o) &= \|f - M \cap \operatorname{span}_n G\| - \|f - M\| \\ &\leq (1 + c\|g^o\|) \sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}}; \end{aligned}$$

- (ii) if M is bounded, $\{\varepsilon_n\}$ is a sequence of positive reals, and for every n , $g_n \in \operatorname{argmin}_{\varepsilon_n} (M \cap \operatorname{span}_n G, e_f^2)$, then

$$\|g_n - g^o\|^2 \leq 2 \operatorname{diam} M \left((1 + c\|g^o\|) \sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}} \right)^2 + \varepsilon_n.$$

Proof. As every closed convex subset of a Hilbert space is Chebyshev [29, p. 35], the problem (M, e_f) has a unique minimum point.

By the triangle inequality, for every $h, g \in X$ we have $|e_f(h) - e_f(g)| \leq \|h - g\|$. So e_f is uniformly continuous on X and its modulus of continuity is $\alpha(t) = t$. Hence, applying Theorem 4.2(i) to (M, e_f) we obtain (i).

To derive (ii), we apply Theorem 4.2(iv) to the functional e_f^2 . As $\|f - g^o\|^2 = \inf_{g \in M} \|f - g\|^2$, g^o is a minimum point of (M, e_f^2) and by Proposition 2.1(iv), the functional $\|\cdot\|^2$ is uniformly convex with a modulus of convexity $\delta(t) = t^2$.

By the triangle inequality, for every $h, g \in X$ we have $|e_f^2(h) - e_f^2(g)| = |(\|f - h\| - \|f - g\|)(\|f - h\| + \|f - g\|)| \leq 2 \operatorname{diam} M \|h - g\|$, and so $\alpha(t) = 2t \operatorname{diam} M$ is an upper bound on the modulus of continuity of e_f^2 . Thus, by Theorem 4.2(iv) we get

$$\|g_n - g^o\|^2 \leq 2 \operatorname{diam} M \left((1 + c\|g^o\|) \sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}} \right)^2 + \varepsilon_n. \quad \square$$

Combining Theorem 4.3 with estimates of moduli of convexity of $\mathcal{L}_p$ -spaces, $p \in (1, \infty)$, we obtain the following upper bounds.

THEOREM 6.2. Let $\Omega \subseteq \mathcal{R}^d$ be open, M and G be subsets of $(\mathcal{L}_p(\Omega), \|\cdot\|_p)$, $p \in (1, \infty)$, G be bounded, $s_G = \sup_{g \in G} \|g\|$, M be closed and convex, $0 \in \operatorname{int} M$, $f \in X$, $q = p/(p-1)$, $\bar{p} = \min(p, q)$, $\bar{q} = \max(p, q)$, and α_p, α_q be moduli of continuity of e_f^p and e_f^q , respectively, at f . Then p_M is Lipschitz on X , and if c is its Lipschitz constant, then there exists a unique minimum point g^o of (M, e_f) , and the following estimates hold for every positive integer n :

- (i) $\inf_{g \in M \cap \operatorname{span}_n G} e_f(g) - e_f(g^o) \leq (1 + c\|g^o\|) \frac{2^{1/\bar{p}} s_G \|g^o\|_G}{n^{1/\bar{q}}}$;
(ii) if $\{\varepsilon_n\}$ is a sequence of positive reals, $p \in (1, 2]$, and $g_n \in \operatorname{argmin}_{\varepsilon_n} (M \cap \operatorname{span}_n G, e_f^q)$, then

$$\|g_n - g^o\|^q \leq 2^{q-2} \alpha_q \left((1 + c\|g^o\|) \frac{2^{1/\bar{p}} s_G \|g^o\|_G}{n^{1/\bar{q}}} \right) + \varepsilon_n;$$

- (iii) if $\{\varepsilon_n\}$ is a sequence of positive reals, $p \geq 2$, and $g_n \in \operatorname{argmin}_{\varepsilon_n} (M \cap \operatorname{span}_n G, e_f^p)$, then

$$\|g_n - g^o\|^p \leq 2^{p-2} \alpha_p \left((1 + c\|g^o\|) \frac{2^{1/\bar{p}} s_G \|g^o\|_G}{n^{1/\bar{q}}} \right) + \varepsilon_n.$$

Proof. Since for all $p \in (1, \infty)$ the space $(\mathcal{L}_p(\Omega), \|\cdot\|_p)$ is uniformly convex [2, 2.29] and every convex best approximation problem in a uniformly convex space is Tikhonov well-posed [34, p. 40], there exists a unique $g^o \in M$ such that $\|f - g^o\|_p = \|f - M\|_p$. Hence, the problem (M, e_f) has a unique minimum point.

(i) By the triangle inequality, for every $h, g \in X$ we have $|e_f(h) - e_f(g)| \leq \|h - g\|_p$. So e_f is uniformly continuous on X and its modulus of continuity is $\alpha(t) = t$. Thus, by Theorem 4.3(i) we obtain (i).

(ii) When $p \in (1, 2]$, the estimate follows from Theorem 4.3(iv) applied to the functional e_f^q with $q = p/(p-1)$, combined with Proposition A.3(i).

(iii) When $p \geq 2$, the estimate follows from Theorem 4.3(iv) applied to the functional e_f^p , combined with Proposition A.3(ii). $\square$

So Theorems 6.1(i) and 6.2(i) extend Theorem 3.1 on approximation by $\text{span}_n G$ to approximation by $M \cap \text{span}_n G$, where M is closed, convex, with zero in its interior, in particular when $M = B_r(\|\cdot\|)$ for some $r > 0$.

COROLLARY 6.3. *Let $(X, \|\cdot\|)$ be a normed linear space, M and G be its subsets, G be bounded, $s_G = \sup_{g \in G} \|g\|$, M be closed and convex, $0 \in \text{int } M$, $f \in M$, and $g^o = \text{argmin}(M, e_f)$. Then p_M is Lipschitz on X , and if c is its Lipschitz constant, then the following estimates hold for every positive integer n :*

(i) *if $(X, \|\cdot\|)$ is a Hilbert space, then*

$$\|f - M \cap \text{span}_n G\| \leq (1 + c\|g^o\|) \sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}} + \|f - g^o\|;$$

(ii) *if $(X, \|\cdot\|) = (\mathcal{L}_p(\Omega), \|\cdot\|_p)$, where $p \in (1, \infty)$, $\Omega \subseteq \mathcal{R}^d$ is open, $q = p/(p-1)$, $\bar{p} = \min(p, q)$, and $\bar{q} = \max(p, q)$, then*

$$\|f - M \cap \text{span}_n G\| \leq (1 + c\|g^o\|) \frac{2^{1/\bar{p}} s_G \|g^o\|_G}{n^{1/\bar{q}}} + \|f - g^o\|.$$

Note that for $M = X$, Corollary 6.3 gives the same estimate as Theorem 3.1, since in such a case the Lipschitz constant of p_M is equal to 0 and $g^o = f$.

7. Application to learning from data. Learning from a sample $\{(x_i, y_i) \in \mathcal{R}^d \times \mathcal{R}, i = 1, \dots, m\}$ of *empirical data* can be modeled as minimization of the *empirical error functional* (also called the *empirical risk functional*), defined as

$$\mathcal{E}(f) = \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2.$$

However, the empirical error does not take into account any global properties of the input/output mapping from which the sample was chosen. Such properties can be expressed through *regularization*, which replaces the functional $\mathcal{E}$ with $\mathcal{E}_{\gamma, \Psi} = \mathcal{E} + \gamma \Psi$, where Ψ is a suitable functional called *stabilizer* and γ is a positive real number called *regularization parameter*. The stabilizer penalizes the solutions with some undesired properties such as high-frequency oscillations, while the regularization parameter allows us to obtain a tradeoff between fitting to empirical data and fitting to properties of solutions expressed by the stabilizer.

An important class of stabilizers are squares of norms on reproducing kernel Hilbert spaces. A *reproducing kernel Hilbert space* (RKHS) $(\mathcal{H}_K(\Omega), \|\cdot\|_K)$ is a Hilbert space of functions defined on a set Ω such that for every $x \in \Omega$, the evaluation functional $\mathcal{F}_x$, defined for any $f \in \mathcal{H}_K(\Omega)$ as $\mathcal{F}_x(f) = f(x)$, is bounded. For any RKHS there exists a unique symmetric, positive semidefinite mapping $K : \Omega \times \Omega \rightarrow \mathcal{R}$, called *kernel*, such that for any $f \in \mathcal{H}_K(\Omega)$ and any $x \in \Omega$, $\mathcal{F}_x(f) = \langle f, K(x, \cdot) \rangle_K$ [7] (recall that a mapping $K : \Omega \times \Omega \rightarrow \mathcal{R}$ is *positive semidefinite* on Ω if for all positive integers m , all $(a_1, \dots, a_m) \in \mathcal{R}^m$, and all $(x_1, \dots, x_m) \in \Omega^m$, $\sum_{i,j=1}^m a_i a_j K(x_i, x_j) \geq 0$).

By the Cauchy–Schwarz inequality, for every $f \in \mathcal{H}_K(\Omega)$ and every $x \in \Omega$ we have $|f(x)| = |\langle f, K(x, \cdot) \rangle_K| \leq \|f\|_K \sqrt{K(x, x)} \leq c_K \|f\|_K$, where $c_K = \sup_{x \in \Omega} \sqrt{K(x, x)}$.

Thus for every set Ω and every kernel K on $\Omega \times \Omega$ we have

$$(7.1) \quad \sup_{x \in \Omega} |f(x)| \leq c_K \|f\|_K.$$

With $\|\cdot\|_K^2$ as a stabilizer, the regularized functional obtained from $\mathcal{E}$ is of the form

$$(7.2) \quad \mathcal{E}_{\gamma,K}(f) = \frac{1}{m} \sum_{i=1}^m |f(x_i) - y_i|^2 + \gamma \|f\|_K^2.$$

The representer theorem (see, e.g., [25, p. 42], [69, pp. 538–539]) states that the problem $(\mathcal{H}_K(\Omega), \mathcal{E}_{\gamma,K})$ has a unique minimum point g^o of the form

$$(7.3) \quad g^o(x) = \sum_{i=1}^m a_i K(x, x_i).$$

It even gives a formula for computing the parameters $a = (a_1, \dots, a_m)$ as the unique solution of the well-posed system of linear equations

$$(7.4) \quad (\mathcal{K}[x] + \gamma m \mathcal{I})a = y,$$

where $y = (y_1, \dots, y_m)$, $\mathcal{K}[x]$ is the $m \times m$ matrix defined as $\mathcal{K}[x]_{ij} = K(x_i, x_j)$, and $\mathcal{I}$ is the identity matrix [25, 69].

Thus, to compute the coefficients $(a_1, \dots, a_m)$ of the linear combination it is necessary to solve the inverse problem (7.4), which may be ill-conditioned. To guarantee for a given m a small condition number [63, p. 33] of the matrix $\mathcal{K}[x] + \gamma m \mathcal{I}$, the regularization parameter γ must be “large” [57]. On the other hand, a “large” γ does not allow good interpolation of the empirical data. This limits the applicability of algorithms for computing the solution of the problem $(\mathcal{H}_K(\Omega), \mathcal{E}_{\gamma,K})$ given by the representer theorem.

It has been argued in [38, p. 219] that the “regularization principles lead to approximation schemes that are equivalent to networks with one layer of hidden units.” Indeed, the unique minimum point of the problem $(\mathcal{H}_K(\Omega), \mathcal{E}_{\gamma,K})$ is in the set $\text{span}_m G_K$, where $G_K = \{K(x, \cdot) : x \in \Omega\}$. Functions from this set can be computed by neural networks with m hidden units. In particular, for the Gaussian kernel, they can be computed by radial-basis-function networks with Gaussian units. A drawback of this elegant result is that the number of network hidden units needed to compute the function minimizing $\mathcal{E}_{\gamma,K}$ is equal to the size of the sample of input/output data. For large data sets, such networks might not be implementable. Moreover, in typical applications of neural networks, a number of hidden units much smaller than the number of data is chosen before learning.

Using Theorem 4.2, we derive an approximate version of the representer theorem. It estimates how quickly approximate solutions achievable by networks with n hidden units converge to the minimum point described by the representer theorem. We first state in the next proposition basic properties of the functional $\mathcal{E}_{\gamma,K}$.

PROPOSITION 7.1. *Let Ω be a nonempty set, $K : \Omega \times \Omega \rightarrow \mathcal{R}$ be a kernel, $c_K = \sup_{x \in \Omega} \sqrt{K(x, x)}$, $\gamma > 0$, m be a positive integer, $\{(x_1, y_1), \dots, (x_m, y_m)\} \subset (\Omega \times \mathcal{R})^m$, $\mathcal{E}_{\gamma,K}(f) = \frac{1}{m} \sum_{i=1}^m |f(x_i) - y_i|^2 + \gamma \|f\|_K^2$, and $y_{\max} = \max\{|y_i| : i = 1, \dots, m\}$. Then*

- (i) *the functional $\mathcal{E}_{\gamma,K}$ is uniformly convex on $\mathcal{H}_K(\Omega)$ with a modulus of convexity $\delta(t) = \gamma t^2$;*

- (ii) at every $f \in \mathcal{H}_K(\Omega)$, $\mathcal{E}_{\gamma,K}$ is continuous with a modulus of continuity bounded from above by $\alpha(t) = a_2 t^2 + a_1 t$, where $a_1 = 2(\|f\|_K c_K^2 + y_{\max} c_K + \gamma \|f\|_K)$ and $a_2 = c_K^2 + \gamma$;
- (iii) when $M \subset \mathcal{H}_K(\Omega)$ is closed, convex, and bounded, or when $M = \mathcal{H}_K(\Omega)$, the problem $(M, \mathcal{E}_{\gamma,K})$ has a unique minimum point g^o ;
- (iv) for any minimum point g^o of $(M, \mathcal{E}_{\gamma,K})$ and any $f \in M$, $\|f - g^o\|_K^2 \leq \frac{|\mathcal{E}_{\gamma,K}(f) - \mathcal{E}_{\gamma,K}(g^o)|}{\gamma}$.

Proof. (i) It is easy to show that $\mathcal{E}$ is convex, so (i) follows from Proposition 2.1(i) and (iv).

(ii) Let $f \in \mathcal{H}_K(\Omega)$, $t > 0$, and $g \in \mathcal{H}_K(\Omega)$ be such that $\|f - g\|_K < t$. Then

$$\begin{aligned} |\mathcal{E}_{\gamma,K}(f) - \mathcal{E}_{\gamma,K}(g)| &= \left| \frac{1}{m} \sum_{i=1}^m ((f(x_i) - y_i)^2 - (g(x_i) - y_i)^2) + \gamma (\|f\|_K^2 - \|g\|_K^2) \right| \\ &\leq \left| \frac{1}{m} \sum_{i=1}^m (f(x_i) - g(x_i)) (f(x_i) + g(x_i) - 2y_i) \right| + \gamma \left| \|f\|_K - \|g\|_K \right| (\|f\|_K + \|g\|_K) \\ &\leq \sup_{x \in \Omega} |f(x) - g(x)| \left| \sup_{x \in \Omega} |f(x) + g(x)| + 2y_{\max} \right| + \gamma \|f - g\|_K (\|f\|_K + \|g\|_K). \end{aligned}$$

Thus, by (7.1) we have $|\mathcal{E}_{\gamma,K}(f) - \mathcal{E}_{\gamma,K}(g)| \leq c_K \|f - g\|_K |c_K \|f + g\|_K + 2y_{\max}| + \gamma \|f - g\|_K (\|f\|_K + \|g\|_K)$. As $\|g\|_K < \|f\|_K + t$, we obtain

$$\begin{aligned} |\mathcal{E}_{\gamma,K}(f) - \mathcal{E}_{\gamma,K}(g)| &< t c_K \left| 2\|f\|_K c_K + t c_K + 2y_{\max} \right| + \gamma t (2\|f\|_K + t) \\ &\leq t c_K \left(2\|f\|_K c_K + t c_K + 2y_{\max} \right) + \gamma t (2\|f\|_K + t) \\ &= t^2 (c_K^2 + \gamma) + 2t \left(\|f\|_K c_K^2 + y_{\max} c_K + \gamma \|f\|_K \right). \end{aligned}$$

Hence, $\|f - g\|_K < t$ implies $|\mathcal{E}_{\gamma,K}(f) - \mathcal{E}_{\gamma,K}(g)| < \alpha(t) = a_2 t^2 + a_1 t$, where $a_2 = c_K^2 + \gamma$ and $a_1 = 2(\|f\|_K c_K^2 + y_{\max} c_K + \gamma \|f\|_K)$.

(iii) When $M \subset \mathcal{H}_K(\Omega)$ is closed, convex, and bounded, the existence of a unique minimum point of $(M, \mathcal{E}_{\gamma,K})$ follows from (i) and [70, Theorem 5]. When $M = \mathcal{H}_K(\Omega)$, it follows from the representer theorem [69, pp. 538–539].

(iv) follows from (i) and Proposition 2.1 (iii). $\square$

So the modulus of continuity of $\mathcal{E}_{\gamma,K}$ at any $f \in \mathcal{H}_K(\Omega)$ is bounded from above by the quadratic function $a_2 t^2 + a_1 t$. Note that a_2 depends on m , c_K , and γ , while a_1 depends, in addition to these values, also on $\|f\|_K$ and $y_{\max}$.

Applying Proposition 7.1 and Theorem 4.2 to the problem $(\mathcal{H}(\Omega), \mathcal{E}_{\gamma,K})$, we obtain the following estimates, which hold for every positive integer n (but are only useful for $n < m$, as the minimum point g^o is in $\text{span}_m G_K$).

THEOREM 7.2. *Let Ω be a nonempty set, $K : \Omega \times \Omega \rightarrow \mathcal{R}$ be a kernel, $c_K = \sup_{x \in \Omega} \sqrt{K(x, x)}$, $(\mathcal{H}_K(\Omega), \|\cdot\|_K)$ be the RKHS defined by K , $G_K = \{K(x, \cdot) : x \in \Omega\}$, $(x_1, \dots, x_m) \in \Omega^m$, $(y_1, \dots, y_m) \in \mathcal{R}^m$, $y_{\max} = \max\{|y_i| : i = 1, \dots, m\}$, $\gamma > 0$, $\mathcal{E}_{\gamma,K}(f) = \frac{1}{m} \sum_{i=1}^m |f(x_i) - y_i|^2 + \gamma \|f\|_K^2$, $g^o(x) = \sum_{i=1}^m w_i K(x, x_i)$ be the unique minimum point of the problem $(\mathcal{H}_K(\Omega), \mathcal{E}_{\gamma,K})$ given by the representer theorem, and $\{\varepsilon_n\}$ be a sequence of positive reals such that $g_n \in \text{argmin}_{\varepsilon_n} (\text{span}_n G_K, \mathcal{E}_{\gamma,K})$. Then the following estimates hold for every positive integer n :*

(i)

$$\inf_{g \in \text{span}_n G_K} \mathcal{E}_{\gamma,K}(g) - \mathcal{E}_{\gamma,K}(g^o) \leq \alpha \left(\sqrt{\frac{(c_K \|g^o\|_{G_K})^2 - \|g^o\|_K^2}{n}} \right);$$

(ii) *if $\|g^o\|_G < \infty$ and $\lim_{n \rightarrow \infty} \varepsilon_n = 0$, then $\{g_n\}$ is an $\mathcal{E}_{\gamma,K}$ -minimizing sequence over $\mathcal{H}_K(\Omega)$ and*

$$\mathcal{E}_{\gamma,K}(g_n) - \mathcal{E}_{\gamma,K}(g^o) \leq \alpha \left(\sqrt{\frac{(c_K \|g^o\|_{G_K})^2 - \|g^o\|_K^2}{n}} \right) + \varepsilon_n;$$

(iii)

$$\gamma \|g_n - g^o\|_K^2 \leq \alpha \left(\sqrt{\frac{(c_K \|g^o\|_{G_K})^2 - \|g^o\|_K^2}{n}} \right) + \varepsilon_n;$$

(iv)

$$\gamma \sup_{x \in \Omega} |g_n(x) - g^o(x)|^2 \leq c_K \left(\alpha \left(\sqrt{\frac{(c_K \|g^o\|_{G_K})^2 - \|g^o\|_K^2}{n}} \right) + \varepsilon_n \right),$$

where $\alpha(t) = a_2 t^2 + a_1 t$, $a_1 = 2(\|g^o\|_K c_K^2 + y_{\max} c_K + \gamma \|g^o\|_K)$ and $a_2 = c_K^2 + \gamma$.

Proof. Statements (i) and (ii) follow from Theorem 4.2 with $M = \mathcal{H}_K(\Omega)$, $c = 0$ (as in this case p_M is the constant functional equal to zero), $\Phi(f) = \mathcal{E}_{\gamma,K}(f)$, $G = G_K$, and $s_G = \sup_{x \in \Omega} \|K(x, \cdot)\|_K = \sup_{x \in \Omega} \sqrt{\langle K(x, \cdot), K(x, \cdot) \rangle_K} = \sup_{x \in \Omega} \sqrt{K(x, x)} = c_K$. By Proposition 7.1(i) and (ii), $\mathcal{E}_{\gamma,K}$ is uniformly convex with a modulus of convexity $\delta(t) = \gamma t^2$ and is continuous at g^o with a modulus of continuity $\alpha(t) = a_2 t^2 + a_1 t$, so (iii) follows from Theorem 4.2(iv). Finally, item (iv) follows from (iii) and inequality (7.1). $\square$

For this application of Theorem 4.2, an explicit formula (7.3) describing the minimum point g^o is given by the representer theorem. Taking advantage of this formula, estimates of $\|g^o\|_{G_K}$ and $\|g^o\|$ in terms of the properties of the sample $\{(x_i, y_i), i = 1, \dots, m\}$ of data, the kernel K , and the regularization parameter γ were derived in [57].

8. Discussion. We have derived upper bounds on rates of approximate optimization by the extended Ritz method for problems (M, Φ) having a minimum point, where Φ is continuous and M is closed, convex, and containing 0 in its interior. The bounds can be applied to a variety of problems with sets of admissible solutions equal to the ambient space, to its subspaces (restating the problems for the subspaces), and to balls of some radii in the ambient norm. Such admissible sets occur, for example, in Tikhonov's and Ivanov's regularizations.

The critical term in the bounds is of the form $n^{-1/2}$ multiplied by the variation norm of the minimum point. To take advantage of these bounds, one needs some insights into the behavior of the variation norm tailored to the basis used for the extended Ritz method. Various methods based on integral representations (such as the Fourier transform [12, 21, 45, 58] and the Radon transform [49, 54]) were proposed to estimate the variation norm. For a survey of properties of G -variation see [53].

The role of variation norms in variable-basis approximation can be clarified by a comparison with the role played by Sobolev norms in linear approximation. Rates of linear approximation of order $\mathcal{O}(n^{-1/2})$ for functions of d variables can be achieved when the approximation is restricted to functions from balls in Sobolev norms of degree $s = d/2$ [67, pp. 232–233]. Similarly, in variable-basis approximation rates bounded from above by $rn^{-1/2}$ can be obtained by restricting the approximation to balls of radii r in G -variation. Note that the “ $\mathcal{O}$ ” notation in estimates of rates of linear approximation of functions from balls in Sobolev norms hides “constants” that may depend on d [67, pp. 232–241]. Moreover with d increasing, balls in Sobolev spaces of degree $s = d/2$ “shrink,” since some d -variable functions in the unit balls in the Sobolev norms $\|\cdot\|_{d/2,p}$ with “large” $((d+1)/2)$ th derivatives cannot be extended to $(d+1)$ -variable functions from the unit balls in the Sobolev norms $\|\cdot\|_{(d+1)/2,p}$. In contrast to the linear approximation case, in variable-basis approximation with certain types of bases (such as those generated by neural-network computational units [53]) there exist families of sets of d -variable functions that can be approximated with rates $n^{-1/2}$ and do not shrink as d increases.

Appendix A. For the reader’s convenience, here we state and prove some results from nonlinear approximation theory, which are used in section 4.

The following theorem states the Maurey–Jones–Barron estimate in a slightly reformulated way. The proof, which is a mild simplification of the argument from [12, Lemma 1, p. 934], is from [53]. By $\text{conv}_n G$ is denoted the set of all convex combinations of at most n elements of the set G , i.e.,

$$\text{conv}_n G = \left\{ \sum_{i=1}^n w_i g_i : w_i \in [0, 1], \sum_{i=1}^n w_i = 1, g_i \in G \right\}.$$

THEOREM A.1. *Let $(X, \|\cdot\|)$ be a Hilbert space, G be its bounded subset, and $s_G = \sup_{g \in G} \|g\|$. Then for every $f \in \text{cl conv } G$ and for every positive integer n ,*

$$\|f - \text{conv}_n G\| \leq \sqrt{\frac{s_G^2 - \|f\|^2}{n}}.$$

Proof. Since the distance from $\text{conv}_n G$ is continuous on $(X, \|\cdot\|)$ [72, p. 391], it is sufficient to verify the statement for $f \in \text{conv } G$. Let $f = \sum_{j=1}^m a_j h_j$ be a representation of f as a convex combination of elements of G . Set $c = s_G^2 - \|f\|^2$. We show by induction that there exists a sequence $\{g_i\}$ of elements of G such that the barycenters $f_n = \sum_{i=1}^n \frac{g_i}{n}$ satisfy $e_n^2 = \|f - f_n\|^2 \leq \frac{c}{n}$.

First we check that there exists $g_1 \in G$ such that $f_1 = g_1$ satisfies $e_1^2 = \|f - f_1\|^2 \leq c$. As $\sum_{j=1}^m a_j \|f - h_j\|^2 = \|f\|^2 - 2 \langle f, \sum_{j=1}^m a_j h_j \rangle + \sum_{j=1}^m a_j \|h_j\|^2 \leq s_G^2 - \|f\|^2 = c$, there must exist at least one $j \in \{1, \dots, m\}$ for which $\|f - h_j\|^2 \leq c$.

Setting $g_1 = h_j$ and assuming that we already have $g_1, \dots, g_n$, we derive the estimate by induction. We express e_{n+1}^2 in terms of e_n^2 as $e_{n+1}^2 = \|f - f_{n+1}\|^2 = \|\frac{n}{n+1}(f - f_n) + \frac{1}{n+1}(f - g_{n+1})\|^2 = \frac{n^2}{(n+1)^2} e_n^2 + \frac{2n}{(n+1)^2} \langle f - f_n, f - g_{n+1} \rangle + \frac{1}{(n+1)^2} \|f - g_{n+1}\|^2$.

Analogously to the first step, we consider a convex combination of the last two terms from the formula expressing e_{n+1}^2 in terms of e_n^2 . Thus we obtain

$$\begin{aligned}
& \sum_{j=1}^m a_j \left(\frac{2n}{(n+1)^2} \langle f - f_n, f - h_j \rangle + \frac{1}{(n+1)^2} \|f - h_j\|^2 \right) \\
&= \frac{2n}{(n+1)^2} \left\langle f - f_n, f - \sum_{j=1}^m a_j h_j \right\rangle \\
&+ \frac{1}{(n+1)^2} \left(\|f\|^2 - 2 \left\langle f, \sum_{j=1}^m a_j h_j \right\rangle + \sum_{j=1}^m a_j \|h_j\|^2 \right) \\
&= \frac{1}{(n+1)^2} \left(\sum_{j=1}^m a_j \|h_j\|^2 - \|f\|^2 \right) \leq \frac{1}{(n+1)^2} (s_G^2 - \|f\|^2) = \frac{c}{(n+1)^2}.
\end{aligned}$$

So there must exist $j \in \{1, \dots, m\}$ such that $\frac{2n}{(n+1)^2} \langle f - f_n, f - h_j \rangle + \frac{1}{(n+1)^2} \|f - h_j\|^2 \leq \frac{c}{(n+1)^2}$.

Setting $g_{n+1} = h_j$, we get $e_{n+1}^2 \leq \frac{n^2}{(n+1)^2} e_n^2 + \frac{c}{(n+1)^2}$. It can be easily verified by induction that this recursive formula together with $e_1^2 \leq c$ gives $e_n^2 \leq \frac{c}{n}$. $\square$

In [28], the Maurey–Jones–Barron estimate was extended to $\mathcal{L}_p$ -spaces, $p \in (1, \infty)$, with a more sophisticated argument, replacing inner products with peak functionals and taking advantage of Clarkson's inequalities stated in the following proposition from [42, pp. 225, 227].

PROPOSITION A.2 (Clarkson's inequalities). *Let $\Omega \subseteq \mathcal{R}^d$, $p \in (1, \infty)$, $f, g \in (\mathcal{L}_p(\Omega), \|\cdot\|_p)$, and $q = p/(p-1)$. Then for $p \in (1, 2]$*

$$(A.1) \quad \left\| \frac{f+g}{2} \right\|_p^q + \left\| \frac{f-g}{2} \right\|_p^q \leq \left(\frac{1}{2} \|f\|_p^p + \frac{1}{2} \|g\|_p^p \right)^{q-1},$$

$$(A.2) \quad \left\| \frac{f+g}{2} \right\|_p^p + \left\| \frac{f-g}{2} \right\|_p^p \geq \frac{1}{2} \|f\|_p^p + \frac{1}{2} \|g\|_p^p,$$

and for $p \geq 2$

$$(A.3) \quad \left\| \frac{f+g}{2} \right\|_p^p + \left\| \frac{f-g}{2} \right\|_p^p \leq \frac{1}{2} \|f\|_p^p + \frac{1}{2} \|g\|_p^p,$$

$$(A.4) \quad \left\| \frac{f+g}{2} \right\|_p^q + \left\| \frac{f-g}{2} \right\|_p^q \geq \left(\frac{1}{2} \|f\|_p^p + \frac{1}{2} \|g\|_p^p \right)^{q-1}.$$

The next estimates follow from Clarkson's inequalities.

PROPOSITION A.3. *Let $\Omega \subseteq \mathcal{R}^d$, $p \in (1, \infty)$, and $q = p/(p-1)$, $\bar{p} = \min(p, q)$.*

Then

- (i) *if $p \in (1, 2]$, the functional e_f^q is uniformly convex with a modulus of convexity $\delta(t) = \frac{t^q}{2^{q-2}}$;*
- (ii) *if $p \geq 2$, the functional e_f^p is uniformly convex with a modulus of convexity $\delta(t) = \frac{t^p}{2^{p-2}}$;*
- (iii) *for all $f, g \in (\mathcal{L}_p(\Omega), \|\cdot\|_p)$, $\|f+g\|_p^{\bar{p}} + \|f-g\|_p^{\bar{p}} \leq 2(\|f\|_p^{\bar{p}} + \|g\|_p^{\bar{p}})$.*

Proof. (i) By [2, Lemma 2.24], for every $1 \leq r < \infty$ and $a, b \geq 0$, $(a + b)^r \leq 2^{r-1}(a^r + b^r)$. Thus, by (A.1) we have

$$\left\| \frac{f+g}{2} \right\|_p^q + \left\| \frac{f-g}{2} \right\|_p^q \leq 2^{q-2} \left(\left(\frac{1}{2} \|f\|_p^p \right)^{q-1} + \left(\frac{1}{2} \|g\|_p^p \right)^{q-1} \right).$$

As $p(q-1) = q$, we obtain

$$\left\| \frac{f+g}{2} \right\|_p^q + \left\| \frac{f-g}{2} \right\|_p^q \leq \frac{2^{q-2}}{2^{q-1}} (\|f\|_p^q + \|g\|_p^q) = \frac{1}{2} (\|f\|_p^q + \|g\|_p^q).$$

(ii) This follows directly from (A.3).

(iii) First suppose that $p \in (1, 2]$. Then $p \leq q$ and so $\bar{p} = \min\{p, q\} = p$. Thus, by (A.2) we have

$$(A.5) \quad \|f\|_p^p + \|g\|_p^p \leq 2 \left(\left\| \frac{f+g}{2} \right\|_p^p + \left\| \frac{f-g}{2} \right\|_p^p \right).$$

Set $\phi = \frac{f+g}{2}$ and $\psi = \frac{f-g}{2}$. Then $f = \phi + \psi$ and $g = \phi - \psi$. So from (A.5) we get $\|\psi + \phi\|_p^p + \|\phi - \psi\|_p^p \leq 2(\|\phi\|_p^p + \|\psi\|_p^p) = \|\psi + \phi\|_p^{\bar{p}} + \|\psi - \phi\|_p^{\bar{p}} \leq 2(\|\psi\|_p^{\bar{p}} + \|\phi\|_p^{\bar{p}})$, which proves (iii) for $p \in (1, 2]$.

Now suppose $p \geq 2$. Then $p \geq q$ and so $\bar{p} = \min\{p, q\} = q$ and $1/(q-1) = p-1$. By (A.4) we have

$$(A.6) \quad (\|f\|_p^p + \|g\|_p^p) \leq 2 \left(\left\| \frac{f+g}{2} \right\|_p^q + \left\| \frac{f-g}{2} \right\|_p^q \right)^{\frac{1}{q-1}} \\ = 2 \left(\left\| \frac{f+g}{2} \right\|_p^q + \left\| \frac{f-g}{2} \right\|_p^q \right)^{p-1}.$$

As above, set $\phi = \frac{f+g}{2}$ and $\psi = \frac{f-g}{2}$. Then $f = \phi + \psi$ and $g = \phi - \psi$. So, from (A.7) we get

$$(A.7) \quad \|\psi + \phi\|_p^p + \|\phi - \psi\|_p^p \leq 2 (\|\phi\|_p^q + \|\psi\|_p^q)^{p-1}.$$

Since for every $r \in [1, \infty)$ and $a, b \geq 0$ we have $(a + b)^r \leq 2^{r-1}(a^r + b^r)$ [2, 2.24], setting $a = \|\phi + \psi\|_p^q$, $b = \|\phi - \psi\|_p^q$, and $r = p/q = p-1$ we obtain

$$(A.8) \quad \|\psi + \phi\|_p^p + \|\phi - \psi\|_p^p = \left(\|\phi + \psi\|_p^{\frac{q}{p}} + \|\phi - \psi\|_p^{\frac{q}{p}} \right) \\ \geq \frac{1}{2^{p-2}} (\|\psi + \phi\|_p^q + \|\phi - \psi\|_p^q)^{p-1}.$$

Thus, by (A.7) and (A.8) we have $(\|\phi + \psi\|_p^q + \|\phi - \psi\|_p^q)^{p-1} \leq 2^{p-1} (\|\phi\|_p^q + \|\psi\|_p^q)^{p-1}$. Hence $\|\phi + \psi\|_p^q + \|\phi - \psi\|_p^q \leq 2 (\|\phi\|_p^q + \|\psi\|_p^q) = 2 (\|\phi\|_p^{\bar{p}} + \|\psi\|_p^{\bar{p}})$, as $q = \bar{p}$. This proves (iii) for $p \geq 2$. $\square$

The next theorem is a slight reformulation of [28, Theorem 5]. The proof is a simplification of the argument from [28, proof of Theorem 5]. For a Banach space $(X, \|\cdot\|)$ and $f \in X$, we denote by Π_f a *peak functional for f* , i.e., a continuous linear functional such that $\|\Pi_f\| = 1$ and $\Pi_f(f) = \|f\|$ [20, p. 1].

THEOREM A.4. Let $\Omega \subseteq \mathcal{R}^d$ be open, G be a subset of $(\mathcal{L}_p(\Omega), \|\cdot\|_p)$, $p \in (1, \infty)$, $f \in \text{cl conv } G$, and $r > 0$ such that $G \subseteq B_r(f, \|\cdot\|)$. Then for every positive integer n , $\|f - \text{span}_n G\|_p \leq \frac{2^{1/\bar{p}} r}{n^{1/\bar{q}}}$, where $q = p/(p-1)$, $\bar{p} = \min(p, q)$, and $\bar{q} = \max(p, q)$.

Proof. As in the proof of Theorem A.1, it is sufficient to verify the statement for $f \in \text{conv } G$. Let $f = \sum_{j=1}^m a_j h_j$ be a representation of f as a convex combination of elements of G . We show by induction that there exists a sequence $\{g_i\}$ of elements of G such that the barycenters $f_n = \sum_{i=1}^n \frac{g_i}{n}$ satisfy $e_n = \|f - f_n\| \leq \frac{2^{1/\bar{p}} r}{n^{1/\bar{q}}}$.

First we check that there exists $g_1 \in G$ such that $f_1 = g_1$ satisfies $e_1 = \|f - f_1\|_p \leq 2^{1/\bar{p}} r$. This holds trivially as $G \subseteq B_r(f, \|\cdot\|)$, so for any $g \in G$ we have $\|f - g\| \leq r < 2^{1/\bar{p}} r$. Hence we can set $f_1 = g_1$ for any $g_1 \in G$.

Assume that we already have $g_1, \dots, g_n$. Then $f_{n+1} = \frac{n}{n+1} f_n + \frac{1}{n+1} g_{n+1} = \frac{1}{n+1} \sum_{i=1}^{n+1} g_i$. We shall express $e_{n+1}^{\bar{p}}$ in terms of $e_n^{\bar{p}}$.

Let Π_n be a peak functional for $f - f_n$. Since $\sum_{j=1}^m a_j (f - h_j) = 0$, by linearity of Π_n we have $0 = \Pi_n (\sum_{j=1}^m a_j (f - h_j)) = \sum_{j=1}^m a_j \Pi_n (f - h_j)$. Thus, there must exist $j \in \{1, \dots, m\}$ such that $\Pi_n (f - h_j) \leq 0$. Set $g_{n+1} = h_j$, so $\Pi_n (f - g_{n+1}) \leq 0$. Hence, by Proposition A.3(iii) we get

$$\begin{aligned} e_{n+1}^{\bar{p}} &= \|f - f_{n+1}\|_p^{\bar{p}} = \left\| \frac{n}{n+1} (f - f_n) + \frac{1}{n+1} (f - g_{n+1}) \right\|_p^{\bar{p}} \\ &\leq 2 \left(\left\| \frac{n}{n+1} (f - f_n) \right\|_p^{\bar{p}} + \left\| \frac{1}{n+1} (f - g_{n+1}) \right\|_p^{\bar{p}} \right) \\ (A.9) \quad &\times \left\| \frac{n}{n+1} (f - f_n) - \frac{1}{n+1} (f - g_{n+1}) \right\|_p^{\bar{p}}. \end{aligned}$$

As $\|\Pi_n\| = 1$ and $\Pi_n (f - g_{n+1}) \leq 0$, we have

$$\begin{aligned} \left\| \frac{n}{n+1} (f - f_n) - \frac{1}{n+1} (f - g_{n+1}) \right\|_p &\geq \left\| \Pi_n \left(\frac{n}{n+1} (f - f_n) - \frac{1}{n+1} (f - g_{n+1}) \right) \right\|_p \\ &\geq \left\| \Pi_n \left(\frac{n}{n+1} (f - f_n) \right) \right\|_p \\ &= \frac{n}{n+1} \|\Pi_n (f - f_n)\|_p = \frac{n}{n+1} \|f - f_n\|_p. \end{aligned}$$

Hence

$$(A.10) \quad - \left\| \frac{n}{n+1} (f - f_n) - \frac{1}{n+1} (f - g_{n+1}) \right\|_p^{\bar{p}} \leq - \left(\frac{n}{n+1} \|f - f_n\|_p \right)^{\bar{p}}.$$

By (A.9) and (A.10),

$$\begin{aligned} e_{n+1}^{\bar{p}} &= \|f - f_{n+1}\|_p^{\bar{p}} \leq 2 \left(\left\| \frac{n}{n+1} (f - f_n) \right\|_p^{\bar{p}} + \left\| \frac{1}{n+1} (f - g_{n+1}) \right\|_p^{\bar{p}} \right) \\ &\quad - \left(\frac{n}{n+1} \|f - f_n\|_p \right)^{\bar{p}} = \frac{2}{(n+1)^{\bar{p}}} \|f - g_{n+1}\|_p^{\bar{p}} + \left(\frac{2}{n+1} \right)^{\bar{p}} \|f - f_n\|_p^{\bar{p}} \\ &= \frac{2}{(n+1)^{\bar{p}}} \|f - g_{n+1}\|_p^{\bar{p}} + \left(\frac{2}{n+1} \right)^{\bar{p}} e_n^{\bar{p}}. \end{aligned}$$

As $e_n = \|f - f_n\| \leq \frac{2^{1/\bar{p}} r}{n^{1/\bar{q}}}$, we get

$$\begin{aligned} e_{n+1}^{\bar{p}} &\leq \frac{2 r^{\bar{p}}}{(n+1)^{\bar{p}}} + \left(\frac{n}{n+1}\right)^{\bar{p}} \left(\frac{2^{1/\bar{p}} r}{n^{1/\bar{q}}}\right)^{\bar{p}} = \frac{2 r^{\bar{p}}}{(n+1)^{\bar{p}}} \left(1 + \frac{n^{\bar{p}}}{n^{\bar{p}/\bar{q}}}\right) \\ &= \frac{2 r^{\bar{p}}}{(n+1)^{\bar{p}}} \left(1 + n^{\bar{p}-\bar{p}/\bar{q}}\right). \end{aligned}$$

It can be easily verified that $\bar{p} - \frac{\bar{p}}{\bar{q}} = 1$ in both cases, $\bar{p} = p$ (and so $\bar{q} = q = \frac{p}{p-1}$), and $\bar{p} = q$ (and so $\bar{q} = p$). Thus $e_{n+1}^{\bar{p}} \leq \frac{2 r^{\bar{p}}}{(n+1)^{\bar{p}}}$. As $\bar{p} - 1 = \frac{\bar{p}}{\bar{q}}$ for both $\bar{p} = p$ (hence $\bar{q} = q$) and $\bar{p} = q$ (hence $\bar{q} = p$), we get $e_{n+1}^{\bar{p}} \leq \frac{2 r^{\bar{p}}}{(n+1)^{\bar{p}-1}} = \frac{2 r^{\bar{p}}}{(n+1)^{\bar{p}/\bar{q}}}$, i.e., $e_{n+1} \leq \frac{2^{1/\bar{p}} r}{(n+1)^{1/\bar{q}}}$. $\square$

Theorem 3.1 is a corollary of Theorems A.1 and A.4 in terms of G -variation. For $t > 0$, we define $G(t) = \{wg : g \in G, w \in \mathcal{R}, |w| \leq t\}$.

Proof of Theorem 3.1. (i) As $\text{span}_n G \supseteq \text{conv}_n G$, by Theorem A.1 applied to $G(\|f\|_G)$ we have $\|f - \text{span}_n G\| \leq \|f - \text{conv}_n G(\|f\|_G)\| \leq \sqrt{\frac{(s_G \|f\|_G)^2 - \|f\|^2}{n}}$.

(ii) As $\text{span}_n G \supseteq \text{conv}_n G$, by applying Theorem A.4 to $G(\|f\|_G)$ we get $\|f - \text{span}_n G\| \leq \frac{2^{1/\bar{p}} r}{n^{1/\bar{q}}}$ for every r such that $G(\|f\|_G) \subseteq B_r(f, \|\cdot\|)$. Set $r = 2 s_G \|f\|_G$. By (3.1), for every $h \in G(\|f\|_G)$ we have $\|h - f\| \leq \|h\| + \|f\| \leq s_G \|h\|_G + s_G \|f\|_G \leq 2 s_G \|f\|_G$. So $G(\|f\|_G) \subseteq B_{2r\|f\|_G}(f, \|\cdot\|)$, and hence $\|f - \text{span}_n G\| \leq \frac{2^{1/\bar{p}} s_G \|f\|_G}{n^{1/\bar{q}}}$.

Acknowledgments. The authors thank R. Zoppoli (University of Genoa) for stimulating their interest in the theoretical investigation of approximate optimization by the extended Ritz method. They are also grateful to P. C. Kainen and A. Vogt (both of Georgetown University) for fruitful comments and discussions.

REFERENCES

- [1] E. AARTS AND J. KORST, *Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing*, John Wiley and Sons, New York, 1989.
- [2] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
- [3] A. ALESSANDRI, M. BAGLIETTO, T. PARISINI, AND R. ZOPPOLI, *A neural state estimator with bounded errors for nonlinear systems*, IEEE Trans. Automat. Control, 44 (1999), pp. 2028–2042.
- [4] A. ALESSANDRI, M. SANGUINETI, AND M. MAGGIORE, *Optimization-based learning with bounded error for feedforward neural networks*, IEEE Trans. Neural Networks, 13 (2002), pp. 261–273.
- [5] A. ALESSANDRI AND M. SANGUINETI, *Optimization of approximating networks for optimal fault diagnosis*, Optim. Methods Softw., 20 (2005), pp. 235–260.
- [6] W. ALT, *On the approximation of infinite optimization problems with an application to optimal control problems*, Appl. Math. Optim., 12 (1984), pp. 15–27.
- [7] N. ARONSZAJN, *Theory of reproducing kernels*, Trans. Amer. Math. Soc., 68 (1950), pp. 337–404.
- [8] M. BAGLIETTO, T. PARISINI, AND R. ZOPPOLI, *Numerical solutions to the Witsenhausen counterexample by approximating networks*, IEEE Trans. Automat. Control, 46 (2001), pp. 1471–1477.
- [9] M. BAGLIETTO, T. PARISINI, AND R. ZOPPOLI, *Distributed-information neural control: The case of dynamic routing in traffic networks*, IEEE Trans. Neural Networks, 12 (2001), pp. 485–502.
- [10] M. BAGLIETTO, M. SANGUINETI, AND R. ZOPPOLI, *Facing the curse of dimensionality by the extended Ritz method in stochastic functional optimization: Dynamic routing in traffic networks*, in High Performance Algorithms and Software for Nonlinear Optimization, G. Di Pillo and A. Murli, eds., Kluwer Academic Publishers, Norwell, MA, 2003, pp. 23–56.

- [11] A. R. BARRON, *Neural net approximation*, in Proceedings of the 7th Yale Workshop on Adaptive and Learning Systems, New Haven, CT, 1992, pp. 69–72.
- [12] A. R. BARRON, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Trans. Inform. Theory, 39 (1993), pp. 930–945.
- [13] R. W. BEARD AND T. W. McLAIN, *Successive Galerkin approximation algorithms for nonlinear optimal and robust control*, Internat. J. Control, 71 (1998), pp. 717–743.
- [14] R. BELLMAN, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.
- [15] M. BERTERO, *Linear inverse and ill-posed problems*, Adv. Electron. Electron. Phys., 75 (1989), pp. 1–120.
- [16] D. P. BERTSEKAS, *A new class of incremental gradient methods for least squares problems*, SIAM J. Optim., 7 (1997), pp. 913–926.
- [17] D. P. BERTSEKAS, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1999.
- [18] D. P. BERTSEKAS AND J. N. TSITSIKLIS, *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA, 1996.
- [19] W. E. BOSARGE, JR., O. G. JOHNSON, R. S. MCKNIGHT, AND W. P. TIMLAKE, *The Ritz-Galerkin procedure for nonlinear control problems*, SIAM J. Numer. Anal., 10 (1973), pp. 94–111.
- [20] D. BRAESS, *Nonlinear Approximation Theory*, Springer, Berlin, 1986.
- [21] L. BREIMAN, *Hinging hyperplanes for regression, classification, and function approximation*, IEEE Trans. Inform. Theory, 39 (1993), pp. 999–1013.
- [22] H. BREZIS, *Analyse Fonctionnelle—Théorie et Applications*, Masson, Paris, 1983.
- [23] F. C. CHEN AND H. KHALIL, *Adaptive control of a class of nonlinear discrete-time systems using multilayer neural networks*, IEEE Trans. Automat. Control, 40 (1995), pp. 791–801.
- [24] V. C. P. CHEN, D. RUPPERT, AND C. A. SHOEMAKER, *Applying experimental design and regression splines to high-dimensional continuous-state stochastic dynamic programming*, Oper. Res., 47 (1999), pp. 38–53.
- [25] F. CUCKER AND S. SMALE, *On the mathematical foundations of learning*, Bull. Amer. Math. Soc. (N.S.), 39 (2001), pp. 1–49.
- [26] J. W. DANIEL, *The Approximate Minimization of Functionals*, Prentice-Hall, Englewood Cliffs, NJ, 1971.
- [27] J. W. DANIEL, *The Ritz-Galerkin method for abstract optimal control problems*, SIAM J. Control, 11 (1973), pp. 53–63.
- [28] C. DARKEN, M. DONAHUE, L. GURVITS, AND E. SONTAG, *Rate of approximation results motivated by robust neural network learning*, in Proceedings of the Sixth Annual ACM Conference on Computational Learning Theory, The Association for Computing Machinery, New York, 1993, pp. 303–309.
- [29] F. DEUTSCH, *Best Approximation in Inner Product Spaces*, Springer, New York, 2001.
- [30] R. A. DEVORE, R. HOWARD, AND C. MICCHELLI, *Optimal nonlinear approximation*, Manuscripta Math., 63 (1989), pp. 469–478.
- [31] R. A. DEVORE AND G. G. LORENTZ, *Constructive Approximation*, Grundlehren Math. Wiss. 303, Springer, Berlin, 1993.
- [32] R. A. DEVORE AND V. N. TEMLYAKOV, *Nonlinear approximation by trigonometric sums*, J. Fourier Anal. Appl., 2 (1995), pp. 29–48.
- [33] A. L. DONTCHEV, *Perturbations, Approximations and Sensitivity Analysis of Optimal Control Systems*, Lecture Notes in Control and Inform. Sci. 52, Springer, Berlin, Heidelberg, 1983.
- [34] A. L. DONTCHEV AND T. ZOLEZZI, *Well-Posed Optimization Problems*, Lecture Notes in Math. 1543, Springer, Berlin, Heidelberg, 1993.
- [35] I. EKELAND AND R. TEMAM, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, Oxford, American Elsevier, New York, 1976.
- [36] U. FELGENHAUER, *On Ritz type discretizations for optimal control problems*, in Proceedings of the 18th IFIP-ICZ Conference, Res. Notes in Math. 386, Chapman and Hall, London, 1999, pp. 91–99.
- [37] I. M. GELFAND AND S. V. FOMIN, *Calculus of Variations*, Prentice-Hall, Englewood Cliffs, NJ, 1963.
- [38] F. GIROSI, M. JONES, AND T. POGGIO, *Regularization theory and neural networks architectures*, Neural Computation, 7 (1995), pp. 219–269.
- [39] D. E. GOLDBERG, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, Reading, MA, 1989.
- [40] L. GRIPPO, *Convergent on-line algorithms for supervised learning in neural networks*, IEEE Trans. Neural Networks, 11 (2000), pp. 1284–1299.
- [41] W. W. HAGER, *The Ritz-Trefftz method for state and control constrained optimal control problems*, SIAM J. Numer. Anal., 12 (1975), pp. 854–867.

- [42] E. HEWITT AND K. STROMBERG, *Abstract Analysis*, Springer, Berlin, 1965.
- [43] R. B. HOLMES, *Approximating best approximations*, Nieuw Archief voor Wiskunde, XIV(3), (1966), pp. 106–113.
- [44] R. B. HOLMES, *A Course on Optimization and Best Approximation*, Lecture Notes in Math. 257, Springer, Berlin, 1972.
- [45] L. K. JONES, *A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training*, Ann. Statist., 20 (1992), pp. 608–613.
- [46] S. A. JOHNSON, J. R. STEDINGER, C. SHOEMAKER, Y. LI, AND J. TEJADA-GUIBERT, *Numerical solution of continuous-state dynamic programs using linear and spline interpolation*, Oper. Res., 41 (1993), pp. 484–500.
- [47] A. JUDITSKY, H. HJALMARSSON, A. BENVENISTE, B. DELYON, L. LJUNG, J. SJÖBERG, AND Q. ZHANG, *Nonlinear black-box models in system identification: Mathematical foundations*, Automatica, 31 (1995), pp. 1725–1750.
- [48] P. C. KAINEN, V. KŮRKOVÁ, AND M. SANGUINETI, *Minimization of error functionals over variable-basis functions*, SIAM J. Optim., 14 (2003), pp. 732–742.
- [49] P. C. KAINEN, V. KŮRKOVÁ, AND A. VOGT, *An integral formula for Heaviside neural networks*, Neural Network World, 10 (2000), pp. 313–319.
- [50] P. C. KAINEN, V. KŮRKOVÁ, AND A. VOGT, *Upper Bounds on Variation with Respect to Half-Spaces*, Research Report ICS-2004-900, Institute of Computer Science, Prague, 2004.
- [51] V. KŮRKOVÁ, *Dimension-independent rates of approximation by neural networks*, in Computer-Intensive Methods in Control and Signal Processing. The Curse of Dimensionality, K. Warwick and M. Kárný, eds., Birkhäuser, Boston, 1997, pp. 261–270.
- [52] V. KŮRKOVÁ, *Neural networks as universal approximators*, in The Handbook of Brain Theory and Neural Networks, M. Arbib, ed., MIT Press, Cambridge, MA, 2002, pp. 1180–1183.
- [53] V. KŮRKOVÁ, *High-dimensional approximation and optimization by neural networks*, in Advances in Learning Theory: Methods, Models, and Applications, J. Suykens, G. Horvath, S. Basu, C. Micchelli, and J. Vandewalle, eds., IOS Press, Amsterdam, 2003, pp. 69–88.
- [54] V. KŮRKOVÁ, P. C. KAINEN, AND V. KREINOVICH, *Estimates of the number of hidden units and variation with respect to half-spaces*, Neural Networks, 10 (1997), pp. 1061–1068.
- [55] V. KŮRKOVÁ AND M. SANGUINETI, *Bounds on rates of variable-basis and neural-network approximation*, IEEE Trans. Inform. Theory, 47 (2001), pp. 2659–2665.
- [56] V. KŮRKOVÁ AND M. SANGUINETI, *Comparison of worst case errors in linear and neural network approximation*, IEEE Trans. Inform. Theory, 48 (2002), pp. 264–275.
- [57] V. KŮRKOVÁ AND M. SANGUINETI, *Learning with generalization capability by kernel methods of bounded complexity*, J. Complexity, to appear.
- [58] V. KŮRKOVÁ, P. SAVICKÝ, AND K. HLAVÁČKOVÁ, *Representations and rates of approximation of real-valued Boolean functions by neural networks*, Neural Networks, 11 (1998), pp. 651–659.
- [59] E. S. LEVITIN AND B. T. POLYAK, *Convergence of minimizing sequences in conditional extremum problems*, Dokl. Akad. Nauk SSSR, 168 (1966), pp. 764–767.
- [60] A. LINNEMANN, *Convergent Ritz approximations of the set of stabilizing controllers*, Systems Control Lett., 36 (1998), pp. 151–156.
- [61] K. S. NARENDRA AND S. MUKHOPADHYAY, *Adaptive control using neural networks and approximate models*, IEEE Trans. Neural Networks, 8 (1997), pp. 475–485.
- [62] K. S. NARENDRA AND K. PARTHASARATHI, *Identification and control of dynamical systems using neural networks*, IEEE Trans. Neural Networks, 4 (1990), pp. 4–26.
- [63] J. M. ORTEGA, *Numerical Analysis: A Second Course*, Classics Appl. Math. 3, SIAM, Philadelphia, 1990.
- [64] T. PARISINI, M. SANGUINETI, AND R. ZOPPOLI, *Nonlinear stabilization by receding-horizon neural regulators*, Internat. J. Control, 70 (1998), pp. 341–362.
- [65] T. PARISINI AND R. ZOPPOLI, *Neural networks for feedback feedforward nonlinear control systems*, IEEE Trans. Neural Networks, 5 (1994), pp. 436–449.
- [66] T. PARISINI AND R. ZOPPOLI, *Neural approximations for multistage optimal control of nonlinear stochastic systems*, IEEE Trans. Automat. Control, 41 (1996), pp. 889–895.
- [67] A. PINKUS, *n-Widths in Approximation Theory*, Springer, Berlin, Heidelberg, 1985.
- [68] G. PISIER, *Remarques sur un résultat non publié de B. Maurey*, Séminaire d'Analyse Fonctionnelle 1980-81, Exposé no. V, École Polytechnique, Centre de Mathématiques, Palaiseau, France, pp. V.1–V.12.
- [69] T. POGGIO AND S. SMALE, *The mathematics of learning: Dealing with data*, Notices Amer. Math. Soc., 50 (2003), pp. 536–544.
- [70] B. T. POLAK, *Existence theorems and convergence of minimizing sequences in extremum prob-*

- lems with restrictions*, Dokl. Akad. Nauk SSSR, 166 (1966), pp. 72–75.
- [71] D. E. RUMELHART, G. E. HINTON, AND R. J. WILLIAMS, *Learning internal representation by error propagation*, in *Parallel Distributed Processing: Explorations in the Microstructures of Cognition, I: Foundations*, D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, eds., MIT, Cambridge, MA, 1986, pp. 318–362.
 - [72] I. SINGER, *Best Approximation in Normed Linear Spaces by Elements of Subspaces*, Springer, Berlin, 1970.
 - [73] H. R. SIRISENA AND F. S. CHOU, *Convergence of the control parametrization Ritz method for nonlinear optimal control problems*, J. Optim. Theory Appl., 29 (1979), pp. 369–382.
 - [74] J. SJÖBERG, Q. ZHANG, L. LJUNG, A. BENVENISTE, P.-Y. GLORENNEC, B. DELYON, H. HJALMARSSON, AND A. JUDITSKY, *Nonlinear black-box modeling in system identification: A unified overview*, Automatica, 31 (1995), pp. 1691–1724.
 - [75] V. B. TJUHTIN, *An error estimate for approximate solutions in one-sided variational problems*, Vestnik Leningrad Univ. Math., 14 (1982), pp. 247–254.
 - [76] P. TSENG, *Incremental gradient(-projection) method with momentum term and adaptive step-size rule*, SIAM J. Optim., 8 (1998), pp. 506–531.
 - [77] J. F. TRAUB AND A. G. WERSCHULZ, *Complexity and Information*, Cambridge University Press, Cambridge, UK, 1998.
 - [78] A. A. VLADIMIROV, YU. E. NESTEROV, AND YU. N. CHEKANOV, *On uniformly convex functionals*, Vestnik Moskov. Univ. Ser. 15 Vychisl. Mat. Kibernet., 3 (1978), pp. 12–23 (in Russian); Moscow University Computational Mathematics and Cybernetics (1979), pp. 10–21 (in English).
 - [79] G. YIN, *Rates of convergence for a class of global stochastic optimization algorithms*, SIAM J. Optim., 10 (1999), pp. 99–120.
 - [80] R. ZOPPOLI AND T. PARISINI, *Learning techniques and neural networks for the solution of N-stage nonlinear nonquadratic optimal control problems*, in *Systems, Models and Feedback: Theory and Applications*, A. Isidori and T. J. Tarn, eds., Birkhäuser, Boston, 1992, pp. 193–210.
 - [81] R. ZOPPOLI, M. SANGUINETI, AND T. PARISINI, *Approximating networks and extended Ritz method for the solution of functional optimization problems*, J. Optim. Theory Appl., 112 (2002), pp. 403–440.

ON CONSTRAINT QUALIFICATION FOR AN INFINITE SYSTEM OF CONVEX INEQUALITIES IN A BANACH SPACE*

CHONG LI[†] AND K. F. NG[‡]

Abstract. For a general infinite system of convex inequalities in a Banach space, we study the basic constraint qualification and its relationship with other fundamental concepts, including various versions of conditions of Slater type, the Mangasarian–Fromovitz constraint qualification, as well as the Pshenichnyi–Levin–Valadier property introduced by Li, Nahak, and Singer. Applications are given in the restricted range approximation problem, constrained optimization problems, as well as in the approximation problem with constraints by conditionally positive semidefinite functions.

Key words. system of convex inequalities, basic constraint qualification, Slater condition, MFCQ, PLV property, optimality condition, best approximation

AMS subject classifications. Primary, 90C34; Secondary, 90C46, 41A29

DOI. 10.1137/S1052623403434693

1. Introduction. Let X be a Banach space over the real field $\mathbb{R}$, C be a closed convex subset of X , and let $\{g_i : i \in I\}$ be a family of continuous convex functions on X , where I is an arbitrary (but nonempty) index-set. We consider the following system of convex inequalities:

$$(1.1) \quad g_i(x) \leq 0, \quad i \in I,$$

and the associated system for the sup-function G ,

$$(1.2) \quad G(x) \leq 0,$$

where G is defined by

$$(1.3) \quad G(x) := \sup_{i \in I} g_i(x) \quad \text{for all } x \in X.$$

In this paper, we always assume that the sup-function $G(\cdot)$ is continuous on X . It is clear that if X is of finite dimension and $G(x)$ is finite for each $x \in X$, or if $\{g_i : i \in I\}$ is locally uniformly bounded, then the condition that $G(\cdot)$ is continuous on X is automatically satisfied. Throughout we use S to denote the solution set of (1.1) (equivalently, of (1.2)):

$$(1.4) \quad S := \{x \in X : g_i(x) \leq 0 \text{ for all } i \in I\} = \{x \in X : G(x) \leq 0\}.$$

Let

$$(1.5) \quad K := C \cap S;$$

*Received by the editors September 17, 2003; accepted for publication (in revised form) June 28, 2004; published electronically February 19, 2005.

<http://www.siam.org/journals/siopt/15-2/43469.html>

[†]Department of Mathematics, Zhejiang University, Hangzhou 310027, P. R. China (cli@zju.edu.cn). This author was supported in part by National Natural Science Foundation of China grant 10271025.

[‡]Department of Mathematics, Chinese University of Hong Kong, Hong Kong, P. R. China (kfng@math.cuhk.edu.hk). This author was supported by a Direct Grant (CUHK) and an Earmarked Grant from the Research Grant Council of Hong Kong.

that is, K is the solution set of (1.1) with the constraint

$$(1.6) \quad x \in C.$$

Many important problems in mathematical programming are based on relationships between the normal cone $N_K(x)$ in relation to $N_C(x)$ and the subdifferentials of G, g_i with $i \in I(x)$. Here $x \in X$, and $I(x)$ is the set of “active indices” at x defined by

$$(1.7) \quad I(x) := \{i \in I : g_i(x) = G(x)\}.$$

In the classical case (that is, when I is finite, X is a Euclidean space, and each g_i is differentiable), many fundamental notions of so-called constraint qualifications for (1.1) have been introduced, such as the basic constraint qualification (BCQ), the Slater condition (SC), and the Mangasarian–Fromovitz constraint qualification (MFCQ); their interrelationships are well known and play a very important part not only in optimization problems but also in many other areas, such as best approximation theory. A big step forward from the classical theory was recently achieved by Li, Nahak, and Singer in [16], where detailed studies on the BCQ and the SC have been made for the case when the index-set I is not restricted to be finite. They introduced and studied the Pshenichnyi–Levin–Valadier (PLV) property and the weak PLV property; in particular, they made use of these properties in extending the classical theory to the case when I is infinite. Another direction of extension was done in [12, 13], where I was assumed to be finite but X was allowed to be an infinite dimensional Hilbert space; moreover, the constraint (1.6) was considered.

In this paper, we propose to study the general case: X is a Banach space, C is a closed convex subset, and I is an arbitrary set. In this general case we introduce and study various notions of constraint qualifications. In particular, in the stated general setting, the notions of the BCQ and the PLV are introduced in section 2. We show in Theorem 2.4 that, in the presence of the PLV, (1.1) satisfies the BCQ if and only if (1.2) satisfies the BCQ. Our main results are presented in section 3, where we establish some sufficient conditions to ensure the BCQ. Along with the extended notions of the SC and the MFCQ, the weak versions (weak SC and weak MFCQ) are also introduced. Some of their relationships are established. We show in particular that, under the weak SC, the PLV implies the BCQ for (1.1). We remark that many of the results obtained in this paper, for example, Theorems 3.8 and 3.10, are new even for the case when $C = X$ is a finite dimensional space. In particular, Theorem 3.10 provides a set of conditions ensuring that the following implication holds:

Each finite subsystem of (1.1) satisfies the BCQ $\implies$ the system (1.1) does.

Meanwhile, Examples 3.2, 3.3, and 3.4 show that the above implication is no longer true if any of the conditions in the theorem are dropped. The last three sections of the paper are devoted to applications of the main results. In section 4, we study the problem of minimizing a real-valued function $f(x)$ subjected to $x \in C \cap S$ and we show that, for an optimal point x_0 , the BCQ of (1.1) relative to C at x_0 is closely related to the KKT property. In section 6, we address an approximation problem studied by H. Strauss. We give a counterexample to show that [22, Theorem 6.2] is not true, and we provide a new version which is established under the added assumption of the BCQ.

One of the reasons why we study the proposed general case instead of the more restrictive case is that there are many approximation and optimization problems which

are in the general Banach space setting with infinitely many convex constraints. For example, it was shown in [14] that the restricted Chebyshev approximation problem (see [11, 21] and references therein) in the Banach space of complex-valued continuous functions on a compact Hausdorff space can be reformulated as a constrained approximation problem with a system of convex inequalities. Section 5 is devoted to another application dealing with the best restricted range approximation problem in L^p space studied by Levasseur and Levis: a main result from [10] is extended and improved under much weaker conditions.

2. Preliminaries, the BCQ, and the PLV property. For a set Z in X (or in $\mathbb{R}^n$), $\text{ext } Z$ denotes the set of all extreme points of Z and the interior (resp., relative interior, closure, convex hull, convex cone hull, linear hull, negative polar, boundary) of Z is defined by $\text{int } Z$ (resp., $\text{ri } Z$, $\bar{Z}$, $\text{conv } Z$, $\text{cone } Z$, $\text{span } Z$, $Z^\ominus$, $\text{bd } Z$). We use $|Z|(\leq \infty)$ to denote the number of elements of Z and use $\dim Z$ to denote the dimension of $\text{span}(Z - z)$, where z is an element of Z . For $x \in X$, in the case when Z is closed and convex, the normal cone of Z at x is denoted by $N_Z(x)$ and defined by $N_Z(x) = (Z - x)^\ominus$. $\mathbb{R}_-$ denotes the subset of $\mathbb{R}$ consisting of all nonpositive real numbers.

For a proper convex function on X and $x \in X$, the subdifferential of f at x is defined by

$$\partial f(x) := \{z^* \in X^* : f(x) + \langle z^*, y - x \rangle \leq f(y) \text{ for all } y \in X\}.$$

In particular, $N_Z(z) = \partial \mathbf{I}_Z(z)$ for each $z \in Z$. Here and throughout, $\mathbf{I}_Z$ denotes the indicator function of Z : $I_Z(x) = 0$ if $x \in Z$ and $I_Z(x) = +\infty$ if $x \notin Z$.

Recall that the directional derivative $f'(x, d)$ of the function f at x in the direction d is defined by

$$(2.1) \quad f'(x, d) = \lim_{t \rightarrow 0+} \frac{f(x + td) - f(x)}{t}.$$

Then (cf. [6, Proposition 2.2.7]) if f is convex and if x is a continuity point of f ,

$$(2.2) \quad \partial f(x) = \{z^* \in X^* : \langle z^*, d \rangle \leq f'(x, d) \text{ for all } d \in X\}$$

and

$$(2.3) \quad f'(x, d) = \max\{\langle z^*, d \rangle : z^* \in \partial f(x)\}.$$

Remark 2.1. For a continuous convex function f on X , it is easy to show that $\text{cone } \partial f(x) \subseteq N_{f^{-1}(\mathbb{R}_-)}(x)$ if $f(x) = 0$ and that the equality holds in both of the following cases: (a) f is affine, (b) x is not a minimizer of f . See [6, Corollary 1, p. 56].

DEFINITION 2.1. Let $x \in X$, and let $I(x)$ be defined by (1.7). Assuming $I(x) \neq \emptyset$, define $D'(x)$ and $N'(x)$ by

$$(2.4) \quad D'(x) := \text{conv} \left(\bigcup_{i \in I(x)} \partial g_i(x) \right),$$

$$(2.5) \quad N'(x) := \text{cone} \left(\bigcup_{i \in I(x)} \partial g_i(x) \right);$$

also we adopt the convention that

$$(2.6) \quad N'(x) = \{0\} \text{ if } I(x) = \emptyset.$$

Part (ii) of the following proposition is well known while the other parts can be verified easily by definitions.

PROPOSITION 2.2. *The following assertions hold:*

(i) *If $i \in I(x)$, then $\partial g_i(x) \subseteq \partial G(x)$. Hence*

$$(2.7) \quad N'(x) \subseteq \text{cone } \partial G(x) \text{ for all } x \in X$$

and

$$(2.8) \quad D'(x) \subseteq \partial G(x) \text{ for all } x \in X.$$

(ii) *If I is finite, then*

$$(2.9) \quad D'(x) = \partial G(x) \text{ for all } x \in X.$$

(iii) *If $G(x) = 0$ (e.g., if $x \in \text{bd } S$), then*

$$(2.10) \quad \partial G(x) \subseteq N_K(x)$$

and

$$(2.11) \quad N_C(x) + N'(x) \subseteq N_C(x) + \text{cone } \partial G(x) \subseteq N_K(x).$$

Below we define the conversed relations of the inclusions given in the preceding proposition.

DEFINITION 2.3. *Let $x \in X$. The system (1.1) is said to satisfy*

(a) *the BCQ at x relative to C if*

$$(2.12) \quad N_C(x) + N'(x) \supseteq N_K(x);$$

(b) *the PLV at x relative to C if*

$$(2.13) \quad N_C(x) + D'(x) \supseteq N_C(x) + \partial G(x).$$

The system (1.1) is said to satisfy the BCQ relative to C if it satisfies the BCQ at each point of K . In addition, the wording “relative to C ” need not be mentioned if $C = X$.

The following theorem (the proof of which is routine and hence omitted) extends a result of Li, Nahak, and Singer for the case when $C = X$ is finite dimensional.

THEOREM 2.4. *Let $x \in C \cap \text{bd } S$. Suppose that the system (1.1) satisfies the PLV at x relative to C . Then the system (1.1) satisfies the BCQ at x relative to C if and only if (1.2) satisfies the BCQ at x relative to C .*

Part (ii) of the following theorem is well known; see, for example, [9, Theorem 4.4.2] and [16, Theorem 3.1]. The proof given in [16] is valid for any compact space I satisfying the first axiom of countability (in the sense that there exists a countable local base at each point of I). For wider scope of applications (such as that in section 5), we do not assume that I is metrizable.

THEOREM 2.5. *Suppose that I is a compact space satisfying the first axiom of countability and that the function: $i \mapsto g_i(x)$ is upper semicontinuous for each $x \in X$. Then $I(x) \neq \emptyset$ for each $x \in X$ and the following assertions hold:*

- (i) $\partial G(x) = \overline{D'(x)}^{w*}$.
- (ii) $\partial G(x) = D'(x)$, provided that $I(x)$ is finite or X is finite dimensional.
- (iii) If $x \in \text{span } C$ and if $\text{span } C$ is finite dimensional, then the system (1.1) satisfies the PLV relative to C at x .

Furthermore, if in addition $\{g_i : i \in I\}$ is equicontinuous on X , then the assumption on I about the first axiom of countability can be dropped.

Proof. Let $x, d \in X$. By assumption, it is easy to verify that $I(x) \neq \emptyset$. For each $t > 0$, define

$$(2.14) \quad I_t := \left\{ i \in I : \frac{g_i(x + td) - G(x)}{t} \geq G'(x, d) \right\}.$$

By convexity, $I(x + td) \subseteq I_t$ and it follows that I_t is nonempty. Moreover, $\{I_t : t > 0\}$ has the “finite intersection property” (and hence $\cap_{t>0} I_t \neq \emptyset$) because the function

$$(2.15) \quad t \mapsto \frac{g_i(x + td) - G(x)}{t} = \frac{g_i(x + td) - g_i(x)}{t} + \frac{g_i(x) - G(x)}{t}$$

is nondecreasing on $(0, +\infty)$ (the functions represented in the two terms of the right-hand side of (2.15) are nondecreasing by the convexity of g_i and the fact that $g_i(x) - G(x) \leq 0$). Pick $i_0 \in \cap_{t>0} I_t$. Then passing to the limits as $t \rightarrow 0$ in

$$(2.16) \quad g_{i_0}(x + td) - G(x) \geq tG'(x, d),$$

we see that $g_{i_0}(x) = G(x)$ (that is, $i_0 \in I(x)$) and so (2.16) entails $g'_{i_0}(x, d) \geq G'(x, d)$. Consequently, it follows from (2.3) that

$$(2.17) \quad G'(x, d) = \max_{i \in I(x)} \{g'_i(x, d)\} = \max_{i \in I(x)} \left\{ \max_{x^* \in \partial g_i(x)} \langle x^*, d \rangle \right\} = \max_{x^* \in D'(x)} \langle x^*, d \rangle.$$

Thus (i) must hold by [6, Proposition 2.1.4]. Note that we have not used the assumption on the first axiom of countability. For (iii), write $\tilde{X}$ for $\text{span } C$; let $\tilde{G}$ and $\tilde{g}_i$, respectively, denote the restrictions of G and g_i to $\tilde{X}$. Clearly, $\tilde{G}$ is the sup-function of $\{\tilde{g}_i : i \in I\}$ and it follows from (ii) that, for $x \in \tilde{X}$,

$$(2.18) \quad \partial \tilde{G}(x) = \text{conv} \{ \partial \tilde{g}_i(x) : i \in I(x) \}.$$

To prove (iii), we need only show that

$$(2.19) \quad \partial G(x) \subseteq N_C(x) + D'(x).$$

To do this, let $y^* \in \partial G(x)$, and let $\tilde{y}^*$ denote the restriction of y^* to $\tilde{X}$. Then $\tilde{y}^* \in \partial \tilde{G}(x)$, and so it follows from (2.18) that $\tilde{y}^* = \sum_{j=1}^m \lambda_j \tilde{y}_j^*$ for some $\{\lambda_j\} \subseteq [0, 1]$ and $\{\tilde{y}_j^*\} \subseteq \tilde{X}^*$ with $\sum_{j=1}^m \lambda_j = 1$, $\tilde{y}_j^* \in \partial \tilde{g}_{i_j}(x)$, and $i_1, i_2, \dots, i_m \in I(x)$. Noting

$$(2.20) \quad \langle \tilde{y}_j^*, z \rangle \leq \tilde{g}'_{i_j}(x, z) = g'_{i_j}(x, z) \quad \text{for all } z \in \tilde{X},$$

one can apply the Hahn–Banach theorem to obtain a continuous linear extension y_j^* of $\tilde{y}_j^*$ such that

$$\langle y_j^*, \cdot \rangle \leq g'_{i_j}(x, \cdot) \quad \text{on } X.$$

Thus $y_j^* \in \partial g_{i_j}(x)$ and

$$\langle y^*, \cdot \rangle = \left\langle \sum_{j=1}^m \lambda_j y_j^*, \cdot \right\rangle \quad \text{on } \tilde{X}.$$

Since $C - x \subseteq \tilde{X}$, this implies that $y^* - \sum_{j=1}^m \lambda_j y_j^* \in N_C(x)$. Therefore $y^* \in N_C(x) + N'(x)$ and (2.19) is proved.

Now we drop the assumption that I satisfies the first axiom of countability but assume that $\{g_i : i \in I\}$ is equicontinuous on X . By the above proof, (iii) will be true if (ii) holds. Hence, to complete the proof it suffices to show that $\partial G(x) = D'(x)$ when X is finite dimensional. For this purpose, we need to introduce a new topology on the set I . For each $i_0 \in I$, $x \in X$, and each $r \in \mathbb{R}, r > 0$, set

$$O_x(i_0, r) = \{i \in I : g_i(x) - g_{i_0}(x) < r\}.$$

Since X is finite dimensional, there exists a countable subset A of X such that the closure $\bar{A}$ is equal to X . Let τ_N denote the topology generated by $\{O_x(i_0, r) : x \in A, i_0 \in I, r \in \mathbb{R}, r > 0\}$. Then, under the topology τ_N , the following assertions hold:

- (1) I is compact;
- (2) I satisfies the first axiom of countability;
- (3) for each $x \in X$, the function $i \mapsto g_i(x)$ is upper semicontinuous.

In fact, (2) is trivial, and (1) holds because I is compact under the original topology, which is stronger than the new topology τ_N , as each function $i \mapsto g_i(x)$ is upper semicontinuous. (3) follows from the density of A in X and the equicontinuity of $\{g_i : i \in I\}$ on X . In fact, let $x \in X$ and $\epsilon > 0$. By the equicontinuity, there exists $\delta > 0$ such that

$$(2.21) \quad |g_i(x) - g_i(y)| < \frac{\epsilon}{3} \quad \text{for each } i \in I \text{ and each } y \text{ with } \|x - y\| < \delta.$$

Pick $x_\epsilon \in A$ such that $\|x_\epsilon - x\| < \delta$. Let $i_0 \in I$. Noting that, by (2.21),

$$(2.22) \quad g_i(x) - g_{i_0}(x) < \frac{2}{3}\epsilon + g_i(x_\epsilon) - g_{i_0}(x_\epsilon),$$

we get

$$(2.23) \quad g_i(x) - g_{i_0}(x) < \epsilon \quad \text{for each } i \in O_{x_\epsilon}(i_0, \epsilon/3).$$

This shows that the function $i \mapsto g_i(x)$ is upper semicontinuous at i_0 , and hence (3) holds. Thus, by the first part of the proof, it is easy to see that $\partial G(x) = D'(x)$ and the proof is complete. $\square$

Remark 2.2. In general, $D'(x)$ need not be w^* -closed, and hence the w^* -closure in (i) cannot be dropped. For example, take $X = l^2$, the Hilbert space consisting of all infinite real sequences (x_i) satisfying $\sum_{j=1}^\infty |x_j|^2 < \infty$. Write

$$e_j = (0, \dots, 0, 1, 0, \dots) \quad \text{for all } j = 1, 2, \dots,$$

that is, the j th component of e_j is 1 while all other components are 0. Let $I = \{0, 1, 1/2, \dots, 1/j, \dots\}$ and define $\{g_i\}_{i \in I}$ as follows:

$$g_i(x) = \begin{cases} 0, & i = 0, \\ \langle i e_{1/i}, x \rangle, & i \in I \setminus \{0\}. \end{cases}$$

Let $x = 0$. Then $I(x) = I$,

$$\partial g_i(x) = \begin{cases} \{0\}, & i = 0, \\ \{ie_{1/i}\}, & i \neq 0, \end{cases}$$

and hence,

$$D'(x) = \left\{ \sum_{j=1}^m \lambda_j \frac{1}{j} e_j : m > 0, \lambda_j \geq 0, \sum_{j=1}^m \lambda_j \leq 1 \right\}.$$

Clearly, $D'(x)$ is not closed.

3. The SC and the MFCQ. We continue to consider the system (1.1) relative to a closed convex set C in X . Notation is as in the introduction. A main aim of this section is to provide sufficient conditions ensuring that (1.1) satisfies the BCQ relative to C . These conditions are of two types (the SC and the MFCQ): one is expressed in terms of the functions g_i , while the other is expressed in terms of the directional derivatives of g_i . Let us first extend several known concepts to our present general case from the semi-infinite case (cf. [2, 12, 15] for the former type and [12, 17, 23] for the latter type). Recall that S and K are defined by (1.4) and (1.5).

DEFINITION 3.1. Let $x \in C \cap \text{bd } S$. An element $d \in X$ is called

(a) a linearized feasible direction of (1.1) at x if

$$(3.1) \quad \langle z^*, d \rangle \leq 0 \quad \text{for all } z^* \in \text{ext } \partial g_i(x) \text{ and for all } i \in I(x);$$

(b) a sequentially feasible direction of K at x if there exist a sequence (d_k) with $(d_k) \rightarrow d$ and a sequence (δ_k) of positive real numbers with $(\delta_k) \rightarrow 0$ such that $\{x + \delta_k d_k\} \subseteq K$.

Let $\text{LFD}(x)$ (resp., $\text{SFD}(x)$) denote the set of all d satisfying (a) (resp., (b)). Note that $\text{LFD}(x)$ and $\text{SFD}(x)$ are both closed convex cones.

DEFINITION 3.2. Let $K_S(x)$, $K_L(x)$ be, respectively, defined by

$$(3.2) \quad K_S(x) = (x + \text{SFD}(x)) \bigcap C$$

and

$$(3.3) \quad K_L(x) = (x + \text{LFD}(x)) \bigcap C.$$

Note that these two sets are closed convex sets.

PROPOSITION 3.3. Let $x \in C \cap \text{bd } S$. Then $\text{SFD}(x) \subseteq \text{LFD}(x)$, and hence

$$(3.4) \quad K \subseteq K_S(x) \subseteq K_L(x).$$

Proof. Since $K \subseteq K_S(x)$, (3.4) follows from the first assertion and the fact that $\text{LFD}(x)$ is closed convex. To prove the first assertion, let $d \in \text{SFD}(x)$, and let $\{d_k\}$, $\{\delta_k\}$ be as in Definition 3.1(b). In particular, for each $i \in I(x)$, one has $g_i(x + \delta_k d_k) \leq 0$ and hence, for each $z^* \in \partial g_i(x)$, we have

$$(3.5) \quad \langle z^*, \delta_k d_k \rangle \leq g_i(x + \delta_k d_k) - g_i(x) \leq 0$$

thanks to the fact that $G(x) = 0$ as $x \in \text{bd } S$. This implies that $\langle z^*, d_k \rangle \leq 0$. Letting $k \rightarrow +\infty$ we obtain $\langle z^*, d \rangle \leq 0$. Since $i \in I(x)$ and $z^* \in \partial g_i(x)$ are arbitrary, $d \in \text{LFD}(x)$, and the proof is complete. $\square$

DEFINITION 3.4. Let $x \in C \cap \text{bd } S$. We say that the system (1.1) satisfies

- (a) the MFCQ relative to C at x if the following conditions are satisfied:
- (1) $(x + \text{LFD}(x)) \cap \text{ri } C \neq \emptyset$;
 - (2) $\{d \in X : g'_i(x, d) < 0 \text{ for all } i \in I(x)\} \cap \text{span}(C - x) \neq \emptyset$;
- (b) the weak MFCQ relative to C at x if there exists a finite subset I_0 of $I(x)$ such that
- (1) $(x + \text{LFD}(x)) \cap \text{ri } C \neq \emptyset$;
 - (2) $\{d \in X : g'_i(x, d) < 0 \text{ for all } i \in I(x) \setminus I_0\} \cap \text{span}(C - x) \neq \emptyset$;
 - (3) g_i is affine for each $i \in I_0$.

We shall also say that x satisfies the MFCQ (resp., the weak MFCQ) for (1.1) relative to C when (a) (resp., (b)) occurs.

For the following definitions, we need additional notation. Let

$$(3.6) \quad \tilde{I}(x) = \{i \in I(x) : g_i(x) = 0\}.$$

Clearly, $\tilde{I}(x) = I(x)$ in the case when $x \in \text{bd } S$.

DEFINITION 3.5. The system (1.1) is said to satisfy

- (a) the SC relative to C if there exists $\bar{x} \in \text{ri } C$ such that $G(\bar{x}) < 0$;
- (b) the weak SC relative to C if there exists $\bar{x} \in \text{ri } C \cap S$ such that
 - (1) $\tilde{I}(\bar{x})$ is finite and g_i is affine for each $i \in \tilde{I}(\bar{x})$;
 - (2) G_0 is continuous and $G_0(\bar{x}) < 0$, where G_0 denotes the sup-function of $\{g_i : i \in I \setminus \tilde{I}(\bar{x})\}$;

$$(3.7) \quad G_0(z) = \sup_{i \in I \setminus \tilde{I}(\bar{x})} g_i(z), \quad z \in X.$$

A point $\bar{x}$ with the property in (a) (resp., (b)) is called a Slater (resp., weak Slater) point for the system (1.1) relative to C .

As the nomenclature suggests, it is clear that (a) $\implies$ (b) in Definitions 3.4 and 3.5.

THEOREM 3.6. Suppose there exists a weak Slater point $\bar{x}$ for the system (1.1) relative to C , and let $x \in C \cap \text{bd } S$. Then (1.1) satisfies the BCQ relative to C at x if (1.1) satisfies the PLV relative to C at x .

Proof. Let G_0 be defined as in Definition 3.5(b). Set

$$S_0 = \{y \in X : g_i(y) \leq 0, i \in I \setminus \tilde{I}(\bar{x})\}$$

and

$$S_1 = \{y \in X : g_i(y) \leq 0, i \in \tilde{I}(\bar{x})\}.$$

Then $G_0(\bar{x}) < 0$, $\tilde{I}(\bar{x})$ is finite, and g_i is affine for each $i \in \tilde{I}(\bar{x})$. Hence

$$N_{S_1}(x) = \begin{cases} \sum_{i \in I(x) \cap \tilde{I}(\bar{x})} \text{cone } \partial g_i(x) & \text{if } I(x) \cap \tilde{I}(\bar{x}) \neq \emptyset, \\ 0 & \text{if } I(x) \cap \tilde{I}(\bar{x}) = \emptyset \end{cases}$$

and

$$N_{S_0}(x) = \begin{cases} \text{cone } \partial G_0(x) & \text{if } G_0(x) = 0, \\ 0 & \text{if } G_0(x) < 0. \end{cases}$$

Note in particular that $N_{S_1}(x)$ and $N_{S_0}(x)$ are contained in $\text{cone } \partial G(x)$. Moreover, by [8, Proposition 2.3 and Theorem 5.1] (the proofs given there are valid for Banach

spaces, though the results were stated in a Hilbert space setting), $\{C, S_1\}$ and $\{C \cap S_1, S_0\}$ have the strong conical hull intersection property (the strong CHIP). Hence

$$\begin{aligned} N_K(x) &= N_{C \cap S_1}(x) + N_{S_0}(x) = N_C(x) + N_{S_1}(x) + N_{S_0}(x) \\ &\subseteq N_C(x) + \text{cone } \partial G(x) \subseteq N_C(x) + N'(x), \end{aligned}$$

where the last inclusion holds because of the PLV property. $\square$

For convenience of reference, we note the following immediate consequence of Theorem 2.5(iii) and Theorem 3.6.

COROLLARY 3.7. *Suppose that I is a compact space satisfying the first axiom of countability, and that the function: $i \mapsto g_i(x)$ is upper semicontinuous for each $x \in X$. Let $x \in K$. Suppose that $\dim C < \infty$ or $I(x)$ is finite and that there exists a weak Slater point for the system (1.1) relative to C . Then (1.1) satisfies the BCQ relative to C at x .*

Furthermore, if in addition $\{g_i : i \in I\}$ is equicontinuous on X , then the assumption on I about the first axiom of countability can be dropped.

THEOREM 3.8. *For the system (1.1) and relative to C , the following assertions hold:*

(i) *If the weak SC is satisfied, then at each point in $C \cap \text{bd } S$ the weak MFCQ is satisfied. If I is finite and if $x \in C \cap \text{bd } S$ satisfies the weak MFCQ, then there exist weak Slater points arbitrarily near x .*

(ii) *If the SC is satisfied then, at each point in $C \cap \text{bd } S$, the MFCQ is satisfied. Conversely, if $x \in C \cap \text{bd } S$ satisfies the MFCQ and if in addition x satisfies the PLV, then there exist Slater points arbitrarily near x (in particular, (1.1) satisfies the BCQ relative to C).*

Proof. (i) Let $\bar{x} \in \text{ri } C \cap S$ be a weak Slater point for the system (1.1) relative to C , and let $x \in C \cap \text{bd } S$ (hence $G(x) = 0$). Set $I_0 = I(x) \cap \tilde{I}(\bar{x})$. Then I_0 is finite (as $\tilde{I}(\bar{x})$), and x with this I_0 satisfies (1)–(3) of Definition 3.4(b). Indeed, (3) is evident. Suppose $I(x) \subseteq \tilde{I}(\bar{x})$. Then (2) of Definition 3.4(b) holds trivially. (1) also holds as $\bar{x} \in (x + \text{LFD}(x)) \cap \text{ri } C$ because, for each $i \in I_0 = I(x) \subseteq \tilde{I}(\bar{x})$, one has

$$g_i(x) = 0 \quad \text{and} \quad g_i(\bar{x}) = G(\bar{x}).$$

It follows from the convexity that $g'_i(x, \bar{x} - x) \leq 0$; that is, $\bar{x} - x \in \text{LFD}(x)$. We can therefore assume henceforth that $I(x) \not\subseteq \tilde{I}(\bar{x})$. Consequently $G_0(x) = G(x) = 0$, where G_0 is the function on X defined by (3.7). Noting $G_0(\bar{x}) < 0$, we have for each $i \in I(x) \setminus I_0$ that

$$g'_i(x, \bar{x} - x) \leq g_i(\bar{x}) - g_i(x) \leq G_0(\bar{x}) - G_0(x) < 0.$$

This implies that the intersection in (2) of Definition 3.4(b) is a nonempty set (containing $\bar{x} - x$) and that $\bar{x} - x \in \text{LFD}(x)$ as g_i is affine for each $i \in I_0$ (because $I_0 \subseteq \tilde{I}(\bar{x})$ and $\bar{x}$ is a weak Slater point). Therefore x is a weak MFCQ point relative to C . Conversely, we suppose that x satisfies the weak MFCQ relative to C and, in addition, that I is finite. Let $I_0 \subseteq I(x)$ satisfy (2) and (3) of Definition 3.4(b). Then by definition, x satisfies the MFCQ relative to C for the subsystem defined by

$$(3.8) \quad g_i(\cdot) \leq 0 \quad \text{for all } i \in I \setminus I_0.$$

Note that x satisfies the PLV for the subsystem (3.8) as I is finite. Hence, assuming the results in (ii), we can apply (ii) to conclude that there exists $\bar{x} \in \text{ri } C$ arbitrarily near x such that $\sup\{g_i(\bar{x}) : i \in I \setminus I_0\} < 0$. Clearly then $\bar{x}$ is a weak Slater point for (1.1) relative to C .

(ii) Let $\bar{x}, x$ be as in the first part of (i) but assume the stronger assumption that $\bar{x}$ is a Slater point relative to C . Then $\tilde{I}(\bar{x})$ must be empty and so must I_0 . In this case, (2) of Definition 3.4(b) coincides with (2) of Definition 3.4(a). Therefore x is an MFCQ point relative to C by (i). Conversely, suppose that x satisfies the MFCQ for the system (1.1) relative to C and, in addition, that x also satisfies the PLV. Then, by the former, there exist d_1, d_2 such that

$$x + d_1 \in (x + \text{LFD}(x)) \cap \text{ri } C, \quad d_2 \in \text{span}(C - x),$$

and

$$(3.9) \quad g'_i(x, d_2) < 0 \quad \text{for all } i \in I(x).$$

Also, by the PLV property,

$$(3.10) \quad \partial G(x) \subseteq N_C(x) + \text{conv} \{ \partial g_i(x) : i \in I(x) \}.$$

Take $t_0 > 0$ small enough such that $x + (1-t_0)d_1 + t_0d_2 \in \text{ri } C$. Set $d := (1-t_0)d_1 + t_0d_2$. Then, for any $i \in I(x)$,

$$(3.11) \quad g'_i(x, d) \leq (1-t_0)g'_i(x, d_1) + t_0g'_i(x, d_2) \leq t_0g'_i(x, d_2) < 0.$$

It follows that

$$(3.12) \quad G'(x, d) < 0.$$

Indeed, by (2.3) there exists $z^* \in \partial G(x)$ such that $G'(x, d) = \langle z^*, x \rangle$. By (3.10), there exist $t_1, \dots, t_m > 0$ and $z_1^*, \dots, z_m^* \in X^*$ with $\sum_{j=1}^m t_j = 1$, and $z_j^* \in \partial g_{i_j}(x)$ where $i_1, \dots, i_m \in I(x)$ such that $z^* - \sum_{j=1}^m t_j z_j^* \in N_C(x)$; in particular, $\langle z^* - \sum_{j=1}^m t_j z_j^*, d \rangle \leq 0$. It follows from (3.11) and (2.2) that

$$\langle z^*, d \rangle \leq \sum_{j=1}^m t_j g'_{i_j}(x, d) < 0;$$

this proves (3.12). Noting $G(x) = 0$ (as $x \in \text{bd } S$), (3.12) implies that $G(x + \bar{t}d) < 0$ for all sufficiently small $\bar{t} \in (0, 1)$. Clearly $x + \bar{t}d \in \text{ri } C$ (provided $\bar{t} < t_0$) and so $x + \bar{t}d$ is a Slater point for the system (1.1) relative to C . $\square$

Remark 3.1. Example 3.1(a) below will show that the finiteness assumption cannot be dropped in the second assertion of Theorem 3.8(i), and the condition of the PLV in the second assertion of Theorem 3.8(ii) cannot be dropped. Example 3.1(b) provides an example of (1.1) which satisfies the PLV and the weak MFCQ relative to C at some point $x_0 \in C \cap \text{bd } S$ but fails for the weak SC and the BCQ relative to C at x_0 . (In particular, the finiteness assumption in the second assertion of Theorem 3.8(i) cannot be replaced by the PLV.)

Example 3.1. (a) Let $X = C = \mathbb{R}$, $I = [-1, 0) \cup (0, +1]$, and let $\{g_i : i \in I\}$ be defined by

$$g_i(x) = \begin{cases} ix, & i \in [-1, 0) \\ 2ix - i^2, & i \in (0, +1] \end{cases} \quad \text{for all } x \in \mathbb{R}.$$

Then the corresponding sup-function is given by

$$(3.13) \quad G(x) = \begin{cases} 2x - 1, & x \geq 1 \\ x^2, & 0 \leq x \leq 1 \\ -x, & x \leq 0 \end{cases} \quad \text{for all } x \in \mathbb{R}.$$

Take $x_0 = 0$. Then, $S = \{0\}$, $I(x_0) = [-1, 0)$, and

$$(3.14) \quad \partial g_i(x_0) = \{i\} \quad \text{for all } i \in [-1, 0).$$

Moreover,

$$(3.15) \quad g'_i(x_0, 1) = i < 0 \quad \text{for all } i \in [-1, 0)$$

and

$$(3.16) \quad \text{LFD}(x_0) = [0, +\infty).$$

For the system (1.1) relative to C at the point x_0 , the following assertions hold:

- (1) The MFCQ is satisfied.
- (2) The PLV is not satisfied.
- (3) The weak SC is not satisfied.
- (4) The BCQ at x_0 is not satisfied.

Indeed, (1) is evident by (3.15) and (3.16). Next, by (3.14),

$$(3.17) \quad D'(x_0) = [-1, 0), \quad N'(x_0) = (-\infty, 0],$$

so that

$$\partial G(x_0) = [-1, 0] = \overline{D'(x_0)} \subseteq (-\infty, 0] = N'(x_0), \quad \partial G(x_0) \neq D'(x_0).$$

Hence, (2) holds. (3) is also evident because $G_0(\bar{x}) = 0$ whenever $\bar{x} \in S$ and G_0 is the sup-function of $\{g_i : i \in I \setminus I_0\}$ for some finite subset I_0 of I .

Finally, (4) holds because

$$N_K(x_0) = N_S(x_0) = \mathbb{R}, \quad N'(x_0) = (-\infty, 0].$$

(b) Let g_0 be defined by $g_0(x) = 0$ for all $x \in X$. We add this function g_0 to the system in (a) above. The new system satisfies the PLV relative to C at x_0 and the weak MFCQ relative to C at x_0 . However, the weak SC on C does not hold; hence the BCQ relative to C at x_0 is not satisfied.

It is well known (see, e.g., [16]) that the following implication is not true in general:

(*) Each finite subsystem of (1.1) satisfies the BCQ $\implies$ the system (1.1) does.

Therefore the remainder of this section is devoted to studying the following natural question: When does this implication hold? The following theorem (which will be useful for our discussion in section 5) provides a set of conditions ensuring the above implication (*). To facilitate the proof of this theorem, we recall the following proposition which was established independently by Deutsch [7] and Rubenstein [20] (see also [4]).

For a closed convex subset Z of X , let P_Z denote the projection operator defined by

$$P_Z(x) = \{y \in Z : \|x - y\| = d_Z(x)\},$$

where $d_Z(x)$ denotes the distance from x to Z . Recall that the duality map J from X to 2^{X^*} is defined by

$$J(x) := \{x^* \in X^* : \langle x^*, x \rangle = \|x\|^2, \|x^*\| = \|x\|\}.$$

In fact, $J(x) = \partial\phi(x)$, where $\phi(x) := \frac{1}{2}\|x\|^2$. Thus a Banach space X is smooth if and only if for each $x \in X$ the duality map is single-valued.

PROPOSITION 3.9. *Let Z be a closed convex set in X . Then for any $x \in X$, $z_0 \in P_Z(x)$ if and only if $z_0 \in Z$ and there exists $x^* \in J(x - z_0)$ such that $\langle x^*, z - z_0 \rangle \leq 0$ for any $z \in Z$, that is, $J(x - z_0) \cap N_Z(z_0) \neq \emptyset$. In particular, when X is smooth, $z_0 \in P_Z(x)$ if and only if $z_0 \in Z$ and $J(x - z_0) \in N_Z(z_0)$.*

THEOREM 3.10. *Let I be a compact metric space and let C be a closed convex subset of X such that $\dim C := l < +\infty$. Suppose that*

- (i) *for each $x \in X$, the function $i \mapsto g_i(x)$ is continuous;*
- (ii) *for any finite subset J of I , the subsystem*

$$(3.18) \quad g_j(x) \leq 0, \quad j \in J,$$

satisfies the BCQ relative to C ;

(iii) *for any finite subset J of I with $|J| \leq l$, the subsystem (3.18) satisfies the SC relative to C .*

Then (1.1) satisfies the BCQ relative to C .

Proof. Since I is compact and metrizable, there exists a sequence $\{I_k\}$ of subsets of I such that

- (a) each I_k is finite;
- (b) $I_k \subseteq I_{k+1}$ for $k = 1, 2, \dots$;
- (c) $\overline{\bigcup_{k=1}^{\infty} I_k} = I$.

Set $S_k = \bigcap_{i \in I_k} \{x \in X : g_i(x) \leq 0\}$ for each k . Then

$$(3.19) \quad K = C \cap S \subseteq C \cap S_k.$$

Let $x_0 \in K$ and $x^* \in N_K(x_0)$. We have to show that $x^* \in N_C(x_0) + N'(x_0)$. We will first show that there exist $\{x_k\} \subseteq C$ with $x_k \rightarrow x_0$ and $x_k^* \in N_{C \cap S_k}(x_k)$ such that $\{x_k^*\}$ is bounded and

$$(3.20) \quad \lim_{k \rightarrow \infty} \langle x_k^*, y \rangle = \langle x^*, y \rangle \quad \text{for each } y \in \text{span}(C - x_0).$$

In fact, since $Z := \text{span}(C - x_0)$ is of finite dimension, we may assume, without loss of generality, that the norm restricted on Z is both strictly convex and smooth. Clearly, we may assume that $x^*|_Z \neq 0$. Take $z_0 \in Z$ such that $\langle x^*, z_0 \rangle = \|x^*|_Z\| \cdot \|z_0\| = \|z_0\|^2$. Write $x = x_0 + z_0$. Then, $x^*|_Z = J(x - x_0)|_Z$ and, by Proposition 3.9, $x_0 = P_K(x)$. Let $x_k = P_{C \cap S_k}(x)$. Then

$$\|x_k\| \leq \|x_k - x\| + \|x\| \leq \|x - x_0\| + \|x\|;$$

hence $\{x_k\} \subset C$ is bounded. Without loss of generality, assume that $x_k \rightarrow \bar{x}$. Let $i \in I$ and $\epsilon > 0$. By (b) and (c), there exists $\{i_k\} \subseteq I$ with $i_k \in I_k$ for each k such that $i_k \rightarrow i$. By assumption (i), there exists an integer k_0 such that

$$(3.21) \quad g_i(\bar{x}) - g_{i_{k_0}}(\bar{x}) < \epsilon.$$

By (b), for each $k > k_0$, $i_{k_0} \in I_{k_0} \subseteq I_k$; hence, by the fact that $x_k \in S_k \subseteq S_{k_0}$,

$$(3.22) \quad g_{i_{k_0}}(\bar{x}) = g_{i_{k_0}}(\bar{x}) - g_{i_{k_0}}(x_k) + g_{i_{k_0}}(x_k) \leq g_{i_{k_0}}(\bar{x}) - g_{i_{k_0}}(x_k).$$

Consequently, by (3.21) and (3.22), we have that, for each $k > k_0$,

$$(3.23) \quad g_i(\bar{x}) < \epsilon + g_{i_{k_0}}(\bar{x}) - g_{i_{k_0}}(x_k).$$

Taking the limit as $k \rightarrow \infty$ we get $g_i(\bar{x}) \leq \epsilon$; hence $\bar{x} \in K$ as $\epsilon > 0$ and $i \in I$ are arbitrary. Because

$$\|x - \bar{x}\| = \lim_{k \rightarrow \infty} \|x - x_k\| \leq \|x - y\| \quad \text{for each } y \in K,$$

$\bar{x} = P_K(x) = x_0$. Let $x_k^* \in X^*$ with $x_k^*|_Z = J(x - x_k)|_Z$ and $\|x_k^*\| = \|x - x_k\|$. Then, by Proposition 3.9, $x_k^* \in N_{C \cap S_k}(x_k)$. By the smoothness, the mapping $x \mapsto J(x)|_Z$ is continuous and so (3.20) holds.

Now let G_k denote the sup-function corresponding to the subsystem (3.18) with $J = I_k$; that is,

$$G_k(x) := \sup_{i \in I_k} g_i(x) \quad \text{for each } x \in X.$$

If there exists a subsequence $\{k_j\}$ of $\{k\}$ such that $G_{k_j}(x_0) < 0$ for each j , then $N_{C \cap S_{k_j}}(x_{k_j}) = N_C(x_{k_j})$, and hence $x_{k_j}^* \in N_C(x_{k_j})$. This implies that $x^* \in N_C(x_{k_j})$ by (3.20) (and so $x^* \in N_C(x_0) + N'(x_0)$). Therefore we may assume that, for each k , $x_k^* \notin N_C(x_k)$, and that $G_k(x_k) = 0$. Then $I_k(x_k) := \{i \in I_k : g_i(x_k) = G_k(x_k)\} \neq \emptyset$. By assumption (ii), $x_k^* \in N_{C \cap S_k}(x_k) = N_C(x_k) + \text{cone}\{\partial g_i(x_k) : i \in I_k(x_k)\}$, and it follows from [19, Corollary 17.2.2] that there exist $z_k^* \in N_C(x_k)$, $i_j^k \in I_k(x_k)$, $y_{i_j^k}^* \in \partial g_{i_j^k}(x_k)$, and $\lambda_j^k \geq 0$ ($j = 1, 2, \dots, l$) such that

$$(3.24) \quad x_k^* = z_k^* + \sum_{j=1}^l \lambda_j^k y_{i_j^k}^* \quad \text{on } Z, \quad k = 1, 2, \dots$$

Let $\lambda^k = \sum_{j=1}^l \lambda_j^k$. Then $\{\lambda^k\}$ is bounded. Indeed, if not, by considering a subsequence if necessary, we have that $\lim_{k \rightarrow \infty} \lambda^k = +\infty$. Thus $\frac{x_k^*}{\lambda^k} \rightarrow 0$ as $k \rightarrow \infty$. Furthermore, without loss of generality, we may assume that, as $k \rightarrow \infty$,

$$(3.25) \quad i_j^k \rightarrow i_j \quad \text{and} \quad \frac{\lambda_j^k}{\lambda^k} \rightarrow \lambda_j, \quad j = 1, 2, \dots, l.$$

Note in particular that $\sum_{j=1}^l \lambda_j = 1$. Since $y_{i_j^k}^* \in \partial g_{i_j^k}(x_k) \subseteq \partial G(x_k)$ and since $x_k \rightarrow x_0$, $\{y_{i_j^k}^*\}$ is bounded. Consequently, by (3.24), $\{\frac{z_k^*}{\lambda^k}|_Z\}$ is bounded too. Thus, we may also assume that there exist $\tilde{z}_0^*$, $\tilde{y}_{i_j}^* \in Z^*$ such that

$$(3.26) \quad \frac{z_k^*}{\lambda^k} \rightarrow \tilde{z}_0^*; \quad y_{i_j^k}^* \rightarrow \tilde{y}_{i_j}^* \quad \text{on } Z$$

as $k \rightarrow \infty$. Then

$$(3.27) \quad \langle \tilde{z}_0^*, z - x_0 \rangle \leq 0 \quad \text{for each } z \in C$$

and, by assumption (i), (3.25), and (3.26), for each j ,

$$(3.28) \quad \langle \tilde{y}_{i_j}^*, z - x_0 \rangle \leq g_{i_j}(z) \leq g_{i_j}(z) - g_{i_j}(x_0) \quad \text{for each } z \in Z + x_0.$$

Hence by the Hahn–Banach theorem, $\tilde{z}_0^*$ and $\tilde{y}_{i_j}^*$ can be extended to $z_0^* \in N_C(x_0)$ and $y_{i_j}^* \in \partial g_{i_j}(x_0)$. By (3.24), (3.25), and (3.26),

$$(3.29) \quad 0 = z_0^* + \sum_{j=1}^l \lambda_j y_{i_j}^* \quad \text{on } Z.$$

By assumption (iii), take $\bar{y} \in C$ such that $g_{i_j}(\bar{y}) < 0$ for each $j = 1, 2, \dots, l$. Consequently, $\langle y_{i_j}^*, \bar{y} - x_0 \rangle < 0$ for each $j = 1, 2, \dots, l$, and hence,

$$\left\langle z_0^* + \sum_{j=1}^l \lambda_j y_{i_j}^*, \bar{y} - x_0 \right\rangle < 0,$$

which contradicts (3.29). Hence $\{\lambda^k\}$ is bounded. Thus, taking the limits on the two sides of (3.24) and using the similar arguments as above (if necessary, use subsequences), we get that

$$(3.30) \quad x^* = z_0^* + \sum_{j=1}^l \lambda_j y_{i_j}^* \quad \text{on } Z$$

for some $\lambda_j \geq 0$, $z_0^* \in N_C(x_0)$, and $y_{i_j}^* \in \partial g_{i_j}(x_0)$ ($j = 1, 2, \dots, l$). Note that, by (3.28), $y_{i_j}^* \in N_{C \cap g_{i_j}^{-1}(\mathbb{R}_-)}(x_0)$ for each $j = 1, 2, \dots, l$. Since $\langle y_{i_j}^*, \bar{y} - x_0 \rangle < 0$, it follows that $y_{i_j}^*|_Z \neq 0$, and hence $i_j \in I(x_0)$ for each $j = 1, 2, \dots, l$. Let $y^* = x^* - z_0^* - \sum_{j=1}^l \lambda_j y_{i_j}^*$. Then $y^* \in N_C(x_0)$ by (3.30), and so $x^* \in N_C(x_0) + N'(x_0)$. The proof is complete. $\square$

Remark 3.2. Example 3.2 below will show that the condition $\dim C < \infty$ cannot be dropped, while Examples 3.3 and 3.4 will show that conditions (i) and (iii) also cannot be dropped. In each of these examples, I is a compact subset of $\mathbb{R}$: $I = \{0, 1, \frac{1}{2}, \dots, \frac{1}{i}, \dots\}$.

Example 3.2. Let $C = X = \{x = (x_1, x_2, \dots, x_k, \dots) : \lim_k x_k \text{ exists}\}$ with the norm defined by

$$\|x\| = \sup_k |x_k|, \quad x = (x_k) \in X.$$

Define

$$g_i(x) = \begin{cases} \lim_k x_k, & i = 0, \\ x_{\frac{1}{i}}, & i \in I \setminus \{0\}, \end{cases} \quad x = (x_k) \in X.$$

Since $|g_i(x) - g_i(y)| \leq \|x - y\|$ for each $i \in I$, the corresponding sup-function G is continuous on X . Note that $G(\bar{x}) = -1 < 0$ for $\bar{x} = (-1) \in X$. Hence conditions (ii) and (iii) in Theorem 3.10 hold. clearly, condition (i) in Theorem 3.10 is satisfied. Let $x_0 = 0$. Then $x_0 \in C \cap S = S$ and $I(x_0) = I$. It is easy to see that $N'(x_0)$ is not closed; hence this system does not satisfy the BCQ at x_0 .

Example 3.3. Let $C = X = \mathbb{R}^2$. Define

$$g_i(x) = \begin{cases} x_1 + x_2, & i = 0, \\ x_1 + ix_2, & i \in I \setminus \{0\}, \end{cases} \quad x = (x_1, x_2) \in X.$$

Note that $|g_i(x) - g_i(y)| \leq 2\|x - y\|$ for each $i \in I$; thus the corresponding sup-function G is continuous on X . Let $x_0 = 0$. Then $K := C \cap S = S = \{(x_1, x_2) : x_1 \leq 0, x_1 + x_2 \leq 0\}$, $x_0 \in \text{bd} S$, and $I(x_0) = I$. Since $G((-1, \frac{1}{2})) = -\frac{1}{2} < 0$, conditions (ii) and (iii) in Theorem 3.10 hold. However, $N_K(x_0) = N_S(x_0) = \{(t_1, t_2) : 0 \leq t_2 \leq t_1\}$ and $N'(x_0) = \{(t_1, t_2) : 0 < t_2 \leq t_1\} \cup \{(0, 0)\}$. Therefore this system does not satisfy the BCQ at x_0 . Note that condition (i) is not satisfied.

Example 3.4. Let $C = X = \mathbb{R}^2$ and define

$$g_i(x) = \begin{cases} x_2, & i = 1, \\ ix_1 - x_2 - i^2, & i \in I \setminus \{1\}, \end{cases} \quad x = (x_1, x_2) \in X.$$

It is easy to verify that $|g_i(x) - g_i(y)| \leq 2\|x - y\|$ for each $i \in I$; thus the corresponding sup-function G is continuous on X . Let $x_0 = 0$. Then $I(x_0) = \{0, 1\}$ and $K = S = \{(x_1, 0) \in \mathbb{R}^2 : x_1 \leq 0\}$. Hence

$$N_K(x_0) = \{(t_1, t_2) \in \mathbb{R}^2 : t_1 \geq 0\}$$

and

$$N'(x_0) = \text{cone}\{(0, -1), (0, 1)\} = \{(t_1, t_2) \in \mathbb{R}^2 : t_1 = 0\}.$$

Consequently, this system does not satisfy the BCQ at x_0 . Note that conditions (i) and (ii) in Theorem 3.10 are satisfied but condition (iii) is not.

4. Optimality conditions for minimization problems. The main purpose of this section is to apply our results of the earlier sections to provide optimality conditions for optimization problems of the following type. For a real-valued function f on X , a typical problem is to

$$(4.1) \quad \begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in C \cap S, \end{array}$$

where C and S are as at the beginning of section 1. We will consider several kinds of functions on X . Let $\mathcal{A}(X)$ (resp., $\mathcal{C}(X)$) denote the class of all continuous real-valued affine (resp., convex) functions on X . Moreover, for $x_0 \in X$, let $\mathcal{G}_{x_0}(X)$ denote the class of real-valued functions f on X such that f is Gateaux differentiable (with the Gateaux derivative $Df(x_0)$) at x_0 and the restriction of f to C is convex. Recall that $K := C \cap S$.

THEOREM 4.1. *Let $x_0 \in K$. Then the following statements are equivalent:*

- (i) *The system (1.1) satisfies the BCQ relative to C at x_0 .*
- (ii) *For any $f \in \mathcal{G}_{x_0}(X)$, x_0 is an optimal solution of (4.1) if and only if*

$$(4.2) \quad Df(x_0) + x^* + \sum_{i \in I_0} \lambda_i x_i^* = 0 \quad \text{on } X$$

for some finite subset I_0 of $I(x_0)$, $x^* \in N_C(x_0)$, $\lambda_i \geq 0$, and $x_i^* \in \partial g_i(x_0)$ (for all $i \in I_0$).

- (ii*) *Same as (ii) but with (4.2) replaced by*

$$(4.3) \quad Df(x_0) + x^* + \sum_{i \in I_0} \lambda_i x_i^* = 0 \quad \text{on } \text{span } C.$$

- (iii) *For any $f \in \mathcal{C}(X)$, x_0 is an optimal solution of (4.1) if and only if*

$$(4.4) \quad y^* + x^* + \sum_{i \in I_0} \lambda_i x_i^* = 0 \quad \text{on } X$$

for some finite subset I_0 of $I(x_0)$, $y^* \in \partial f(x_0)$, $x^* \in N_C(x_0)$, $\lambda_i \geq 0$, and $x_i^* \in \partial g_i(x_0)$ (for all $i \in I_0$).

(iii*) Same as (ii*) but with (4.4) is replaced by

$$(4.5) \quad y^* + x^* + \sum_{i \in I_0} \lambda_i x_i^* = 0 \quad \text{on } \text{span } C.$$

(iv) Same as (iii) but with $f \in \mathcal{A}(X)$.

(iv*) Same as (iii*) but with $f \in \mathcal{A}(X)$.

Proof. Since each of the statements in the above list is true (with $I_0 = \emptyset$) when $x_0 \in C \cap \text{int } S$, we assume henceforth that $x_0 \in C \cap \text{bd } S$.

(i) $\implies$ (ii) and (ii*). Let $f \in \mathcal{G}_{x_0}(X)$ define

$$F(x) = \langle Df(x_0), x - x_0 \rangle + I_K(x), \quad x \in X.$$

Then, by [6, Corollary 1, p. 105]

$$(4.6) \quad \partial F(x_0) = Df(x_0) + N_K(x_0).$$

It follows from (i) that

$$(4.7) \quad \partial F(x_0) = Df(x_0) + N_C(x_0) + N'(x_0).$$

If x_0 is an optimal solution of (4.1), x_0 is a minimal point of F on X . In fact, it is clear that $F(x) \geq F(x_0)$ for any $x \in X \setminus K$. Moreover, if $x \in K$, then

$$(4.8) \quad F(x) = \langle Df(x_0), x - x_0 \rangle = \lim_{t \rightarrow 0+} \frac{f(x_0 + t(x - x_0)) - f(x_0)}{t} \geq 0 = F(x_0)$$

because $x_0 + t(x - x_0) \in K$ for each $0 < t < 1$ (and x_0 is an optimal solution of (4.1)). Hence, by [6, Proposition 2.4.11], $0 \in \partial F(x_0)$. Consequently, (4.7) entails (4.2) (and (4.3)) for appropriate x^*, λ_i, x_i^* as stated in (ii) and (ii*). Conversely, suppose that (4.3) holds for some $I_0, x^*, \lambda_i, x_i^*$ as stated in (ii*). Writing z^* for $x^* + \sum_{i \in I_0} \lambda_i x_i^*$, one has from (2.11) that $z^* \in N_K(x_0)$ and from (4.3) that

$$\langle Df(x_0), \cdot \rangle = \langle -z^*, \cdot \rangle \quad \text{on } \text{span } C.$$

Therefore, by convexity, we have for each $x \in K$ that

$$f(x) - f(x_0) \geq \langle Df(x_0), x - x_0 \rangle = \langle -z^*, x - x_0 \rangle \geq 0.$$

Thus the implications (i) $\implies$ (ii) and (ii*) are proved.

The implication (i) $\implies$ (iii) is a direct consequences of [6, Corollary, p. 52]. The necessity part of (iii*) follows from (iii) while the sufficiency part of (iii*) is proved similarly to the proof of the corresponding part of (ii*).

Further, since the implications (ii) $\implies$ (iv), (ii*) $\implies$ (iv*), (iii) $\implies$ (iv), and (iii*) $\implies$ (iv*) are trivial, it remains to show that (iv) $\implies$ (i) and (iv*) $\implies$ (i). To show these, let $y^* \in N_K(x_0)$. Then $-y^* \in \mathcal{A}(X)$ and x_0 is an optimal solution of (4.1) with $f := -y^*$. Thus $\partial f(x_0) = -y^*$ and, assuming (iv) or (iv*), there exist a finite subset I_0 of $I(x_0)$, $x^* \in N_C(x_0)$, $\lambda_i > 0$, and $x_i^* \in \partial g_i(x_0)$ ($i \in I_0$) such that

$$-y^* + x^* + \sum_{i \in I_0} \lambda_i x_i^* = 0 \quad \text{on } \text{span } C,$$

that is,

$$\langle y^*, \cdot \rangle = \langle z^*, \cdot \rangle \quad \text{on } \text{span } C,$$

where $z^* := x^* + \sum_{i \in I_0} \lambda_i x_i^*$. Since $x_0 \in C$, it follows that $y^* - z^* \in N_C(x_0)$. Consequently,

$$y^* \in N_C(x_0) + x^* + \sum_{i \in I_0} \lambda_i x_i^* \subseteq N_C(x_0) + N'(x_0),$$

and (i) holds. This completes the proof. $\square$

5. Best restricted range approximation problem in L^p spaces. Let $\Omega := T \cup A$, where A is a compact Hausdorff space and T is a measure space with a measure μ . As usual, let $C_R(A)$ denote the Banach space of all real-valued continuous functions on A equipped with the uniform norm denoted by $\|\cdot\|_A$.

Let X be the space consisting of all real functions x on Ω such that the restrictions $x|_T \in L_p(T, \mu)$ and $x|_A \in C_R(A)$, where $p \in [1, +\infty)$. Then X is a Banach space under the norm $\|\cdot\|$ defined by

$$\|x\| = \|x|_T\|_p + \|x|_A\|_A \quad \text{for each } x \in X.$$

For the so-called restricted range approximation problem in L^p spaces (studied by Levasseur and Levis in [10]), one is given the following data: Let l and u be a pair of real functions which are, respectively, upper and lower semicontinuous on A such that $l(t) \leq u(t)$ for each $t \in A$. Let L be a finite dimensional vector subspace of X generated by $h_1, h_2, \dots, h_n$. Let $z \in X$ be given. Then the problem to be considered is

$$(5.1) \quad \text{minimize} \quad \int_T |z(t) - x(t)|^p d\mu$$

subject to

$$(5.2) \quad x \in L, \quad l(t) \leq x(t) \leq u(t) \quad \text{for all } t \in A$$

(one can of course replace μ by μ' of the form $d\mu' = w d\mu$ for some positive $w \in L^\infty(T, \mu)$).

The aim of this section is to apply results of the preceding section to establish the following theorem, which was proved for the case when $p = 2$ (or p is an even integer) in [10] using a very different method and under assumptions H₁–H₆ given in the following:

H₁: A is a compact subset of $\mathbb{R}$ and T is a finite union of compact intervals of $\mathbb{R}$.

H₂: z is a real continuous function on Ω .

H₃: l and u are continuous on A with $l(t) < u(t)$ for all $t \in A$.

H₄: $\{h_1, h_2, \dots, h_n\}$ is linearly independent on T .

H₅: There exists $h \in X$ such that (5.2) is satisfied with $x = h$.

H₆: For each $k = 1, 2, \dots, n$, the set $\{h_1, h_2, \dots, h_k\}$ is a Chebyshev system of order k on A ; that is, if $t_1, t_2, \dots, t_k$ are distinct points in A , then the determinant $\det(h_j(t_i)) \neq 0$.

Instead of assumptions H₁–H₆, we consider the following two conditions:

D₁: There exists an element $h_0 \in L$ satisfying $l(t) < h_0(t) < u(t)$ for all $t \in A$.

D₂: A is metrizable (in addition to being compact Hausdorff), l and u are continuous on A , and, for any $k \leq n$ and any distinct points $t_1, t_2, \dots, t_k$ in A , there exists an element $\bar{h} \in L$ satisfying $l(t_i) < \bar{h}(t_i) < u(t_i)$ for all $i = 1, 2, \dots, k$.

We remark that $H_1 + H_6 \implies D_2$.

We use $Z(x)$ to denote the set of all points in T such that $x(t) = 0$. Note that if $x = \tilde{x}$ a.e. on T , then the “symmetric difference” $(Z(x) \setminus Z(\tilde{x})) \cup (Z(\tilde{x}) \setminus Z(x))$ is of μ -measure zero; hence $\int_{Z(x)} h(t) d\mu = \int_{Z(\tilde{x})} h(t) d\mu$ for each integrable function h on T . We shall need the following well-known results; see, for example, [3, Example 2.19] for a proof of Lemma 5.1. The other two lemmas follow immediately from Lemma 5.1 and the Riesz representation theorem.

LEMMA 5.1. *Let W be a normed vector space. Then, for any $w \in W \setminus \{0\}$, the subdifferential $\partial\|\cdot\|(w)$ of the norm function at w equals the set*

$$\{w^* \in W^* : \langle w^*, w \rangle = \|w\|\}.$$

LEMMA 5.2. *Let $W = L^1(T, \mu)$, where here and throughout, T denotes a measure space with measure μ . Let $w \in W \setminus \{0\}$. Then, as a subset of $L^\infty(T)$, $\partial\|\cdot\|_1(w)$ consists of all $\beta \in L^\infty(T)$ satisfying the properties that*

$$\begin{cases} |\beta(t)| \leq 1 & \text{a.e. on } T, \\ \beta(t) = \operatorname{sgn}[w(t)] & \text{a.e. on } T \setminus Z(w). \end{cases}$$

LEMMA 5.3. *Let $W = L^p(T, \mu)$ with $p \in (1, +\infty)$. Let $w \in W \setminus \{0\}$. Then, the norm function is Fréchet differentiable at w with the derivative $\|\cdot\|'_p(w)$, which is represented by $\xi \in L^q(T, \mu)$ ($\frac{1}{p} + \frac{1}{q} = 1$) with*

$$\xi(t) = \|w\|_p^{1-p} \cdot |w(t)|^{p-1} \cdot \operatorname{sgn}[w(t)] \quad \text{a.e. on } T.$$

In particular, $\partial\|\cdot\|_p(w) = \{\xi\}$.

Our main result in this section is now stated as follows.

THEOREM 5.4. *Suppose that D_1 or D_2 holds. Let $x_0 \in X$ satisfy (5.2) and let $z \in X$. Then x_0 is an optimal solution of the minimization problem (5.1) subject to (5.2) if and only if there exist a finite subset $\{t_1, t_2, \dots, t_k\}$ of A and a finite subset $\{\lambda_{t_1}, \lambda_{t_2}, \dots, \lambda_{t_k}\}$ of $\mathbb{R}$ ($0 \leq k \leq n$) with the following properties:*

- (a) $x_0(t_i) = u(t_i)$ or $x_0(t_i) = l(t_i)$ for each $i = 1, 2, \dots, k$;
- (b) for each $i = 1, 2, \dots, k$,

$$\operatorname{sgn} \lambda_{t_i} = \begin{cases} 1 & \text{if } x_0(t_i) = u(t_i), \\ -1 & \text{if } x_0(t_i) = l(t_i); \end{cases}$$

- (c) for each $j = 1, 2, \dots, n$,

$$(5.3) \quad \int_T |z(t) - x_0(t)|^{p-1} \operatorname{sgn}[z(t) - x_0(t)] h_j(t) d\mu = \sum_{i=1}^k \lambda_{t_i} h_j(t_i), \quad p > 1;$$

$$(5.4) \quad \int_T \operatorname{sgn}[z(t) - x_0(t)] h_j(t) d\mu + \int_{Z(z-x_0)} \beta(t) h_j(t) d\mu = \sum_{i=1}^k \lambda_{t_i} h_j(t_i), \quad p = 1,$$

for some $\beta \in L^\infty$ with $|\beta(t)| \leq 1$ a.e. on $Z(z - x_0)$. (Note: we adopt the convention that, if $k = 0$, the sums of the right-hand sides of (5.3) and (5.4) are read as zero.)

In order to prove this theorem, we first do some preparation. We make two homeomorphic copies of A : one is still denoted by A while the other is denoted by $\check{A} = \{\check{a} : a \in A\}$, where $a \mapsto \check{a}$ is a bijection. Let $I = A \cup \check{A}$. Thus I can be

topologized such that $A, \check{A}$ are two disjoint compact subsets of I . For each $t \in A$, define $g_t : X \rightarrow \mathbb{R}$ by

$$(5.5) \quad g_t(x) = e_t(x) - u(t), \quad x \in X,$$

where $e_t : X \rightarrow \mathbb{R}$ is the “unit point functional” at t defined by

$$e_t(x) = x(t), \quad x \in X.$$

Similarly, for each $\check{t} \in \check{A}$, define $g_{\check{t}} : X \rightarrow \mathbb{R}$ by

$$(5.6) \quad g_{\check{t}}(x) = l(\check{t}) - e_t(x), \quad x \in X.$$

Thus $\{g_i : i \in I\}$ is a family of affine functions on X such that, for each $i \in I$,

$$(5.7) \quad |g_i(x) - g_i(y)| \leq \|x - y\|, \quad x, y \in X.$$

In particular, $\{g_i : i \in I\}$ is equicontinuous on X , and the corresponding sup-function $x \mapsto G(x) := \sup\{g_i(x) : i \in I\}$ is continuous. Note furthermore that the function $i \mapsto g_i(x)$ is upper semicontinuous for each $x \in X$. Clearly, $x \in X$ satisfies (5.2) if and only if $x \in L$ and $g_i(x) \leq 0$ for each $i \in I$. Assuming D_1 , it is easy to verify that $h_0 \in L$ given in D_1 satisfies the strict inequality $G(h_0) < 0$ (thanks to the compactness of A); thus h_0 is a Slater point for the system $g_i(\cdot) \leq 0, i \in I$. Thus by Corollary 3.7, this system satisfies the BCQ relative to L at each point x in $L \cap S$, where $S := \{x \in X : g_i(x) \leq 0, i \in I\}$ (the assertion being trivial if $x \in L \cap \text{int } S$). Similarly, if D_2 holds, then it follows from Theorem 3.10 that this system satisfies the BCQ relative to L at each point x in $L \cap S$.

Note that if $x \in \text{bd } S$, then $G(x) = 0$, and $I(x)$ defined in (1.7) is equal to the set $\tilde{I}(x) = \{i \in I : g_i(x) = G(x) = 0\}$ defined in (3.6); hence

$$(5.8) \quad I(x) = \{t \in A : x(t) = u(t)\} \cup \{\check{t} \in \check{A} : x(t) = l(\check{t})\}$$

whenever $x \in \text{bd } S$. Note that, since $l(\cdot) < u(\cdot)$ on A , $t \in I(x)$ implies $\check{t} \notin I(x)$; similarly, $\check{t} \in I(x)$ implies $t \notin I(x)$. For easy reference, we record the following trivial fact:

$$(5.9) \quad g'_i(x) = \begin{cases} e_t, & i = t \in A, \\ -e_t, & i = \check{t} \in \check{A}. \end{cases}$$

Now define $f : X \rightarrow \mathbb{R}$ by

$$(5.10) \quad f(x) = \left(\int_T |z(t) - x(t)|^p d\mu \right)^{\frac{1}{p}} \quad \text{for each } x \in X.$$

Then f is a continuous convex function on X , and x_0 is a solution of the minimization problem (5.1) subject to (5.2) if and only if it is a solution of the problem of minimizing f subject to

$$(5.11) \quad x \in L, \quad g_i(x) \leq 0 \quad \text{for each } i \in I,$$

where $\{g_i : i \in I\}$ is defined by (5.5) and (5.6).

Since $f(x)$ defined in (5.10) depends only on the restriction of x to T , the following result follows from Lemmas 5.2 and 5.3 together with the chain rule.

PROPOSITION 5.5. *Suppose that $(z - x_0)|_T$ is not the zero-element of $L^p(T, \mu)$. Then the following assertions hold:*

(a) *The case when $p = 1$. A functional $y^* \in \partial f(x_0)$ if and only if there exists $\beta \in L^\infty(T, \mu)$ such that*

$$(5.12) \quad |\beta(t)| \leq 1 \quad \text{a.e. on } T,$$

$$(5.13) \quad \beta(t) = \text{sgn}[z(t) - x_0(t)] \quad \text{a.e. on } T \setminus Z(z - x),$$

and

$$(5.14) \quad \langle y^*, x \rangle = - \int_T x(t) \beta(t) d\mu \quad \text{for each } x \in X.$$

(b) *The case when $p \in (1, +\infty)$. A functional $y^* \in \partial f(x_0)$ if and only if*

$$(5.15) \quad \langle y^*, x \rangle = -c \int_T x(t) |z(t) - x_0(t)|^{p-1} \text{sgn}[z(t) - x_0(t)] d\mu \quad \text{for each } x \in X,$$

where $c := \|(z - x_0)|_T\|_p^{1-p}$.

Proof of Theorem 5.4. Let f, g_i, I , and S be the same as those introduced after the statement of Theorem 5.4. Since the proof for the case when $x_0 \in L \cap \text{int } S$ is easier (in this case one can take $k = 0$, and the right-hand sides of (5.3) and (5.4) are replaced by zero), it suffices to consider the case when $x_0 \in L \cap \text{bd } S$. As already noted, by assumption D₁ or D₂, the system of inequalities $g_i(\cdot) \leq 0, i \in I$, satisfies the BCQ at x_0 relative to L . Noting that $x^* = 0$ on L whenever $x^* \in N_L(x_0)$, it follows from Theorem 4.1 that the following statements are equivalent:

(i) x_0 is a minimizer of f subject to (5.11).

(ii) There exist a finite subset I_0 of $I(x_0)$, $y^* \in \partial f(x_0)$, $\mu_i > 0$, and $x_i^* \in \partial g_i(x_0)$ (for all $i \in I_0$) such that

$$(5.16) \quad y^* + \sum_{i \in I_0} \mu_i x_i^* = 0 \quad \text{on } L.$$

(ii*) Same as (ii) but with (5.16) replaced by

$$(5.17) \quad cy^* + \sum_{i \in I_0} \mu_i x_i^* = 0 \quad \text{on } L,$$

where c is a positive constant.

In view of (5.8) and (5.9), we note that, in (ii) and (ii*), each x_i^* can be expressed as either e_t or $-e_t$:

$$x_i^* = \begin{cases} e_t & \text{if } i = t \in I(x_0) \cap A \text{ (i.e., if } x_0(t) = u(t)), \\ -e_t & \text{if } i = \tilde{t} \in I(x_0) \cap \tilde{A} \text{ (i.e., if } x_0(t) = l(t)). \end{cases}$$

Thus $\mu_i x_i^*$ can be expressed in the form $\lambda_{t_i} e_{t_i}$ for some $\lambda_{t_i} \in \mathbb{R}$ and some $t_i \in A$ satisfying $x_0(t_i) = u(t_i)$ or $x_0(t_i) = l(t_i)$. Explicitly, λ_{t_i} is defined by

$$\lambda_{t_i} = \begin{cases} \mu_{t_i} & \text{if } x_0(t_i) = u(t_i) \text{ and } t_i \in I_0, \\ -\mu_{t_i} & \text{if } x_0(t_i) = l(t_i) \text{ and } \tilde{t}_i \in I_0. \end{cases}$$

Moreover, when (ii) (or (ii*)) holds, we can suppose I_0 satisfies an additional property that $|I_0|$, the number of elements of I_0 , is not greater than $\dim L$, the dimension of

L . To see this, we suppose, without loss of generality, that $h_1, h_2, \dots, h_n$ are linearly independent, so $\dim L = n$. Let $\mathcal{V}$ denote the subset of $\mathbb{R}^n$ defined by

$$\{(\langle x_i^*, h_1 \rangle, \langle x_i^*, h_2 \rangle, \dots, \langle x_i^*, h_n \rangle) : i \in I_0\}$$

and $\mathcal{U} := \mathcal{V} \cup \{(\langle y^*, h_1 \rangle, \langle y^*, h_2 \rangle, \dots, \langle y^*, h_n \rangle)\}$. Since L is spanned by $h_1, h_2, \dots, h_n$, (ii) implies that $0 \in \text{co}\mathcal{U}$. Note that

$$(5.18) \quad \langle x_i^*, h_0 - x_0 \rangle < 0 \quad \text{for each } i \in I_0.$$

Indeed, if $i = \check{t} \in I_0$ with some $t \in A$, then $x_0(t) = l(t)$ and

$$\langle x_i^*, h_0 \rangle = \langle -e_t, h_0 \rangle = -h_0(t) < -l(t) = \langle -e_t, x_0 \rangle = \langle x_i^*, x_0 \rangle;$$

thus (5.18) is true, as the case when $i = t \in I_0$ with some $t \in A$ can be considered similarly. Since $h_0 - x_0 \in L$ and L is spanned by $h_1, h_2, \dots, h_n$, it follows from (5.18) that $0 \notin \text{co}\mathcal{V}$; thus $0 \in \text{co}\mathcal{U} \setminus \text{co}\mathcal{V}$.

By the Carathéodory theorem (cf. [5]), there exist $\bar{\mu}_0 > 0$, $\bar{\mu}_{i_j} \geq 0$ ($j = 1, 2, \dots, n$) with $\bar{\mu}_0 + \sum_{j=1}^n \bar{\mu}_{i_j} = 1$ and $i_1, i_2, \dots, i_n \in I_0$ such that

$$\bar{\mu}_0 y^* + \sum_{j=1}^n \bar{\mu}_{i_j} x_{i_j}^* = 0 \quad \text{at } h_1, h_2, \dots, h_n.$$

Therefore, I_0 in (ii) can be replaced by its subset with at most n elements. Making use of Proposition 5.5 and noting that L is spanned by $h_1, h_2, \dots, h_n$, we now see that (ii) (or (ii*)) holds if and only if (a), (b), and (c) of Theorem 5.4 hold for some finite subsets $\{t_1, t_2, \dots, t_k\} \subseteq A$ and $\{\lambda_{t_1}, \lambda_{t_2}, \dots, \lambda_{t_k}\} \subseteq \mathbb{R}$, where $k \leq n$. Finally, as we already noted, (i) holds if and only if x_0 is an optimal solution of the minimization problem (5.1) subject to (5.2). This completes the proof of Theorem 5.4. $\square$

6. Approximation problem with constraints by conditionally positive semidefinite functions. This problem is studied in a recent paper [22] by Strauss. Following his approach, let Q be a nonempty subset of a Euclidean space $\mathbb{R}^d$ and let $F(Q)$ denote the class of all real-valued functions on Q . Let H be a vector subspace of $F(Q)$ and suppose that H is equipped with a seminorm induced by a semi-inner product $\langle \cdot, \cdot \rangle$ on $H \times H$. Let $U = \text{span}\{u_1, u_2, \dots, u_m\}$ be an m -dimensional subspace of H , and let K be a reproducing kernel with respect to U (for undefined terms, see [1, 18]). Let $A = \{z_1, z_2, \dots, z_n\} \subseteq Q$ and let C_A denote the $(m \times n)$ -matrix defined by

$$C_A = (u_j(z_i))_{j=1, i=1}^{m, n}.$$

Let W_n, W_m , respectively, denote an $(n \times n)$ -matrix and an $(m \times m)$ -matrix; we assume that W_n and W_m are positive semidefinite. Define a seminorm $\|\cdot\|$ on $\mathbb{R}^n \times \mathbb{R}^m$ by

$$\|(\mathbf{a}, \mathbf{b})\|_W^2 = \mathbf{a}^T W_n \mathbf{a} + \mathbf{b}^T W_m \mathbf{b} \quad \text{for all } \mathbf{a} \in \mathbb{R}^n, \mathbf{b} \in \mathbb{R}^m.$$

Suppose that we are given the following data: $h \in H$, $\{\xi_i : i \in I\} \subseteq \mathbb{R}$, $\{p_i : i \in I\} \subseteq \mathbb{R}^n$, and $\{q_i : i \in I\} \subseteq \mathbb{R}^m$, where I is an index set. Let

$$Z_I = \{(\mathbf{a}, \mathbf{b}) : \mathbf{a} \in \mathbb{R}^n, \mathbf{b} \in \mathbb{R}^m, C_A \mathbf{a} = 0, p_i^T \mathbf{a} + q_i^T \mathbf{b} \leq \xi_i, i \in I\}$$

and let $\hat{d} : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ be defined by

$$\hat{d}(\mathbf{a}, \mathbf{b}) = \frac{1}{2} \left(\left\| h - \sum_{i=1}^n a_i K(\cdot, z_i) - \sum_{j=1}^m b_j u_j \right\|^2 + \|(\mathbf{a}, \mathbf{b})\|_W^2 \right),$$

where $\mathbf{a} = (a_1, a_2, \dots, a_n) \in \mathbb{R}^n$ and $\mathbf{b} = (b_1, b_2, \dots, b_m) \in \mathbb{R}^m$. Under the obvious modifications, Z_I and $\hat{d}$ can be defined when $m = 0$, and our discussions below are valid for this case as well as for the case when $m \neq 0$.

The problem to be considered here is

$$(6.1) \quad \text{minimize } \hat{d}(\mathbf{a}, \mathbf{b}) \quad \text{subject to } (\mathbf{a}, \mathbf{b}) \in Z_I.$$

Define

$$(6.2) \quad U_A^\perp = \{(\mathbf{a}, \mathbf{b}) \in \mathbb{R}^n \times \mathbb{R}^m : C_A \mathbf{a} = 0\},$$

$$(6.3) \quad K_A = (K(z_i, z_j))_{i,j=1}^n,$$

$$(6.4) \quad \mathbf{h}_A = (h(z_1), \dots, h(z_n))^T,$$

and

$$(6.5) \quad \hat{e}(\mathbf{a}, \mathbf{b}) = \frac{1}{2} (\mathbf{a}^T (K_A + W_n) \mathbf{a} + \mathbf{b}^T W_m \mathbf{b}) - \mathbf{h}_A^T \mathbf{a} + \frac{1}{2} \langle h, h \rangle.$$

By Proposition 2.3 and the proof of Proposition 4.1 in [22], we have

$$(6.6) \quad \hat{d}(\mathbf{a}, \mathbf{b}) = \hat{e}(\mathbf{a}, \mathbf{b}) \quad \text{for all } (\mathbf{a}, \mathbf{b}) \in U_A^\perp$$

and

$$(6.7) \quad K_A \text{ is positive semidefinite with respect to } U_A^T$$

in the sense that $\mathbf{a}^T K_A \mathbf{a} \geq 0$ whenever $(\mathbf{a}, \mathbf{b}) \in U_A^T$. This implies of course that $K_A + W_n$ is also positive semidefinite, and so the quadratic function $\hat{e}$ is convex on U_A^T .

We always assume that (6.1) has a solution: let $(\mathbf{a}_0, \mathbf{b}_0)$ denote a solution of (6.1). Let $r(C_A)$ denote the rank of the matrix C_A . Finally, $|J|$ denotes the number of elements of a finite set J . The following result was asserted as a theorem ([22, Theorem 6.2]):

Suppose that $(\mathbf{a}_0, \mathbf{b}_0)$ is a solution of (6.1), $r(C_A) = m$, and that there exists $(\bar{\mathbf{a}}, \bar{\mathbf{b}}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that

$$(6.8) \quad C_A \bar{\mathbf{a}} = 0, \quad \mathbf{p}_i^T \bar{\mathbf{a}} + \mathbf{q}_i^T \bar{\mathbf{b}} < \xi_i \quad \text{for all } i \in I.$$

Then there exists a finite subset I_0 of I with $|I_0| \leq n$ such that $(\mathbf{a}_0, \mathbf{b}_0)$ also minimizes $\hat{d}(\mathbf{a}, \mathbf{b})$ subject to $(\mathbf{a}, \mathbf{b}) \in Z_{I_0}$.

While we will give a counterexample below to show this result is false, we put forward the following corrected version.

THEOREM 6.1. *Let $(\mathbf{a}_0, \mathbf{b}_0)$ be a solution of (6.1). Assume that the system of linear inequalities*

$$(6.9) \quad \mathbf{p}_i^T \mathbf{a} + \mathbf{q}_i^T \mathbf{b} \leq \xi_i, \quad i \in I,$$

satisfies the BCQ relative to $U_A^\perp$ at $(\mathbf{a}_0, \mathbf{b}_0)$. Then there exists a finite subset I_0 of I with $|I_0| \leq m + n - r(C_A)$ such that $(\mathbf{a}_0, \mathbf{b}_0)$ also minimizes $\hat{d}(\mathbf{a}, \mathbf{b})$ subject to $(\mathbf{a}, \mathbf{b}) \in Z_{I_0}$.

Proof. By assumption and (6.6), $(\mathbf{a}_0, \mathbf{b}_0)$ is a solution of the following minimization problem:

$$\text{Minimize } \hat{e}(\mathbf{a}, \mathbf{b}) \quad \text{subject to } (\mathbf{a}, \mathbf{b}) \in U_A^\perp \text{ and (6.9).}$$

We shall use $D\hat{e}(\mathbf{a}_0, \mathbf{b}_0)|_{U_A^\perp}$ to denote the restriction to $U_A^\perp$ of the Gateaux (equivalently, Fréchet) derivative of the function $\hat{e}$ at $(\mathbf{a}_0, \mathbf{b}_0)$. Since, by assumption, the system (6.9) in $X := \mathbb{R}^n \times \mathbb{R}^m$ satisfies the BCQ relative to $U_A^\perp$ at $(\mathbf{a}_0, \mathbf{b}_0)$, one can apply the implication (i) $\implies$ (ii*) of Theorem 4.1 to conclude that there exist a finite subset I'_0 of $I(\mathbf{a}_0, \mathbf{b}_0)$ and a collection $\{\lambda'_i : i \in I'_0\}$ of nonnegative real numbers such that

$$(6.10) \quad D\hat{e}(\mathbf{a}_0, \mathbf{b}_0)|_{U_A^\perp} + \sum_{i \in I'_0} \lambda'_i (\mathbf{p}_i^T, \mathbf{q}_i^T)|_{U_A^\perp} = 0,$$

because $x^*|_{U_A^\perp} = 0$ whenever $x^* \in N_{U_A^\perp}(\mathbf{a}_0, \mathbf{b}_0)$. We write μ for the sum $\sum_{i \in I'_0} \lambda'_i$ and assume that $\mu \neq 0$ (otherwise, $(\mathbf{a}_0, \mathbf{b}_0)$ must be a minimal point of $\hat{e}(=\hat{d})$ on $U_A^\perp$ by convexity of $\hat{e}$ on $U_A^\perp$). Thus

$$(6.11) \quad -D\hat{e}(\mathbf{a}_0, \mathbf{b}_0)|_{U_A^\perp} \in \text{cone}\{(\mathbf{p}_i^T, \mathbf{q}_i^T)|_{U_A^\perp} : i \in I'_0\}.$$

Since $\dim U_A^\perp = n + m - r(C_A)$, it follows from [19, Corollary 17.1.2] that I'_0 in (6.11) can be replaced by a subset I_0 with $|I_0| \leq n + m - r(C_A)$. Note that (6.10) continues to hold when I'_0 is replaced by I_0 (replace the family $\{\lambda'_i\}$ by a suitable new one if necessary). Consequently, by (ii*) of Theorem 4.1, $(\mathbf{a}_0, \mathbf{b}_0)$ minimizes $\hat{e}(\mathbf{a}, \mathbf{b})$ on $U_A^\perp$ subject to the system of linear inequalities

$$\mathbf{p}_i^T \mathbf{a} + \mathbf{q}_i^T \mathbf{b} \leq \xi_i, \quad i \in I_0.$$

This completes the proof as before. $\square$

Example 6.1. Let $A = Q$ consist of two distinct elements z_1, z_2 . We identify $F(Q)$ with $\mathbb{R}^2$ in the obvious manner. Let $H = F(Q)$ with the usual inner product. Take $h = (1, 0)^T \in H$, $m = 0$, $W_m = 0$, $U = \{0\}$, $C_A = 0$, and $U_A^\perp = \mathbb{R}^2$. Define W_n by $W_n = (\frac{1}{4}, \frac{0}{4})$ and fix the reproducing kernel $K : Q \mapsto \mathbb{R}$ defined by

$$K(x, y) = \begin{cases} 1 & \text{if } x = y, \\ 0 & \text{if } x \neq y. \end{cases}$$

Take $I = \{1, 2, \dots\}$, $\xi_i = 0$, and $\mathbf{p}_i = (1, \frac{1}{i})$ for $i \in I$. Then, (6.8) is satisfied for appropriate $\bar{\mathbf{a}}$. Note

$$\hat{d}(\mathbf{a}) := \hat{d}(\mathbf{a}, \mathbf{b}) = \frac{1}{2} \left[\|h - \mathbf{a}\|^2 + \frac{1}{4} \|\mathbf{a}\|^2 \right]$$

and

$$Z_I = \left\{ \mathbf{a} \in \mathbb{R}^2 : a_1 + \frac{1}{i} a_2 \leq 0 \text{ for all } i \in I \right\}.$$

Clearly, Z_I is the convex set of $\mathbb{R}^2$ containing the 4th quadrant and bounded by the two half-lines, respectively, contained in $a_1 + a_2 = 0$ and $a_1 = 0$. It follows easily that the origin $0 \in \mathbb{R}^2$ is the nearest point to h from Z_I . On the other hand, if I_0 is a finite subset of I , then Z_{I_0} is also bounded by the two half-lines, respectively, contained in $a_1 + \frac{1}{i_-} a_2 = 0$ and $a_1 + \frac{1}{i_+} a_2 = 0$, where $i_- := \min I_0$, $i_+ := \max I_0$. Let $\mathbf{a}_{I_0}$ denote the nearest point to h from Z_{I_0} . Then $\mathbf{a}_{I_0} = \frac{1}{1+i_+^2} (1, -i_+)^T$ and $\|h\|^2 = \|\mathbf{a}_{I_0}\|^2 + \|h - \mathbf{a}_{I_0}\|^2$. This implies that

$$\hat{d}(0) = \frac{1}{2} \|h\|^2 > \frac{1}{2} \left[\frac{1}{4} \|\mathbf{a}_{I_0}\|^2 + \|h - \mathbf{a}_{I_0}\|^2 \right] = \hat{d}(\mathbf{a}_{I_0}),$$

while

$$\hat{d}(0) = \frac{1}{2} \|h - 0\|^2 \leq \frac{1}{2} \|h - \mathbf{a}\|^2 \leq \hat{d}(\mathbf{a}) \quad \text{for all } \mathbf{a} \in Z_I.$$

Therefore [22, Theorem 6.2] mentioned earlier is not true.

Remark 6.1. Let l_i be the function on $\mathbb{R}^n \times \mathbb{R}^m$ defined by

$$l_i(\mathbf{a}, \mathbf{b}) = \mathbf{p}_i^T \mathbf{a}_i + \mathbf{q}_i^T \mathbf{b} - \xi_i, \quad (\mathbf{a}, \mathbf{b}) \in \mathbb{R}^n \times \mathbb{R}^m.$$

Suppose that I can be topologized in such a way that it is a compact space satisfying the first axiom of countability and that $i \mapsto l_i(\mathbf{a}, \mathbf{b})$ is upper semicontinuous on I for each $(\mathbf{a}, \mathbf{b}) \in U_A^\perp$. Then, by Corollary 3.7, the linear inequality system (6.9) satisfies the BCQ relative to $U_A^\perp$ at $(\mathbf{a}_0, \mathbf{b}_0)$ if (6.8) is satisfied. In this particular case, the conclusion of [22, Theorem 6.2] mentioned earlier does hold. Similarly, if I can be topologized in such a way that it is a compact metric space and that $i \mapsto l_i(\mathbf{a}, \mathbf{b})$ is continuous on $i \in I$ for each $(\mathbf{a}, \mathbf{b}) \in U_A^\perp$, then, by Theorem 3.10, the aforesaid result is also valid even if (6.8) is replaced by a weaker condition given in the following: for each subset J of I with $|J| \leq m + n - r(C_A)$, there exists $(\bar{\mathbf{a}}, \bar{\mathbf{b}}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that

$$(6.12) \quad C_A \bar{\mathbf{a}} = 0, \quad \mathbf{p}_i^T \bar{\mathbf{a}} + \mathbf{q}_i^T \bar{\mathbf{b}} < \xi_i \quad \text{for all } i \in J.$$

Acknowledgments. We wish to thank two referees for their valuable comments and suggestions.

REFERENCES

- [1] M. ATTEIA, *Hilbertian Kernels and Spline Functions*, North-Holland, Amsterdam, 1992.
- [2] H. BAUSCHKE, J. BORWEIN, AND W. LI, *Strong conical hull intersection property, bounded linear regularity, Jameson's property (G), and error bounds in convex optimization*, Math. Program. Ser. A, 86 (1999), pp. 135–160.
- [3] J. F. BONNANS AND A. SHAPIRO, *Perturbation Analysis of Optimization Problems*, Springer-Verlag New York, 2000.
- [4] D. BRAESS, *Nonlinear Approximation Theory*, Springer-Verlag, New York, 1986.
- [5] E. W. CHENEY, *Introduction to Approximation Theory*, McGraw-Hill, New York, 1966.
- [6] F. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.
- [7] F. DEUTSCH, *Some Applications of Functional Analysis to Approximation Theory*, Doctoral dissertation, Brown University, Providence, RI, 1965.

- [8] F. DEUTSCH, W. LI, AND J. D. WARD, *Best approximation from the intersection of a closed convex set and a polyhedron in Hilbert space, weak Slater conditions, and the strong conical hull intersection property*, SIAM J. Optim., 10 (1999), pp. 252–268.
- [9] J.-B. HIRIART-URRUTY AND C. LEMARÉCHAL, *Convex Analysis and Minimization Algorithms. I. Fundamentals* Grundlehren der Mathematischen Wissenschaften 305, Springer-Verlag, Berlin, 1993.
- [10] K. M. LEVASSEUR AND J. T. LEVIS, *Least-squares approximation with restricted range*, J. Approx. Theory, 36 (1982), pp. 304–316.
- [11] C. LI, *On best uniform restricted range approximation in complex-valued continuous function spaces*, J. Approx. Theory, 120 (2003), pp. 71–84.
- [12] C. LI AND X.-Q. JIN, *Nonlinearly constrained best approximation in Hilbert spaces: The strong CHIP and the basic constraint qualification*, SIAM J. Optim., 13 (2002), pp. 228–239.
- [13] C. LI AND K. F. NG, *On best approximation by nonconvex sets and perturbation of nonconvex inequality systems in Hilbert spaces*, SIAM J. Optim., 13 (2002), pp. 726–744.
- [14] C. LI AND K. F. NG, *Constraint qualification, the strong CHIP, and best approximation with convex constraints in Banach spaces*, SIAM J. Optim., 14 (2003), pp. 584–607.
- [15] W. LI, *Abadie's constraint qualification, metric regularity, and error bounds for differentiable convex inequalities*, SIAM J. Optim., 7 (1997), pp. 966–978.
- [16] W. LI, C. NAHAK, AND I. SINGER, *Constraint qualification for semi-infinite systems of convex inequalities*, SIAM J. Optim., 11 (2000), pp. 31–52.
- [17] O. MANGASARIAN, *Nonlinear Programming*, McGraw-Hill, New York, 1969.
- [18] W. R. MADYCH AND S. A. NELSON, *Multivariate interpolation and conditionally positive definite functions*, Approx. Theory Appl., 4 (1988), pp. 77–89.
- [19] R. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [20] G. SH. RUBENSTEIN, *On an extremal problem in a linear normed space*, Sibirsk. Mat. Z., 6 (1965), pp. 711–714. (in Russian).
- [21] G. S. SMIRNOV AND R. G. SMIRNOV, *Best uniform restricted ranges approximation II*, C. R. Acad. Sci. Paris Ser. I Math., 330 (2000), pp. 1059–1064.
- [22] H. STRAUSS, *Approximation with constraints by conditionally positive definite functions*, Numer. Algorithms, 30 (2002), pp. 185–198.
- [23] Y. YUAN AND W. SUN, *Optimization Theory and Methods*, Science Press, Beijing, 1997 (in Chinese).

CONVERGENCE OF MINIMAL SETS IN CONVEX VECTOR OPTIMIZATION*

ENRICO MIGLIERINA[†] AND ELENA MOLHO[‡]

Abstract. We study the behavior of the minimal sets of a sequence of convex sets $\{A_n\}$ converging to a given set A . The main feature of the present work is the use of convexity properties of the sets A_n and A to obtain upper and lower convergence of the minimal frontiers. We emphasize that we study both Kuratowski–Painlevé convergence and Attouch–Wets convergence of minimal sets. Moreover, we prove stability results that hold in a normed linear space ordered by a general cone, in order to deal with the most common spaces ordered by their natural nonnegative orthants (e.g., $C([a, b])$, l^p , and $L^p(\mathbb{R})$ for $1 \leq p \leq \infty$). We also make a comparison with the existing results related to the topics considered in our work.

Key words. stability, vector optimization, minimal points, set-convergences, convex sets

AMS subject classifications. 90C29, 90C31, 54B20

DOI. 10.1137/030602642

1. Introduction. The study of the behavior of minimal sets under perturbations plays a fundamental role in vector optimization (see, e.g., the recent books [9] and [10]). The classical way to deal with stability in vector optimization is to study the continuity properties of the optimal multifunction (see, e.g., [17] and, among recent works, [3]). Our approach is based on the notion of set-convergence: we consider a sequence of perturbed sets $\{A_n\}$ converging to a given set A and we study the convergence of the minimal sets $\text{Min}A_n$. The same problem is discussed in [2] and in [14] (see also the references therein).

The distinguishing feature of our work lies in the use of convexity assumptions to establish stability results in the sense of convergence of minimal sets in a normed linear space. In a finite-dimensional setting the stability of minimal sets under convexity has already been studied in [19] and [15]. The fact that convexity plays an important role in stability theory in vector optimization is not surprising; also in the scalar case, convexity is deeply related to stability.

In section 3, we consider the lower convergence of minimal sets. The domination property, a quite common tool in the theory of vector optimization (see, e.g., [13]), plays an important role in this section. Whenever the ordering cone has a sequentially weakly compact base, the Mosco convergence of a sequence of perturbed sets A_n , together with the convexity of A_n for every n implies the lower Kuratowski–Painlevé convergence of $\text{Min}A_n$ to $\text{Min}A$. It is known (see [7]) that the assumptions on the order structure are unduly restrictive since they are not satisfied by the nonnegative orthants of the most common infinite-dimensional normed vector lattices (e.g., $C([a, b])$, l^p , and $L^p(\mathbb{R})$, $1 \leq p \leq \infty$).

In the more general setting of a normed linear space X , ordered by a closed convex pointed cone, an example shows that the Mosco convergence of the perturbed

*Received by the editors December 22, 2003; accepted for publication (in revised form) May 24, 2004; published electronically February 19, 2005.

<http://www.siam.org/journals/siopt/15-2/60264.html>

[†]Dipartimento di Economia, Università degli Studi dell’Insubria, via Ravasi 2, 21100 Varese, Italy (emiglierina@eco.uninsubria.it).

[‡]Dipartimento di Ricerche Aziendali, Università degli Studi di Pavia, via San Felice 5, 27100 Pavia, Italy (molhoe@eco.unipv.it).

sets A_n to A is not enough to ensure the lower convergence of minimal sets. Even if we strengthen the assumptions to the Attouch–Wets convergence of A_n to A , we just obtain the lower Kuratowski–Painlevé convergence to $\text{StMin}A$, the set of the strictly minimal points of A .

It is not surprising that strict minimality plays a central role in stability problems in vector optimization, as originally emphasized by Bednarczuk in [3]. Moreover, we underline the relevance of the notion of strict minimality in establishing the well-posedness of the vector optimization problem (see [20] and [16]).

To obtain a lower Kuratowski–Painlevé convergence result to the larger set $\text{Min}A$, we require an additional convexity restriction on the limit set A ; i.e., we assume that A is uniformly rotund. Finally, in order to obtain the Attouch–Wets lower convergence of minimal sets, we add a compactness assumption on the set of minimal points of A .

In section 4, we study the upper convergence of minimal sets. It is well known that, without any convexity assumption, we have the upper convergence of $\text{WMin}A_n$ to $\text{WMin}A$. If we add a rotundity assumption on the limit set A , we can prove the upper Kuratowski–Painlevé convergence of $\text{Min}A_n$ to $\text{Min}A$. We remark that here the role of weak minimality is crucial. Hence, all the previous results require that the ordering cone has nonempty interior. This hypothesis is unduly restrictive since non-negative orthants of many common infinite-dimensional spaces (e.g., l^p and $L^p(\mathbb{R})$, $1 \leq p < \infty$) have an empty interior. If we add the requirement that all the perturbed sets A_n are convex and that the sequence $\{A_n\}$ converges to A with respect to Attouch–Wets convergence, we can prove directly that $\text{Min}A$ is the upper Kuratowski–Painlevé limit of $\text{Min}A_n$. Moreover, with an additional compactness assumption involving the whole sequence A_n , we obtain the upper Attouch–Wets convergence of $\text{Min}A_n$ to $\text{Min}A$.

We emphasize that our approach, based on convexity assumptions, allows us to obtain a unified framework for both lower and upper convergence of minimal sets. Finally, we remark that a strengthening of the convexity requirements on the limit set A corresponds to stronger stability properties of the set $\text{Min}A$.

In the last section we compare our results with some existing theorems related to similar topics.

2. Preliminaries. Let X be a normed linear space, by B_ρ we denote the closed ball centered at 0 and with radius ρ . The space X is ordered by a pointed (i.e., $K \cap (-K) = \{0\}$) closed convex cone K . Let X^* be the topological dual space of X ; we define the *polar* (respectively, *strict polar*) cone of K :

$$K^+ = \{\phi \in X^* : \phi(x) \geq 0 \text{ for every } x \in K\},$$

$$(K^{+i} = \{\phi \in X^* : \phi(x) > 0 \text{ for every } x \in K \setminus \{0\}\}).$$

A cone K is said to have a *base* G if $G \subset X$ is convex, $0 \notin \text{cl}G$, and

$$K = \{\lambda g : \lambda \in \mathbb{R}^+, g \in G\}.$$

A based cone is necessarily pointed and convex.

Given a subset A of X we define the set

$$\text{Min}A = \{a \in A : A \cap (a - K) = \{a\}\};$$

the elements of the set $\text{Min}A$ are called the *minimal points* of A . When $\text{int}K$ is nonempty we can also consider the set

$$\text{WMin}A = \{a \in A : A \cap (a - \text{int}K) = \emptyset\};$$

the elements of the set $\text{WMin}A$ are named *weakly minimal points* of A .

In the theory of vector optimization some refinements of the notion of minimality were introduced in order to avoid anomalous features. For a survey and a comparison of these notions, see, e.g., [11] and [20]. In what follows we will consider positive proper minimality and strict minimality. The point $a_0 \in A$ is a *positive properly minimal point* of A if there exists some $\phi \in K^{+i}$ such that $\phi(a) \geq \phi(a_0)$ for all $a \in A$. We denote the set of positive properly minimal points of A by $\text{Pos}A$. The definition of strictly minimal point was introduced in [4] and [3] by Bednarczuk. The point $a_0 \in A$ is a *strictly minimal point* of A when for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$(A - a_0) \cap (B_\delta - K) \subset B_\varepsilon.$$

We denote the set of strictly minimal points of A by $\text{StMin}A$. It is easy to see that both the previous definitions imply minimality. If the ordering cone has a bounded base, we recall that positive proper minimality implies strict minimality. The same situation does not hold when the cone K does not have a bounded base, as shown in Example 3.6 below.

A subset A of X satisfies the *domination property* if, for every $a \in A$, there is $a_0 \in \text{Min}A$ such that $a \in a_0 + K$ (see, e.g., [13]). The importance of the domination property in the stability of the vector optimization problem is well known.

Various notions of convergence of a sequence of sets have been developed in the literature (for a complete overview on this topic see [5]). Recently these notions were widely employed to study the stability of an optimization problem. Here we will consider only the following four definitions. We recall that, given a sequence of sets $\{A_n\}$ in X , the *Kuratowski–Painlevé lower* and *upper limits* are defined as

$$\begin{aligned} \text{Li}A_n &= \left\{ x \in X : x = \lim_{n \rightarrow \infty} x_n, x_n \in A_n, \text{ for all large } n \right\}, \\ \text{Ls}A_n &= \left\{ x \in X : x = \lim_{s \rightarrow \infty} x_s, x_s \in A_{n_s}, \{n_s\} \text{ subsequence of } \{n\} \right\}. \end{aligned}$$

If a subset A of X satisfies the conditions $\text{Ls}A_n \subset A \subset \text{Li}A_n$, we say that the sequence $\{A_n\}$ converges to A in the sense of Kuratowski–Painlevé and we denote it by $A_n \xrightarrow{K} A$. When we consider the limits in the weak topology on X rather than in the original norm topology, we denote the lower and the upper limits above by $\text{w-Li}A_n$ and $\text{w-Ls}A_n$. When $\text{w-Ls}A_n \subset A \subset \text{w-Li}A_n$ we say that the sequence $\{A_n\}$ converges to A in the sense of *Kuratowski–Painlevé with respect to weak topology* and we denote it by $A_n \xrightarrow{w-K} A$. Moreover, when $\text{w-Ls}A_n \subset A \subset \text{Li}A_n$ we say that $\{A_n\}$ converges to A in the sense of *Mosco* ($A_n \xrightarrow{M} A$). It is easy to observe that $A_n \xrightarrow{M} A$ if and only if $A_n \xrightarrow{K} A$ both in the original and in the weak topology.

The last notion of set-convergence that we recall is the Attouch–Wets convergence (also known as bounded Hausdorff convergence). To illustrate this notion we introduce some notations. Given $x \in X$ and two nonempty subsets A and B of X , we have

$$\begin{aligned} d(x, A) &= \inf_{a \in A} \|x - a\| & (d(x, \emptyset) &= \infty), \\ e(A, B) &= \sup_{a \in A} d(a, B) & (e(\emptyset, A) = 0 \quad e(\emptyset, \emptyset) = 0, \quad e(A, \emptyset) = \infty), \\ e_\rho(A, B) &= e(A \cap B_\rho, B), \\ h_\rho(A, B) &= \max \{e_\rho(A, B), e_\rho(B, A)\}. \end{aligned}$$

We say that the sequence $\{A_n\}$ converges to A in the sense of Attouch–Wets when

$$\lim_{n \rightarrow \infty} h_\rho(A_n, A) = 0 \quad \text{for all } \rho > 0.$$

As usual, we can split this notion of convergence into an upper part and a lower part as follows:

$$\begin{aligned} \lim_{n \rightarrow \infty} e_\rho(A_n, A) &= 0, \\ (\lim_{n \rightarrow \infty} e_\rho(A, A_n) &= 0). \end{aligned}$$

For the relationships between the various notions of set convergence introduced here, see, e.g., [18] (especially for sequences of nonempty closed convex sets in normed linear spaces, see the detailed scheme of p. 342). We recall that, if X is a finite-dimensional space, the four quoted notions of set-convergence coincide whenever we consider a sequence $\{A_n\}$ of closed sets.

We remark that the limit set of a Kuratowski–Painlevé converging sequence of convex sets is convex (see, e.g., [1, Prop. 3.1]).

Finally, we recall two definitions concerning some restrictions of the notion of convexity of a set A (see, e.g., [12]). A nonempty convex set $A \subset X$ is said to be *rotund* when the boundary of A does not contain line segments. Moreover, it is said to be *uniformly rotund* when there exists a nondecreasing function $\delta : [0, +\infty) \rightarrow [0, +\infty)$ satisfying $\delta(0) = 0$, $\delta(t) > 0$ for every $t > 0$ and such that for every $x_1, x_2 \in A$, we have

$$\frac{1}{2}(x_1 + x_2) + B_{\delta(\|x_1 - x_2\|)} \subset A.$$

3. Lower convergence of minimal sets. In this section we prove some results on the lower convergence of the minimal points of a sequence of convex sets in an infinite-dimensional setting.

We begin to study the normed linear spaces ordered by a closed convex cone with a sequentially weakly compact base. In this framework we prove the lower Kuratowski–Painlevé convergence of the minimal sets of a sequence of closed convex sets A_n to $\text{Min}A$, whenever A is the Mosco limit of A_n .

THEOREM 3.1. *Let X be a normed linear space, K be a closed convex cone with a sequentially weakly compact base, and $\{A_n\}_{n \in \mathbb{N}}$ be a sequence of subsets of X . If*

1. A_n is a closed convex set for every n ,
2. A_n satisfies the domination property for every n ,
3. $A_n \xrightarrow{M} A$,

then $\text{Min}A \subset \text{Li Min}A_n$.

If we substitute assumption 3 in the previous theorem with Kuratowski–Painlevé convergence with respect to weak topology of the sets A_n to A , we can obtain the following result.

THEOREM 3.2. *Let X be a normed linear space, K be a closed cone with a sequentially weakly compact base, and $\{A_n\}_{n \in \mathbb{N}}$ be a sequence of subsets of X . If*

1. A_n is a closed convex set for every n ,
2. A_n satisfies the domination property for every n ,
3. $A_n \xrightarrow{w-K} A$,

then $\text{Min}A \subset \text{w-Li Min}A_n$.

Now we give a proof of Theorem 3.1 that holds, with minor changes, also for Theorem 3.2. We explicitly add the necessary changes for Theorem 3.2 in brackets.

Proof. If $\text{Min}A$ is empty, the thesis follows trivially. If $\text{Min}A \neq \emptyset$, we suppose to the contrary that there exists $a \in \text{Min}A$ such that $a \notin \text{Li Min}A_n$ ($a \notin \text{w-Li Min}A_n$). Without loss of generality we can choose $a = 0$. By hypothesis 3 there exists a sequence $\{a_n\}$ such that $a_n \in A_n$ and $\lim_n a_n = 0$ ($\text{w-lim}_n a_n = 0$). Since $0 \notin \text{Li Min}A_n$ ($0 \notin \text{w-Li Min}A_n$) we can find a subsequence $\{a_{n_j}\}$ of the sequence $\{a_n\}$ such that $a_{n_j} \in A_{n_j} \setminus \text{Min}A_{n_j}$. Now there exists a sequence $\{b_j\}$ such that $b_j \in \text{Min}A_{n_j}$ and $a_{n_j} - b_j \in K$. We consider two distinct cases.

- The sequence $\{b_j\}$ (weakly) converges to 0. The thesis follows immediately.
- The sequence $\{b_j\}$ does not (weakly) converge to 0. We recall that $b_j \in a_{n_j} - K$ and $\lim_j a_{n_j} = 0$ ($\text{w-lim}_j a_{n_j} = 0$). Let

$$[a_{n_j}, b_j] := \{x \in X : x = \alpha a_{n_j} + (1 - \alpha)b_j, 0 \leq \alpha \leq 1\};$$

then we have $[a_{n_j}, b_j] \subset A_{n_j} \cap (a_{n_j} - K)$.

Let G be a sequentially weakly compact base; then $b_j = a_{n_j} - \lambda_j g_j$, where $g_j \in G$ and $\lambda_j > 0$. If $\lambda_j \rightarrow 0$, then $\{b_j\} = \{a_{n_j} - \lambda_j g_j\}$ (weakly) converges to 0, a contradiction. Hence $\{\lambda_j\}$ does not converge to 0. Up to a subsequence, it is $\lambda_j > \epsilon$, where ϵ is a positive real number. Now take $\tilde{\lambda}_j = \lambda_j/\epsilon$ and $\tilde{g}_j = \epsilon g_j \in \tilde{G}$, where $\tilde{G}$ is the base ϵG . Let $c_j = \alpha_j a_{n_j} + (1 - \alpha_j)b_j$, where $\alpha_j \in [0, 1]$; then $c_j = a_{n_j} - (1 - \alpha_j)\tilde{\lambda}_j \tilde{g}_j$. Now we can choose $\alpha_j = 1 - \frac{1}{\tilde{\lambda}_j}$, and hence $c_j = a_{n_j} - \tilde{g}_j$, where we recall that $\lim_j a_{n_j} = 0$ ($\text{w-lim}_j a_{n_j} = 0$). Since $\{\tilde{g}_j\}$ admits a weakly convergent subsequence to $\bar{g} \in \tilde{G}$, there exists a subsequence $\{c_{j_k}\}$ weakly converging to $c = -\bar{g}$. It is $c \neq 0$. Since $c_j \in A_j \cap (a_{n_j} - K)$, it is

$$c \in A \cap (-K)$$

against the minimality of 0. $\square$

Remark 3.3. In a finite-dimensional setting, assumption 2 in Theorem 3.1 can be dropped. Let us consider only the nontrivial case where $\text{Min}A$ is a nonempty set. Since A is a closed convex set we have $A_\infty \cap (-K) = \{0\}$, where

$$A_\infty = \{x \in X : \exists \{t_n\} \text{ with } t_n > 0, t_n \rightarrow 0, \exists \{a_n\} \subset A \text{ s.t. } t_n a_n \rightarrow x\}.$$

Now, since $Ls (A_n)_\infty \subset A_\infty$, it follows that there exists n_0 such that $(A_n)_\infty \cap (-K) = \{0\}$ for $n \geq n_0$. By Corollary 3.2.27 in [10], A_n has the domination property for $n \geq n_0$. Thus we can obtain the same result as in Theorem 3.1 in [15] as a consequence of Theorem 3.1.

The assumption that the ordering cone K has a sequentially weakly compact base is widely used in vector optimization. Nevertheless, it is known (see, e.g., [7]) that in normed spaces the assumption that the ordering cone K has sequentially weakly compact base is restrictive: for instance, the nonnegative orthants in l^p and L^p , $1 \leq p \leq \infty$, do not have a sequentially weakly compact base. The following example shows that the previous hypothesis cannot be relaxed even in a separable Hilbert space.

Example 3.4. Let $X = l^2$ endowed with the usual norm and let K be the nonnegative orthant. Let $\{e_n\}_{n \in \mathbb{N}}$ be the canonical orthonormal base of X , i.e., $e_n = (0, \dots, 0, \underset{\text{nth}}{1}, 0, \dots)$. We consider the sequence of sets $\{A_n\}_{n \in \mathbb{N}}$ defined as follows:

$$A_n = [-ne_n, 0].$$

We observe that $A_n \xrightarrow{M} \{0\}$ but $\text{Min}A_n = \{-ne_n\}$ and $\text{w-Li Min}A_n = \emptyset$ while $A = \text{Min}A = \{0\}$.

In what follows we drop the assumption of weak compactness on the base of the ordering cone K . If we strengthen assumption 3 to the Attouch–Wets convergence we obtain the lower convergence of $\text{Min}A_n$ to $\text{StMin}A$, the set of the strictly minimal points of A .

THEOREM 3.5. *Let X be a normed linear space, K be a closed convex pointed cone, and $\{A_n\}_{n \in \mathbb{N}}$ be a sequence of subsets of X . If*

1. A_n is a closed convex set for every n ,
2. A_n satisfies the domination property for every n ,
3. $A_n \xrightarrow{AW} A$,

then $\text{StMin}A \subset \text{Li Min}A_n$.

Proof. If $\text{StMin}A$ is empty, the theorem is trivial. If $\text{StMin}A \neq \emptyset$, we prove the thesis by contradiction. Let $a \in \text{StMin}A \setminus \text{Li Min}A_n$.

Without loss of generality, let $a = 0$. Since the lower part of the Attouch–Wets convergence implies the lower Kuratowski–Painlevé convergence of $\{A_n\}$ to A , there exists a sequence $\{a_n\}$ such that $a_n \in A_n$ and $a_n \rightarrow 0$. Since $0 \notin \text{Li Min}A_n$ we can find a subsequence $\{a_{n_j}\}$ of the sequence $\{a_n\}$ such that $a_{n_j} \in A_{n_j} \setminus \text{Min}A_{n_j}$. Now, by assumption 2, there exists a sequence $\{b_j\}$ such that $b_j \in \text{Min}A_{n_j}$ and $b_j \in a_{n_j} - K$. By convexity, it is, for any $j \in \mathbb{N}$,

$$[a_{n_j}, b_j] \subset A_{n_j} \cap (a_{n_j} - K),$$

where $[a_{n_j}, b_j] = \{x \in X : x = \alpha a_{n_j} + (1 - \alpha)b_j, \alpha \in [0, 1]\}$.

Now, recalling that $a_n \rightarrow 0$ and using the upper part of the Attouch–Wets convergence, for any $n \in \mathbb{N}$ and for any $\rho > 0$, we have

$$(3.1) \quad [a_{n_j}, b_j] \cap B_\rho \subset (A + B_{\eta_j}) \cap (B_{\eta_j} - K),$$

where $\{\eta_j\}$ is a sequence of real numbers such that $\eta_j \rightarrow 0$.

Since $0 \in \text{StMin}A$, for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$A \cap (B_\delta - K) \subset B_\varepsilon.$$

Then there exists j_0 such that for every $j \geq j_0$, it holds

$$(3.2) \quad (A + B_{\eta_j}) \cap (B_{\eta_j} - K) \subset B_\varepsilon.$$

Since $0 \notin \text{Li Min}A_n$ the sequence $\{b_j\}$ has a subsequence not converging to 0. Then we obtain a contradiction (against (3.1) and (3.2)). $\square$

The following example shows that we cannot obtain the same convergence result for the larger set $\text{Min}A$, without additional assumptions.

Example 3.6. Let $X = l^2$ endowed with the usual norm and let $K = l_+^2$ be the nonnegative orthant. Let $\{e_k\}_{k \in \mathbb{N}}$ be the canonical orthonormal base of X . Let

$$A = \{x \in l^2 : \langle x^*, x \rangle = 0\} \cap B_2,$$

where $x^* = (\frac{1}{k})_{k \in \mathbb{N} \setminus \{0\}}$. Since $\langle x^*, x \rangle = 0$, for all $x \in A$ and $x^* \in K^{+i}$, it is easy to see that $\text{Pos } A = \text{Min } A = A$. Now we show that $0 \notin \text{StMin } A$. Let us consider the sequence $\{x_s\} \subset A$,

$$x_s = \frac{1}{s}e_1 - e_s, \quad s = 2, 3, \dots$$

Then $x_s \in B_{1/s} - K$ and $\|x_s\| > 1$. Hence $0 \notin \text{StMin } A$.

Let us consider the sequence of sets

$$A_n = \{x \in l^2 : \langle x_n^*, x \rangle = 0\} \cap B_2,$$

where $x_n^* = x^* - \frac{1}{n-1}e_n$. The sets A_n satisfy the assumptions of Theorem 3.5 (for assumption 3 see Theorem 3.4.1 in [5]). Since $x_n^* \notin K^+$, we have that $\text{Min}A_n \subset \text{bd}B_2$. Therefore $0 \in \text{Min } A$ and $0 \notin \text{Li Min}A_n$.

It is not surprising that strict minimality plays a role in our stability result; indeed, in [3] the notion of strict efficiency was related to lower continuity of the optimal multifunction. In [3] and [20] the notion of strict minimality was compared with some kinds of proper minimality. We remark that in Example 3.6 there is a positive properly minimal point that is not even a strictly minimal point. This is an interesting fact since, under the assumption of weak compactness of the base of the ordering cone, positive proper minimality is the most restrictive notion of proper minimality (see, e.g., [11]), while in a more general framework the situation is completely different.

If we add a compactness assumption involving only the minimal set of A , we obtain a stronger lower convergence of $\text{Min}A_n$ to the strictly minimal set of A .

THEOREM 3.7. *Let X be a normed linear space, K be a closed convex cone, and $\{A_n\}_{n \in \mathbb{N}}$ be a sequence of subsets of X . If*

1. A_n is a closed convex set for every n ,
2. A_n satisfies the domination property for every n ,
3. $A_n \xrightarrow{AW} A$,
4. $\text{StMin}A \cap B_\rho$ is a relatively compact set for every $\rho > 0$,

then for each $\rho > 0$, $\lim_{n \rightarrow \infty} e_\rho(\text{StMin}A, \text{Min}A_n) = 0$.

Proof. If $\text{StMin}A$ is empty, the theorem is trivial. If $\text{StMin}A \neq \emptyset$, let us suppose that the thesis of the theorem does not hold. Then there exist $\rho > 0$, $\varepsilon > 0$, and a subsequence $\{A_{n_k}\}_{k \in \mathbb{N}}$ of $\{A_n\}$, such that

$$e_\rho(\text{StMin}A, \text{Min}A_{n_k}) > 2\varepsilon$$

for every k . This fact implies that, for every k , there exists $m_k \in \text{StMin}A \cap B_\rho$ such that

$$(3.3) \quad d(m_k, \text{Min}A_{n_k}) > 2\varepsilon.$$

Since $\text{StMin}A \cap B_\rho$ is relatively compact, without loss of generality, we can suppose that $\{m_k\}$ converges to an element $a \in \text{cl}(\text{StMin}A \cap B_\rho)$ (if this is not the case, it is enough to choose the sequence $\{n_k\}$ appropriately). Then Theorem 3.5 implies that, for every k , there exists a sequence $\{a_s^{(k)}\}_{s \in \mathbb{N}}$ such that $a_s^{(k)} \in \text{Min}A_{n_s}$ and $a_s^{(k)} \rightarrow m_k$. Now we can choose an increasing function $f : \mathbb{N} \rightarrow \mathbb{N}$ such that

$$(3.4) \quad a_{f(k)}^{(k)} \rightarrow a.$$

By (3.3), we have

$$d(a, \text{Min}A_{n_k}) > \varepsilon$$

for large k . It is easy to see that (3.4) contradicts the last assertion. $\square$

Remark 3.8. The convergence of the sequence of sets $\{A_n\}$ actually used in the proof of the previous theorems is weaker than the Attouch–Wets convergence.

We need only the upper part of Attouch–Wets convergence, while for the lower part Kuratowski–Painlevé convergence is enough. This type of convergence is known in the literature as b-proximal convergence (see, e.g., [18]). Some interesting results involving b-proximal convergence can be found in [6]. In the same work a meaningful example can be found that shows that the b-proximal convergence is significantly weaker than the Attouch–Wets convergence.

Clearly, whenever $\text{Min}A \subset \text{cl StMin}A$, we obtain lower convergence of $\text{Min}A_n$ to $\text{Min}A$. In [4] a sufficient condition to ensure that the former inclusion holds for weakly compact sets is given. The following result gives a condition based on an enforcement of convexity that is more suitable to our setting.

PROPOSITION 3.9. *Let X be a normed linear space, K be a closed convex cone, and A be a nonempty subset of X . If A is uniformly rotund, then $\text{Min} A = \text{StMin} A$.*

Proof. It is sufficient to prove $\text{Min}A \subset \text{StMin}A$. By contradiction, let us suppose that there exists a point $a \in \text{Min}A \setminus \text{StMin}A$. Without loss of generality we suppose that $a = 0$. Then there exists a real number $\eta > 0$ and two sequences $\{q_n\} \subset X$ and $\{k_n\} \subset K$ such that $q_n \rightarrow 0$, $q_n - k_n \in A$, and $\|q_n - k_n\| > \eta$ for every n . Let $x_n = q_n - k_n$. It holds that

$$d\left(\frac{1}{2}x_n, -K\right) \rightarrow 0;$$

hence, by the minimality of 0,

$$d\left(\frac{1}{2}x_n, X \setminus A\right) \rightarrow 0.$$

By the uniform rotundity of A , there exists a positive real number $\delta(\eta)$ such that $d\left(\frac{1}{2}x_n, X \setminus A\right) \geq \delta(\eta)$, a contradiction. $\square$

Remark 3.10. In the previous proposition the uniform rotundity of set A can be substituted by the following weaker assumption: for every $x_0 \in \text{Min}A$, there exists a nondecreasing function $\delta_{x_0} : [0, +\infty) \rightarrow [0, +\infty)$ satisfying $\delta_{x_0}(0) = 0$, $\delta_{x_0}(t) > 0$ for every $t > 0$ and such that for every $x \in A$

$$\frac{1}{2}(x_0 + x) + B_{\delta_{x_0}(\|x_0 - x\|)} \subset A.$$

Using Proposition 3.9, from Theorems 3.5 and 3.7 we obtain the following results on the lower convergence of $\text{Min}A_n$ to $\text{Min}A$, without any assumption involving the base of the ordering cone.

COROLLARY 3.11. *Let X be a normed linear space, K be a closed convex pointed cone, and $\{A_n\}_{n \in \mathbb{N}}$ be a sequence of subsets of X . If*

1. A_n is a closed convex set for every n ,
2. A_n satisfies the domination property for every n ,
3. $A_n \xrightarrow{AW} A$,
4. A is a uniformly rotund set,

then $\text{Min}A \subset \text{Li Min}A_n$.

COROLLARY 3.12. *Let X be a normed linear space, K be a closed convex pointed cone, and $\{A_n\}_{n \in \mathbb{N}}$ be a sequence of subsets of X . If*

1. A_n is a closed convex set for every n ,
2. A_n satisfies the domination property for every n ,
3. $A_n \xrightarrow{AW} A$,

4. $\text{Min}A \cap B_\rho$ is a relatively compact set for every $\rho > 0$,
 5. A is a uniformly rotund set,
- then for each $\rho > 0$, $\lim_{n \rightarrow \infty} e_\rho(\text{Min}A, \text{Min}A_n) = 0$.

4. Upper convergence of minimal sets. In this section we want to study the upper convergence of the set of minimal points.

The following simple example shows that, in a finite-dimensional setting, $\text{Min}A_n$ do not converge to $\text{Min}A$, even if $\{A_n\}$ is a sequence of compact convex sets.

Example 4.1. Let $X = \mathbb{R}^2$ and $K = \mathbb{R}_+^2$. Let

$$A_n = \{(x_1, x_2) \in \mathbb{R}^2 : x_2 = (-1/n)x_1, n \in \mathbb{N} \setminus \{0\}\} \cap B_1.$$

A_n is a compact convex set for every n . It is $A_n \xrightarrow{K} A = \{(x_1, x_2) \in \mathbb{R}^2 : x_2 = 0\} \cap B_1$. Now $\text{Min}A_n = A_n$ and $\text{Min}A = \{(-1, 0)\}$.

The example suggests that without additional assumptions an upper convergence result can be obtained just for the set of the weakly minimal points of A . Indeed, the following classical result [17], [8] allows us to study the upper convergence of the weakly minimal frontiers without any convexity assumption. For the convenience of the reader, we explicitly quote the proof as stated in [15], which holds also in a more general setting.

THEOREM 4.2. *Let X be a normed linear space, K be a closed convex cone with $\text{int}K \neq \emptyset$, and $\{A_n\}_{n \in \mathbb{N}}$ be a sequence of subsets of X . If*

1. A_n is a closed set for every n ,
2. $A_n \xrightarrow{K} A$,

then $\text{Ls WMin}A_n \subset \text{WMin}A$.

Proof. Let $\{a_k\}$ be a sequence such that $a_k \in \text{WMin } A_{n_k}$ and $a_k \rightarrow a$. Clearly, $a \in A$. We suppose, by contradiction, that there exists $y \in A$ such that $y \in a - \text{int}K$. Now take $e \in \text{int}K$ and $\sigma > 0$ so small that $y + \sigma e \in a - \sigma e - \text{int}K$, and take a sequence $\{y_k\}$ such that $y_k \in A_{n_k}$ and $y_k \rightarrow y$. Then eventually $y_k \in y + \sigma e - K$ and $a - \sigma e \in a_k - K$, and this contradicts the weak minimality in A_{n_k} of the elements a_k . $\square$

In order to deal with the stability of the set of minimal points of A , we state a result that ensures the coincidence of minima and weak minima under the requirement of rotundity of the set A .

PROPOSITION 4.3. *Let X be a normed space and K be a closed convex pointed cone with $\text{int}K \neq \emptyset$. If $A \subset X$ is a rotund set, then $\text{WMin}A = \text{Min}A$.*

Proof. By contradiction we suppose there exists an element $y_0 \in \text{WMin}A \setminus \text{Min}A$. Without loss of generality we can consider $y_0 = 0$. Then there exists $y \in (-K \setminus (-\text{int}K \cup \{0\})) \cap A$. By the rotundity of A the line segment

$$(0, y] = \{x \in X : x = \alpha y, \alpha \in (0, 1]\}$$

does not lie entirely in the boundary of A . Hence there exists $z \in \text{int}A \cap (0, y]$. Now we can choose $k \in -\text{int}K$ such that $z + k \in A \cap -\text{int}K$, against the weak minimality of 0. $\square$

The next result follows immediately from Theorem 4.2 and Proposition 4.3. Thus, using a refinement of the notion of convexity, we obtain an upper convergence result for the set of minimal points.

THEOREM 4.4. *Let X be a normed linear space, K be a closed convex cone with $\text{int}K \neq \emptyset$, and $\{A_n\}_{n \in \mathbb{N}}$ be a sequence of subsets of X . If*

1. A_n is a closed set for every n ,

2. $A_n \xrightarrow{K} A$,
3. A is a rotund set,

then $\text{Ls Min}A_n \subset \text{Min}A$.

To prove the previous result, we consider the weakly minimal points. Therefore the assumption that the interior of the ordering cone is nonempty plays an essential role. In fact, it is a restrictive hypothesis because the nonnegative orthants in many common normed spaces, such as $l^p(\mathbb{R})$ and $L^p(\mathbb{R})$, $1 \leq p < \infty$, have empty interior.

One can drop the assumptions on the interior of the ordering cone if one adds the requirement of convexity on the sets A_n and considers a stronger notion of convergence of A_n to A .

THEOREM 4.5. *Let X be a normed linear space, K be a closed convex cone, and $\{A_n\}_{n \in \mathbb{N}}$ be a sequence of subsets of X . If*

1. A_n is a closed convex set for every n ,
2. $A_n \xrightarrow{AW} A$,
3. A is a closed rotund set,

then $\text{Ls Min}A_n \subset \text{Min}A$.

Proof. By contradiction we suppose that $a \in \text{Ls Min}A_n \setminus \text{Min}A$. Without loss of generality let $a = 0$. Hence there exists a sequence $\{a_k\}$ such that $a_k \in \text{Min}A_{n_k}$ and $a_k \rightarrow 0$. By contradiction we suppose that $0 \notin \text{Min}A$; i.e., there exists $y \in A \cap (-K \setminus \{0\})$. By the rotundity of A the line segment,

$$(0, y] = \{x \in X : x = \alpha y, \alpha \in (0, 1]\}$$

does not lie entirely in the boundary of A . Hence there exists $z \in \text{int}A \cap (0, y] \cap (-K)$. We have $a_k + z \rightarrow z$. By a consequence of the Rådström cancellation principle (see Theorem 7.4.2 in [5]) we obtain that $a_k + z \in A_{n_k}$ eventually, against the minimality of a_k in A_{n_k} . $\square$

If we add a compactness assumption we can also obtain the upper Attouch–Wets convergence of minimal sets of A_n to $\text{Min}A$.

THEOREM 4.6. *Let X be a normed linear space, K be a closed convex cone, and $\{A_n\}_{n \in \mathbb{N}}$ be a sequence of subsets of X . If*

1. A_n is a closed convex set for every n ,
2. $A_n \xrightarrow{AW} A$,
3. A is a closed rotund set,
4. $\bigcup_{n=1}^{\infty} \text{Min}A_n \cap B_\rho$ is relatively compact for every $\rho > 0$,

then for each $\rho > 0$, $\lim_{n \rightarrow \infty} e_\rho(\text{Min}A_n, \text{Min}A) = 0$.

Proof. If $\text{Min}A_n$ is empty, eventually the thesis follows trivially. Otherwise we suppose that the thesis of the theorem does not hold. Then there exists a subsequence $\{A_{n_k}\}$ and two real positive numbers ρ and ε such that

$$e_\rho(\text{Min}A_{n_k}, \text{Min}A) > \varepsilon;$$

i.e., there exists a sequence $\{m_k\}$ such that, for every k , $m_k \in \text{Min}A_{n_k}$ and

$$d(m_k, \text{Min}A) > \varepsilon.$$

By assumption 4, the sequence $\{m_k\}$ contains a subsequence $\{m_{k_s}\}$ such that $m_{k_s} \rightarrow m$. Since, by Theorem 4.5, $m \in \text{Min}A$, we have a contradiction. $\square$

We remark that in [14] assumption 4 was used to prove the lower convergence of minimal sets.

5. Conclusions. In this section we use the results proved in sections 3 and 4 to obtain two theorems concerning the convergence (in both the lower and the upper sense) of the minimal sets.

By Corollary 3.11 and Theorem 4.5 we can easily prove the following result.

THEOREM 5.1. *Let X be a normed linear space, K be a closed convex cone, and $\{A_n\}_{n \in \mathbb{N}}$ be a sequence of subsets of X . If*

1. A_n is a closed convex set for every n ,
2. A_n satisfies the domination property for every n ,
3. $A_n \xrightarrow{AW} A$,
4. A is a closed, uniformly rotund set,

then $\text{Min}A_n \xrightarrow{K} \text{Min}A$.

Adding a compactness assumption, by Corollary 3.12 and Theorem 4.6, we can have the Attouch–Wets convergence of minimal sets.

THEOREM 5.2. *Let X be a normed linear space, K be a closed convex cone, and $\{A_n\}_{n \in \mathbb{N}}$ be a sequence of subsets of X . If*

1. A_n is a closed convex set for every n ,
2. A_n satisfies the domination property for every n ,
3. $A_n \xrightarrow{AW} A$,
4. A is a closed, uniformly rotund set,
5. $\text{cl}(\bigcup_{n=1}^{\infty} \text{Min}A_n) \cap B_{\rho}$ is a compact set for every $\rho > 0$,

then $\text{Min}A_n \xrightarrow{AW} \text{Min}A$.

Proof. It is enough to show that assumption 4 in Corollary 3.12 and assumption 4 in Theorem 4.6 are satisfied. Indeed, by Corollary 3.11, it holds that, for every $\rho > 0$,

$$\text{cl}(\text{Min}A \cap B_{\rho}) \subset \text{Li } \text{Min}A_n \cap B_{\rho}.$$

Now $\text{Li } \text{Min}A_n \cap B_{\rho} \subset \text{cl}(\bigcup_{n=1}^{\infty} \text{Min}A_n) \cap B_{\rho}$, which is a compact set. Hence $\text{Min}A \cap B_{\rho}$ is a relatively compact set. Moreover,

$$\text{cl}\left(\bigcup_{n=1}^{\infty} \text{Min}A_n \cap B_{\rho}\right) \subset \text{cl}\left(\bigcup_{n=1}^{\infty} \text{Min}A_n\right) \cap B_{\rho};$$

then $\bigcup_{n=1}^{\infty} \text{Min}A_n \cap B_{\rho}$ is relatively compact. $\square$

We emphasize that the set of minimal points of a convex set A has a valuable stability property (with respect to convex perturbations) whenever A satisfies a geometrical regularity condition that strengthens the convexity assumption (i.e., A is a uniformly rotund set). Thus, under convexity assumptions, we obtain a unified framework to deal with both upper and lower convergence of minimal sets.

6. Comparison with the existing results. Among the results on the stability of minimal sets, initially we focus on those that can be formulated in terms of the lower part of some notion of set-convergence. In this section we present some existing theorems in order to make a comparison with our results. We state these results as in [14].

THEOREM 6.1 (see [17], [8]). *Let X be a normed space and K be a closed convex cone. Assume that the following conditions hold:*

1. $A_n \xrightarrow{K} A$.
2. A_n satisfies the domination property for every n .
3. *If $a_n \in A_n$ is such that $\lim_{n \rightarrow +\infty} a_n$ exists and $e_n \in \text{Min}A_n \cap (a_n - K)$, then admits a convergent subsequence.*

Then $\text{Min}A \subset \text{Li Min}A_n$.

If we consider a Banach space we have the following results.

THEOREM 6.2 (see [2]). *Let X be a Banach space, K be such that*

$$-K \subset \{x \in X : l(x) + \varepsilon \|x\| \leq 0\}$$

for some $l \in X^*, \varepsilon > 0$, and $\{A_n\}_{n \in \mathbb{N}}$ be a sequence of subsets of X . If

1. $A_n \xrightarrow{K} A$,
2. $\inf_{n \in \mathbb{N}} \inf_{x \in A_n} l(x) > -\infty$,
3. *for every $\rho > 0$, there exists a compact subset $C_\rho \subset X$, such that, for every $n \in \mathbb{N}$, $\text{Min}A_n \cap B_\rho \subset C_\rho$,*

then $\text{Min}A \subset \text{Li Min}A_n$.

The original version of this theorem also gives an existence condition for the minimal points under the same hypotheses. We underline that the assumption made about the ordering cone K is equivalent to the requirement that K has a bounded base if $K \neq \{0\}$. In a reflexive Banach space the compactness assumption 3 can be dropped using Mosco convergence of $\{A_n\}$ to A , instead of Kuratowski–Painlevé convergence (see [2, Theorem 3.5]). We recall also that in this setting the boundedness of the base implies its weak compactness.

THEOREM 6.3 (see [2]). *Let X be a reflexive Banach space, K be a closed convex cone such that*

$$-K \subset \{x \in X : l(x) + \varepsilon \|x\| \leq 0\}$$

for some $l \in X^*, \varepsilon > 0$, and $\{A_n\}_{n \in \mathbb{N}}$ be a sequence of subsets of X . If

1. $A_n \xrightarrow{M} A$,
2. $\inf_{n \in \mathbb{N}} \inf_{x \in A_n} l(x) > -\infty$,

then $\text{Min}A \subset \text{Li Min}A_n$.

The last result that we quote gives a stronger lower convergence of minimal sets; i.e., it provides some conditions to assure the Attouch–Wets convergence of minimal sets of A_n to $\text{Min}A$.

THEOREM 6.4 (see [14]). *Assume that the following conditions hold:*

1. $A_n \xrightarrow{K} A$.
2. A_n satisfies the domination property for every n .
3. *If $a_n \in A_n$ is such that $\lim_{n \rightarrow +\infty} a_n$ exists and $e_n \in \text{Min}A_n \cap (a_n - K)$, then $\{e_n\}$ admits a convergent subsequence.*
4. *For every $\rho > 0$, $\text{Min}A \cap B_\rho$ is relatively compact.*

Then for each $\rho > 0$, $\lim_{n \rightarrow \infty} e_\rho(\text{Min}A, \text{Min}A_n) = 0$.

We recall that in Theorems 6.1 and 6.4 the authors also take into account perturbations of the ordering cone. We just consider the case of a constant cone K , for the sake of easier comparisons with our results.

The hypotheses used in the present work have a completely different nature. The results quoted above are based on some “compactness” requirement of the entire sequence $\{A_n\}$ (hypothesis 3 in Theorem 6.1, 3 in Theorem 6.2, or 3 in Theorem 6.4); and also in Theorem 6.3, where there is not a compactness assumption, the boundedness condition 2 concerns the whole sequence $\{A_n\}$. We remark that, except for Theorem 4.6, we avoid any requirement involving the family $\{A_n\}_{n \in \mathbb{N}}$ as a whole by the use of a convexity assumption on each set A_n . We show by some examples that, even in a euclidean space, the two approaches are distinct.

The following example proves that, even in $\mathbb{R}^n$, our assumptions do not imply the compactness assumption 3 of Theorems 6.1 and 6.4.

Example 6.5. Let $X = \mathbb{R}^2$ and $K = \mathbb{R}_+^2$. Let

$$A_n = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 = -1/n, x_2 \geq -n\}.$$

A_n is a closed convex set that satisfies the domination property for every n . It is $A_n \xrightarrow{K} A = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 = 0\}$ (hence $A_n \xrightarrow{AW} A$). We can see that condition 3 in Theorem 6.1 does not hold. As a matter of fact we consider the sequence $\{a_n\} = \{(-1/n, 0)\}$. It is $\lim_{n \rightarrow +\infty} a_n = (0, 0)$ and the only sequence $\{e_n\}$ such that $e_n \in \text{Min}A_n \cap (a_n - K)$ is the sequence $\{(-1/n, -n)\}$, which has no convergent subsequence.

We underline that this example is not conclusive since $\text{Min}A$ is empty; hence the comparison of the two distinct sets of assumptions in the case $\text{Min}A \neq \emptyset$ is still an open question.

The following example shows that our assumptions do not imply the boundedness condition 2 of Theorems 6.2 and 6.3.

Example 6.6. Let $X = \mathbb{R}^2$ and $K = \mathbb{R}_+^2$. Let

$$A_n = \{(x_1, x_2) \in \mathbb{R}^2 : x_2 = (-1 + 1/n)x_1\}.$$

A_n is a closed convex set that satisfies the domination property for every n . It is $A_n \xrightarrow{K} A = \{(x_1, x_2) \in \mathbb{R}^2 : x_2 = -x_1\}$. It can be easily verified that

$$\inf_{n \in \mathbb{N}} \inf_{x \in A_n} l(x) = -\infty$$

for every linear functional l .

In [3] the lower semicontinuity of the optimal multifunction is studied. Although in her work Bednarczuk does not explicitly use set convergences, her results are closely related to ours. Here we recall only that her assumptions are not comparable with those introduced in the present work, since a notion of Hausdorff convergence is considered.

In the existing literature there are some results on the upper convergence of the weakly minimal points. To our knowledge the only result on the upper convergence of minimal points is the following.

THEOREM 6.7 (see [14]). *Let X be a reflexive space and let K be a pointed convex cone. If*

1. $A_n \xrightarrow{wK} A$,
2. $w\text{-cl}(X \setminus K) \subset (X \setminus K) \cup \{0\}$,

then $\liminf d(\text{Min}A_n, x) \geq d(\text{Min}A, x)$ for every $x \in X$.

We quote here just a version of the original theorem adapted to the case of a fixed ordering cone.

The previous result only apparently concerns the minimal points of A . In fact, whenever $K \neq \{0\}$, assumption 2 implies that $K \setminus \{0\} \subset \text{int}K$, hence $\text{Min}A = \text{WMin}A$. Moreover, we remark that assumption 2 is severe since it rules out the case of a closed nontrivial ordering cone K and cannot be dropped as shown in the first example in section 4 [14]. In Theorem 4.5 the upper convergence of minimal sets is established under stronger assumptions on the convergence and on the structure of the sets A_n and A . Nevertheless, our result holds in a more general space ordered by a closed convex cone, and hence it can be applied to the most common normed linear spaces ordered by their nonnegative orthants.

Acknowledgments. The authors wish to thank R. Lucchetti for useful discussions during the preparation of the paper. Moreover, the authors thank the anonymous referees of a previous version of this work, whose remarks and suggestions helped them to substantially improve this work.

REFERENCES

- [1] H. ATTOUCH, *Variational Convergence for Functions and Operators*, Pitman, Boston, 1984.
- [2] H. ATTOUCH AND H. RIHAI, *Stability results for Ekeland's ϵ -variational principle and cone extremal solutions*, Math. Oper. Res., 18 (1993), pp. 173–201.
- [3] E. M. BEDNARCZUK, *A note on lower semicontinuity of minimal points*, Nonlinear Anal., 50 (2002), pp. 285–287.
- [4] E. M. BEDNARCZUK AND W. SONG, *PC points and their application to vector optimization*, Pliska Stud. Math. Bulgar., 12 (1998), pp. 21–30.
- [5] G. BEER, *Topologies on Closed and Closed Convex Sets*, Mathematics and Its Applications, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993.
- [6] G. BEER AND R. LUCCHETTI, *Well-posed optimization problems and a new topology for the closed subsets of a metric space*, Rocky Mountain J. Math., 23 (1993), pp. 1197–1220.
- [7] J. P. DAUER AND R. J. GALLAGHER, *Positive proper efficient points and related cone results in vector optimization theory*, SIAM J. Control Optim., 28 (1990), pp. 158–172.
- [8] S. DOLECKI AND C. MALIVERT, *Stability of efficient sets: Continuity of mobile polarities*, Nonlinear Anal., 12 (1988), pp. 1461–1486.
- [9] M. EHROGOTT AND X. GANDIBLEUX, EDS., *Multiple Criteria Optimization. State of the Art Annotated Bibliographic Surveys*, Kluwer Academic Publishers, Boston, 2002.
- [10] A. GÖPFERT, H. RIAHI, C. TAMMER, AND C. ZĂLINESCU, *Variational Methods in Partially Ordered Spaces*, CMS Books Math. 17, Springer-Verlag, New York, 2003.
- [11] A. GUERRAGGIO, E. MOLHO, AND A. ZAFFARONI, *On the notion of proper efficiency in vector optimization*, J. Optim. Theory Appl., 82 (1994), pp. 1–21.
- [12] R. B. HOLMES, *Geometric Functional Analysis and Its Applications*, Grad. Texts in Math. 24, Springer-Verlag, New York, 1975.
- [13] D. T. LUC, *Theory of Vector Optimization*, Lecture Notes in Econom. and Math. Systems 319, Springer-Verlag, Berlin, Heidelberg, 1989.
- [14] D. T. LUC, R. LUCCHETTI, AND C. MALIVERT, *Convergence of the efficient sets*, Set-Valued Anal., 2 (1994), pp. 207–218.
- [15] R. LUCCHETTI AND E. MIGLIERINA, *Stability for convex vector optimization problems*, Optimization, to appear.
- [16] E. MIGLIERINA AND E. MOLHO, *Well-posedness and convexity in vector optimization*, Math. Methods Oper. Res., 58 (2003), pp. 375–385.
- [17] J. P. PENOT AND A. STERNA-KARWAT, *Parametrized multicriteria optimization: Continuity and closedness of optimal multifunction*, J. Math. Anal. Appl., 120 (1986), pp. 150–168.
- [18] Y. SONNTAG AND C. ZĂLINESCU, *Set convergences: A survey and a classification*, Set-Valued Anal., 2 (1994), pp. 339–356.
- [19] T. TANINO, *Stability and sensitivity analysis in convex vector optimization*, SIAM J. Control Optim., 26 (1988), pp. 521–536.
- [20] A. ZAFFARONI, *Degrees of efficiency and degrees of minimality*, SIAM J. Control Optim., 42 (2003), pp. 1071–1086.

CONVERGENCE AND APPROXIMATION OF OPTIMIZATION PROBLEMS*

JORGE ALVAREZ-MENA[†] AND ONÉSIMO HERNÁNDEZ-LERMA[‡]

Abstract. In this paper we consider a general optimization problem (OP) and study the convergence and approximation of optimal values and optimal solutions to changes in the cost function and the set of feasible solutions. By a “general” OP we mean that the cost function and the constraints are defined on a Hausdorff topological space. First we obtain convergence results for a general OP, and then we present an application of these results on the approximation of the optimal value and the optimal solutions for the so-called *general capacity* problem in metric spaces.

Key words. minimization problem, approximation, general capacity problem

AMS subject classification. 90C48

DOI. 10.1137/S1052623402413938

1. Introduction. In this paper we consider a general constrained optimization problem (OP)

$$\mathbb{P}_\infty : \text{minimize } \{f_\infty(x) : x \in \mathcal{F}_\infty\},$$

and a sequence of approximating problems

$$\mathbb{P}_n : \text{minimize } \{f_n(x) : x \in \mathcal{F}_n\}.$$

(The notation is explained in section 2.) The questions we are interested in are

- (i) the convergence of the sequence of optimal values $\{\min \mathbb{P}_n\}$ —or subsequences thereof—to $\min \mathbb{P}_\infty$ and
- (ii) the convergence of sequences of optimal solutions of $\{\mathbb{P}_n\}$ —or subsequences thereof—to optimal solutions of $\mathbb{P}_\infty$.

We give conditions under which the convergence in (i) and (ii) holds; see Theorem 2.3. As an application of these results we study the approximation of the optimal value and the optimal solutions for the so-called *general capacity* (GC) problem in metric spaces.

General OPs appear in many branches of mathematics, including probability theory, numerical analysis, optimal control, game theory, mathematical economics, and operations research, to name just a few [1, 2, 3, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20]. Of course, by their very nature, these problems are in general difficult to solve in *closed* form, and so one is forced to use some kind of approximation procedure, which in turn is closely related to the convergence issues (i) and (ii).

The problem of finding conditions under which (i) and (ii) hold is of great interest, and it has been studied in many different settings; see, e.g., [1, 5, 6, 9, 12, 13, 15, 17, 18, 19, 20] and their references. In particular, the problem can be studied via

*Received by the editors August 30, 2002; accepted for publication (in revised form) May 12, 2004; published electronically February 19, 2005.

<http://www.siam.org/journals/siopt/15-2/41393.html>

[†]Programa de Investigación en Matemáticas Aplicadas y Computación, IMP, A. Postal 14-805, México D.F. 07730, México (jamena@imp.mx).

[‡]Departamento de Matemáticas, CINVESTAV-IPN, A. Postal 14-740, México D.F. 07000, México (ohernand@math.cinvestav.mx). The research of this author was partially supported by CONACyT grant 37355-3.

the notion of Γ -convergence (or epiconvergence) of sequences of functionals [4, 8]. However, the approach used in this paper is more direct and generalizes several known results [4, 8, 9, 15, 17, 18]; see Remark 2.4. Even more, Example 3.7 shows that our assumptions are strictly weaker than those considered in [4, 8, 9, 15, 17, 18]. Namely, Example 3.7 studies a particular GC problem in which our assumptions are satisfied, but the assumptions considered in those references fail to hold. The GC problem has been previously analyzed in, e.g., [2, 3, 11] from different viewpoints.

The remainder of the paper is organized as follows. In section 2 we present our main results on the convergence and approximation of general OPs. These results are applied in section 3 to the GC problem, and a particular case of the GC problem is analyzed.

2. Convergence of general OPs. We shall use the notation $\mathbb{N} := \{1, 2, \dots\}$, $\bar{\mathbb{N}} := \mathbb{N} \cup \{\infty\}$ and $\bar{\mathbb{R}} := \mathbb{R} \cup \{\infty, -\infty\}$.

Let $\mathcal{X}$ be a *Hausdorff topological space*. For each $n \in \bar{\mathbb{N}}$, consider a function $f_n : \mathcal{X} \rightarrow \bar{\mathbb{R}}$, a set $\mathcal{F}_n \subset \mathcal{X}$, and the optimization problem

$$\begin{aligned} \mathbb{P}_n : \quad & \text{minimize } f_n(x) \\ & \text{subject to } x \in \mathcal{F}_n. \end{aligned}$$

We call $\mathcal{F}_n$ the set of *feasible solutions* for $\mathbb{P}_n$. If $\mathcal{F}_n$ is nonempty, the (optimum) *value* of $\mathbb{P}_n$ is defined as $\inf \mathbb{P}_n := \inf\{f_n(x) \mid x \in \mathcal{F}_n\}$; otherwise, $\inf \mathbb{P}_n := +\infty$. The problem $\mathbb{P}_n$ is said to be *solvable* if there is a feasible solution x^* that achieves the optimum value. In this case, x^* is called an *optimal solution* for $\mathbb{P}_n$, and the value $\inf \mathbb{P}_n$ is then written as $\min \mathbb{P}_n = f_n(x^*)$. We shall denote by $\mathcal{M}_n$ the *minimum set*, that is, the set of optimal solutions for $\mathbb{P}_n$.

To state our assumptions we will use Kuratowski's [16] concept of *outer* and *inner limits* of $\{\mathcal{F}_n\}$, denoted by $\mathbf{OL}\{\mathcal{F}_n\}$ and $\mathbf{IL}\{\mathcal{F}_n\}$, respectively, and defined as follows:

$$\begin{aligned} \mathbf{OL}\{\mathcal{F}_n\} := \{x \in \mathcal{X} \mid x = \lim_{i \rightarrow \infty} x_{n_i}, \text{ where } \{n_i\} \subset \mathbb{N} \text{ is an increasing} \\ \text{sequence such that } x_{n_i} \in \mathcal{F}_{n_i} \ \forall i\}. \end{aligned}$$

Thus a point $x \in \mathcal{X}$ is in $\mathbf{OL}\{\mathcal{F}_n\}$ if x is an accumulation point of a sequence $\{x_n\}$ with $x_n \in \mathcal{F}_n$ for all n . On the other hand, if x is the limit of the sequence $\{x_n\}$ itself, then x is in the inner limit $\mathbf{IL}\{\mathcal{F}_n\}$, i.e.,

$$\begin{aligned} \mathbf{IL}\{\mathcal{F}_n\} := \{x \in \mathcal{X} \mid x = \lim_{n \rightarrow \infty} x_n, \text{ where } x_n \in \mathcal{F}_n \text{ for all } n \text{ but} \\ \text{a finite number of } n\text{'s}\}. \end{aligned}$$

In these definitions we may, of course, replace $\{\mathcal{F}_n\}$ with any other sequence of subsets of $\mathcal{X}$. Also note that $\mathbf{IL}\{\cdot\} \subset \mathbf{OL}\{\cdot\}$.

We shall consider two sets of hypotheses.

Assumption 2.1.

(a) The minimum sets $\mathcal{M}_n$ satisfy that

$$(2.1) \quad \mathbf{OL}\{\mathcal{M}_n\} \subset \mathcal{F}_\infty.$$

(b) If x_{n_i} is in $\mathcal{M}_{n_i}$ for all i and $x_{n_i} \rightarrow x$ (so that x is in $\mathbf{OL}\{\mathcal{M}_n\}$), then

$$(2.2) \quad \liminf_{i \rightarrow \infty} f_{n_i}(x_{n_i}) \geq f_\infty(x).$$

- (c) For each $x \in \mathcal{F}_\infty$ there exist $N \in \mathbb{N}$ and a sequence $\{x_n\}$ with $x_n \in \mathcal{F}_n$ for all $n \geq N$, and such that $x_n \rightarrow x$ and $\lim_{n \rightarrow \infty} f_n(x_n) = f_\infty(x)$.

Assumption 2.2. Parts (b) and (c) are the same as in Assumption 2.1. Moreover,
(a) the minimum sets $\mathcal{M}_n$ satisfy

$$(2.3) \quad \mathbf{IL}\{\mathcal{M}_n\} \subset \mathcal{F}_\infty.$$

Note that Assumption 2.1(c) implies, in particular, $\mathcal{F}_\infty \subset \mathbf{IL}\{\mathcal{F}_n\}$, but the equality does not necessarily hold. In fact, in section 3 we give an example in which Assumption 2.1 is satisfied, in particular $\mathcal{F}_\infty \subset \mathbf{IL}\{\mathcal{F}_n\}$, but $\mathcal{F}_\infty \neq \mathbf{IL}\{\mathcal{F}_n\}$ (see Example 3.7).

On the other hand, note that Assumptions 2.2 (a) and (c) yield that

$$\mathbf{IL}\{\mathcal{M}_n\} \subset \mathcal{F}_\infty \subset \mathbf{IL}\{\mathcal{F}_n\}.$$

THEOREM 2.3. (a) *If Assumption 2.1 holds, then*

$$(2.4) \quad \mathbf{OL}\{\mathcal{M}_n\} \subset \mathcal{M}_\infty.$$

In other words, if $\{x_n\}$ is a sequence of minimizers of $\{\mathbb{P}_n\}$, and a subsequence $\{x_{n_i}\}$ of $\{x_n\}$ converges to $x \in X$, then x is optimal for $\mathbb{P}_\infty$. Furthermore, the optimal values of $\mathbb{P}_{n_i}$ converge to the optimal value of $\mathbb{P}_\infty$, that is,

$$(2.5) \quad \min \mathbb{P}_{n_i} = f_{n_i}(x_{n_i}) \rightarrow f_\infty(x) = \min \mathbb{P}_\infty.$$

(b) *Suppose that Assumption 2.2 holds. Then*

$$\mathbf{IL}\{\mathcal{M}_n\} \subset \mathcal{M}_\infty.$$

If in addition $\mathbf{IL}\{\mathcal{M}_n\}$ is nonempty, then

$$(2.6) \quad \min \mathbb{P}_n \rightarrow \min \mathbb{P}_\infty.$$

Proof. We prove only (a) because the proof of (b) is quite similar.

To prove (a), let $x \in \mathcal{X}$ be in the outer limit $\mathbf{OL}\{\mathcal{M}_n\}$. Then there is a sequence $\{n_i\} \subset \mathbb{N}$ and $x_{n_i} \in \mathcal{M}_{n_i}$ for all i such that

$$(2.7) \quad x_{n_i} \rightarrow x.$$

Moreover, by Assumption 2.1(a), x is in $\mathcal{F}_\infty$. To prove that x is in $\mathcal{M}_\infty$, choose an arbitrary $x' \in \mathcal{F}_\infty$ and let $\{x'_n\}$ and N be as in Assumption 2.1(c) for x' , that is, x'_n is in $\mathcal{F}_n$ for all $n \geq N$, $x'_n \rightarrow x'$, and $f_n(x'_n) \rightarrow f_\infty(x')$. Furthermore, if $\{n_i\} \subset \mathbb{N}$ is as in (2.7), then the subsequence $\{x'_{n_i}\}$ of $\{x'_n\}$ also satisfies the following:

$$(2.8) \quad x'_{n_i} \text{ is in } \mathcal{F}_{n_i}, \quad x'_{n_i} \rightarrow x', \quad \text{and} \quad f_{n_i}(x'_{n_i}) \rightarrow f_\infty(x').$$

Combining the latter fact with Assumption 2.1(b) and the optimality of each x_{n_i} we get

$$\begin{aligned} f_\infty(x) &\leq \liminf_{i \rightarrow \infty} f_{n_i}(x_{n_i}) && \text{(by (2.2))} \\ &\leq \liminf_{i \rightarrow \infty} f_{n_i}(x'_{n_i}) \\ &= f_\infty(x') && \text{(by (2.8)).} \end{aligned}$$

Hence, as $x' \in \mathcal{F}_\infty$ was arbitrary, it follows that x is in $\mathcal{M}_\infty$, that is, (2.4) holds.

To prove (2.5), suppose again that x is in $\mathbf{OL}\{\mathcal{M}_n\}$ and let $x_{n_i} \in \mathcal{M}_{n_i}$ be as in (2.7). By Assumption 2.1(c), there exists a sequence $x'_{n_i} \in \mathcal{F}_{n_i}$ that satisfies (2.8) for x instead of x' ; thus

$$\begin{aligned} f_\infty(x) &\leq \liminf_{i \rightarrow \infty} f_{n_i}(x_{n_i}) && \text{(by (2.2))} \\ &\leq \limsup_{i \rightarrow \infty} f_{n_i}(x_{n_i}) \\ &\leq \limsup_{i \rightarrow \infty} f_{n_i}(x'_{n_i}) \\ &= f_\infty(x) && \text{(by (2.8)).} \end{aligned}$$

This proves (2.5). $\square$

In Theorem 2.3 the solvability of each $\mathbb{P}_n$ is not assumed. Thus $\mathbf{OL}\{\mathcal{M}_n\}$ might be empty; in fact, it might be empty even if each $\mathbb{P}_n$ is solvable. In this case, the (convergence of minimizers) inclusion (2.4) trivially holds. In the convergence of the optimal values (2.5) and (2.6), unlike the convergence of minimizers, it is implicitly assumed that $\mathbf{OL}\{\mathcal{M}_n\}$ is nonempty.

Remark 2.4. (i) Parts (a) and (b) of Theorem 2.3 generalize in particular some results in [17, 18] and [9, 15], respectively. Indeed, using our notation, in [9, 15, 17, 18] it is assumed that the cost functions f_n are *continuous* and converge *uniformly* to f_∞ . On the other hand, with respect to the feasible sets $\mathcal{F}_n$, in [9] it is assumed that $\mathbf{IL}\{\mathcal{F}_n\} = \mathcal{F}_\infty$, whereas in [15, 17, 18] it is required that $\mathcal{F}_n \rightarrow \mathcal{F}_\infty$ in the Hausdorff metric. These hypotheses trivially yield the following conditions:

(C₁) The *inner* and/or the *outer limit* of the feasible sets $\mathcal{F}_n$ coincide with $\mathcal{F}_\infty$, i.e.,

$$(2.9) \quad \mathbf{IL}\{\mathcal{F}_n\} = \mathcal{F}_\infty$$

or

$$(2.10) \quad \mathbf{OL}\{\mathcal{F}_n\} = \mathbf{IL}\{\mathcal{F}_n\} = \mathcal{F}_\infty.$$

(C₂) For every x in $\mathcal{X}$ and for every sequence $\{x_n\}$ in $\mathcal{X}$ converging to x , it holds that

$$(2.11) \quad \lim_{n \rightarrow \infty} f_n(x_n) = f_\infty(x).$$

However, instead of (2.10) and (2.11) we require (the weaker) Assumption 2.1, and instead of (2.9) and (2.11) we require (the weaker) Assumption 2.2.

(ii) Theorem 2.3 generalizes the results in [4, 8], where the notion of Γ -convergence is used. Indeed, [4, 8] study problems of the form

$$(2.12) \quad \min_{x \in \mathcal{X}} F_n(x).$$

Each of our problems $\mathbb{P}_n$ can be put in the form (2.12) by letting

$$F_n(x) := \begin{cases} f_n(x) & \text{if } x \in \mathcal{F}_n, \\ \infty & \text{if } x \notin \mathcal{F}_n, \end{cases}$$

and then, when the space $\mathcal{X}$ is *first countable*, the assumptions in [4, 8] can be translated to this context as follows: the sequence $\{F_n\}$ Γ -converges to F_∞ ; see Theorems 7.8 and 7.18 in [8]. On the other hand, when $\mathcal{X}$ is first countable, the sequence $\{F_n\}$ Γ -converges to F_∞ if and only if

(C₃) for every x in $\mathcal{X}$ and for every sequence $\{x_n\}$ in $\mathcal{X}$ converging to x , it holds that

$$\liminf_{n \rightarrow \infty} F_n(x_n) \geq F_\infty(x);$$

(C₄) for every x in $\mathcal{X}$ there exists a sequence $\{x_n\}$ in $\mathcal{X}$ converging to x such that

$$\lim_{n \rightarrow \infty} F_n(x_n) = F_\infty(x).$$

See Proposition 8.1 in [8]. It is natural to assume that $f_n(x) < \infty$ for each $x \in \mathcal{F}_n$ and $n \in \overline{\mathbb{N}}$, and that $\mathcal{F}_\infty$ is nonempty. In this case, (C₃) implies part (b) of Assumptions 2.1 and 2.2, (C₄) implies part (c), and (C₃) together with (C₄) imply part (a). Indeed, the last statement can be proved as follows. Let $x \in \mathcal{X}$ be in $\mathbf{OL}\{\mathcal{M}_n\}$. Then there is a sequence $\{n_i\} \subset \mathbb{N}$ and $x_{n_i} \in \mathcal{M}_{n_i}$ for all i such that $x_{n_i} \rightarrow x$. Now, as $\mathcal{F}_\infty$ is nonempty, we can take x' in $\mathcal{F}_\infty$. For this x' , let $\{x'_n\} \in \mathcal{X}$ be as in (C₄), so that

$$\begin{aligned} F_\infty(x) &\leq \liminf_{i \rightarrow \infty} F_{n_i}(x_{n_i}) && \text{(by (C}_3\text{))} \\ &\leq \liminf_{i \rightarrow \infty} F_{n_i}(x'_{n_i}) && \text{(because } x_{n_i} \text{ is in } \mathcal{M}_{n_i}\text{)} \\ &= \lim_{i \rightarrow \infty} F_{n_i}(x'_n) && \text{(by (C}_4\text{))} \\ &= F_\infty(x') < \infty && \text{(by (C}_4\text{)).} \end{aligned}$$

Hence x is in $\mathcal{F}_\infty$.

On the other hand, if in addition we assume that $f_n(x) \leq K$ for all $x \in \mathcal{F}_n$, $n \in \mathbb{N}$, and some $K \in \mathbb{R}$, then (C₃) and (C₄) imply (2.10). In fact, (C₄) implies the inclusion $\mathcal{F}_\infty \subset \mathbf{IL}\{\mathcal{F}_n\} \subset \mathbf{OL}\{\mathcal{F}_n\}$, and (C₃) together with the uniform boundedness condition imply the reverse inclusion.

In the next section we present an application of Theorem 2.3. We also show, in Example 3.7, a particular problem in which Assumption 2.1 is satisfied, but the assumptions considered in [4, 8, 9, 15, 17, 18] do not hold.

3. The GC problem. The GC problem is related to the problem of determining the electrostatic capacity of a conducting body. In fact, it originated in the above-mentioned electrostatic capacity problem; see, for instance, [2, 3].

Let X and Y be metric spaces endowed with their corresponding Borel σ -algebras $\mathcal{B}(X)$ and $\mathcal{B}(Y)$. We denote by $M(Y)$ the vector space of finite signed measures on Y , and by $M^+(Y)$ the cone of nonnegative measures in $M(Y)$.

Now let $b : X \rightarrow \mathbb{R}$, $c : Y \rightarrow \mathbb{R}$, and $g : X \times Y \rightarrow \mathbb{R}$ be nonnegative Borel-measurable functions. Then the GC problem can be stated as follows:

$$\begin{aligned} \text{GC: minimize } & \int_Y c(y) \mu(dy) \\ \text{subject to } & \int_Y g(x, y) \mu(dy) \geq b(x) \quad \forall x \in X, \mu \in M^+(Y). \end{aligned}$$

In this section we study the convergence problem (see (i) and (ii) in section 1) in which g and c are replaced with sequences of nonnegative measurable functions $g_n : X \times Y \rightarrow \mathbb{R}$ and $c_n : Y \rightarrow \mathbb{R}$, for $n \in \overline{\mathbb{N}}$, such that $g_n \rightarrow g_\infty =: g$ and $c_n \rightarrow c_\infty =: c$ uniformly.

Thus we shall deal with GC problems

$$(3.1) \quad \begin{aligned} & \text{GC}_n: \text{ minimize } \int_Y c_n(y) \mu(dy) \\ & \text{subject to } \int_Y g_n(x, y) \mu(dy) \geq b(x) \quad \forall x \in X, \mu \in M^+(Y) \end{aligned}$$

for $n \in \bar{\mathbb{N}}$. For each $n \in \bar{\mathbb{N}}$, we denote by $\mathcal{F}_n$ the set of feasible solutions for GC_n , that is, the set of measures μ that satisfy (3.1), but in addition $\int_Y c_n d\mu < \infty$.

Convergence. We shall study the convergence issue via Theorem 2.3. First, we introduce assumptions that guarantee the solvability of the GC problems. We shall distinguish two cases for the cost functions c_n , the bounded case and the unbounded case, which require slightly different hypotheses. For the bounded case we suppose the following.

Assumption 3.1 (bounded case). For each $n \in \bar{\mathbb{N}}$, we have the following:

- (a) $\mathcal{F}_n$ is nonempty.
- (b) The function $g_n(x, \cdot)$ is bounded above and upper semicontinuous (u.s.c.) for each $x \in X$.
- (c) The function c_n is bounded and lower semicontinuous (l.s.c.). Further c_n is *bounded away from zero*; that is, there exist $\delta_n > 0$ such that $c_n(y) \geq \delta_n$ for all $y \in Y$.

In addition,

- (d) the space Y is compact.

For the unbounded case, we replace parts (c) and (d) with an inf-compactness hypothesis.

Assumption 3.2 (unbounded case). Parts (a)–(c) are the same as in Assumption 3.1. Moreover,

- (d) for each $n \in \bar{\mathbb{N}}$, the function c_n is *inf-compact*, which means that, for each $r \in \mathbb{R}$, the set $\{y \in Y | c_n(y) \leq r\}$ is compact. Further, c_n is bounded away from zero.

Observe that the inf-compactness condition implies that c_n is l.s.c.

We next introduce the assumptions for our convergence and approximation results. As above, we require two sets of assumptions depending on whether the cost functions c_n are bounded or unbounded. (See Remark 3.12 for alternative sets of assumptions.)

Assumption 3.3 (bounded case).

- (a) (*Slater condition*). There exist $\mu \in \mathcal{F}_\infty$ and $\eta > 0$ such that

$$\int_Y g_\infty(x, y) \mu(dy) \geq b(x) + \eta \quad \forall x \in X.$$

- (b) $g_n \rightarrow g_\infty$ uniformly on $X \times Y$.
- (c) $c_n \rightarrow c_\infty$ uniformly on Y .

Assumption 3.4 (unbounded case). Parts (a) and (b) are the same as in Assumption 3.3. Moreover,

- (c) $c_n \downarrow c_\infty$ uniformly on Y .

Before stating our main result for the GC problem we recall some facts on the weak convergence of measures (for further details see [7] or Chapter 12 in [13], for instance).

DEFINITION 3.5. Let Y , $M(Y)$, and $M^+(Y)$ be as at the beginning of this section. A sequence $\{\mu_n\}$ in $M^+(Y)$ is said to be bounded if there exists a constant m such

that $\mu_n \leq m$ for all n . Let $C_b(Y)$ be the vector space of continuous bounded functions on Y . We say that a sequence $\{\mu_n\}$ in $M(Y)$ converges weakly to $\mu \in M(Y)$ if $\mu_n \rightarrow \mu$ in the weak topology $\sigma(M(Y), C_b(Y))$, i.e.,

$$\int_Y u \, d\mu_n \rightarrow \int_Y u \, d\mu \quad \forall u \in C_b(Y).$$

A subset M_0 of $M^+(Y)$ is said to be relatively compact if for any sequence $\{\mu_n\}$ in M_0 there is a subsequence $\{\mu_m\}$ of $\{\mu_n\}$ and a measure μ in $M^+(Y)$ (but not necessarily in M_0) such that $\mu_m \rightarrow \mu$ weakly. In the latter case, we say that μ is a weak accumulation point of $\{\mu_n\}$.

We now state our main result in this section.

THEOREM 3.6. *Suppose that either Assumptions 3.1 and 3.3, or 3.2 and 3.4, hold. Then*

- (a) GC_n is solvable for every $n \in \bar{\mathbb{N}}$.
- (b) The optimal value of GC_n converges to the optimal value of GC_∞ , i.e.,

$$(3.2) \quad \min \text{GC}_n \longrightarrow \min \text{GC}_\infty.$$

Furthermore, if $\mu_n \in M^+(Y)$ is an optimal solution for GC_n for each $n \in \mathbb{N}$, then the sequence $\{\mu_n\}$ is relatively compact, and every weak accumulation point of $\{\mu_n\}$ is an optimal solution for GC_∞ .

- (c) If GC_∞ has a unique optimal solution, say μ , then for any μ_n in the set of optimal solutions for GC_n , with $n \in \mathbb{N}$, the sequence $\{\mu_n\}$ converges weakly to μ .

The proof of Theorem 3.6 is given below after some technical preliminaries. In the meantime we give an example which satisfies our hypotheses, Assumption 2.1, but the hypotheses used in [4, 8, 9, 15, 17, 18] do *not* hold.

Example 3.7. This example shows a particular GC problem in the *unbounded case*, in which our assumptions are satisfied, but the conditions used in [4, 8, 9, 15, 17, 18] do *not* hold. Hence, this example shows that our assumptions are *strictly weaker* than those considered in [4, 8, 9, 15, 17, 18].

We first compare our hypotheses with those in [9, 15, 17, 18]. In our notation, the latter conditions are as follows (see Remark 2.4).

- (i) For each sequence $\mu_n \in M^+(Y)$ such that $\mu_n \rightarrow \mu$ weakly we have

$$\lim_{n \rightarrow \infty} \int_Y c_n \, d\mu_n = \int_Y c_\infty \, d\mu.$$

- (ii) $\mathbf{IL}\{\mathcal{F}_n\} = \mathcal{F}_\infty$.

Consider the spaces $X = Y = [0, 2]$, and for each $n \in \bar{\mathbb{N}}$ let $g_n \equiv 1$, and

$$c_n(x) := \begin{cases} 1 & \text{if } x = 0, \\ \frac{1}{x} & \text{if } x \in (0, 1], \\ 1 & \text{if } x \in (1, 2]. \end{cases}$$

Let $b \equiv 0$. With these elements the set $\mathcal{F}_n$ of feasible solutions for each problem GC_n is given by

$$\mathcal{F}_n := \left\{ \mu \in M^+([0, 2]) : \int g_n \, d\mu = \mu([0, 2]) \geq b, \int c_n \, d\mu < \infty \right\}.$$

As the cost functions c_n are unbounded, we consider the Assumptions 3.2 and 3.4, which are obviously true in the present case, and which in turn imply Assumption 2.1; see Lemma 3.11 below. Next we show that (i) and (ii) *do not hold*.

Let μ be the Lebesgue measure on $Y = [0, 2]$, and for each $n \in \mathbb{N}$ let μ_n be the restriction of μ to $[1/n, 2]$, i.e., $\mu_n(B) := \mu(B \cap [1/n, 2])$ for all $B \in \mathcal{B}(Y)$. Thus μ_n is in $\mathcal{F}_n$ for each $n \geq 2$, and $\mu_n \rightarrow \mu$ weakly. Therefore, μ is in $\text{IL}\{\mathcal{F}_n\}$, but μ is *not* in $\mathcal{F}_\infty$ because $\int c_\infty d\mu = \infty$. Hence (ii) *does not hold*.

Similarly, let $\mu'_n := (1/k_n)\mu_n$ with $k_n := 1 + \ln(n)$. Then μ'_n is in $\mathcal{F}_n$ for all $n \geq 2$, and $\mu'_n \rightarrow 0 =: \mu'$ weakly, but

$$\int c_n d\mu'_n = 1 \not\rightarrow \int c_\infty d\mu' = 0.$$

Thus (i) *is not satisfied*.

Now we compare our assumptions with those in [4, 8]. This can be done because, as $M([0, 2])$ is metrizable, the space $\mathcal{X} = M([0, 2])$ is first countable. See Remark 2.4.

We take X, Y, c_n, g_n , and $\mathcal{F}_n$ as above, but now we take $b = 1/2$. As in the former case, Assumption 2.1 holds. Now we slightly modify the set of feasible solutions by

$$\tilde{\mathcal{F}}_n := \left\{ \mu \in M^+([0, 2]) : \int g_n d\mu = \mu([0, 2]) \geq \frac{1}{2}, \int c_n d\mu < 1 \right\}.$$

Notice that $\tilde{\mathcal{F}}_n \not\subseteq \emptyset$ and $\tilde{\mathcal{F}}_n \subset \mathcal{F}_n$ for all $n \geq 2$. The sequence of modified GC problems, say $\{GC_n\}$, also satisfies Assumption 2.1. Indeed, parts (a) and (b) of Assumption 2.1 hold because the minimum sets have not changed ($\mathcal{M}_n = \tilde{\mathcal{M}}_n$), and part (c) is true since it holds for $\mathcal{F}_\infty$, and $\tilde{\mathcal{F}}_\infty \subset \mathcal{F}_\infty$. Next we show that condition (C₃) in Remark 2.4 *does not hold*. For each $n \in \mathbb{N}$, let

$$F_n(\mu) := \begin{cases} \int c_n d\mu & \text{if } \mu \in \tilde{\mathcal{F}}_n, \\ \infty & \text{if } \mu \notin \tilde{\mathcal{F}}_n. \end{cases}$$

Moreover, for each $n \in \mathbb{N}$, let μ''_n be the restriction of μ to $[1 + 1/n, 2]$, and let μ'' be the restriction of μ to $[1, 2]$. Hence we have $\mu''_n([0, 2]) = (n-1)/n \geq 1/2$ and $\int c_n d\mu''_n = (n-1)/n < 1$ for all $n \geq 2$. Therefore, μ''_n is in $\tilde{\mathcal{F}}_n$ for each $n \geq 2$, and $\mu''_n \rightarrow \mu''$ weakly. Thus μ'' is in $\text{IL}\{\tilde{\mathcal{F}}_n\}$, but μ'' is *not* in $\tilde{\mathcal{F}}_\infty$ because $\int c_\infty d\mu'' = 1$. Hence

$$\liminf_{n \rightarrow \infty} F_n(\mu''_n) = \liminf_{n \rightarrow \infty} \int c_n d\mu''_n = 1 < \infty = F_\infty(\mu''),$$

and so (C₃) *is not satisfied*. It follows that the F_n do *not* Γ -converge to F_∞ ; that is, the assumptions in [4, 8] *do not hold*.

Technical preliminaries. For completeness and ease of reference, we state here some well-known results that are needed to prove Theorem 3.6.

LEMMA 3.8. *Let $u : Y \rightarrow \mathbb{R}$ be a measurable function, and $\{\mu_n\}$ a sequence in $M^+(Y)$ such that $\mu_n \rightarrow \mu$ weakly.*

(a) *If u is l.s.c. and bounded below, then*

$$\liminf_{n \rightarrow \infty} \int_Y u d\mu_n \geq \int_Y u d\mu.$$

(b) If u is u.s.c. and bounded above, then

$$(3.3) \quad \limsup_{n \rightarrow \infty} \int_Y u \, d\mu_n \leq \int_Y u \, d\mu.$$

(c) Let $\{u_n\}$ be a sequence of measurable functions $u_n : Y \rightarrow \mathbb{R}$ such that $u_n \downarrow u$ pointwise. If the limit function u is l.s.c. and bounded below, then

$$(3.4) \quad \liminf_{n \rightarrow \infty} \int_Y u_n \, d\mu_n \geq \int_Y u \, d\mu.$$

Lemma 3.8 is well known; see, for instance, statement (12.3.37) on page 243 of [13].

We denote by $\mathbb{B}(Y)$ the Banach space of bounded measurable functions on Y with the supremum norm $\|\cdot\|$.

LEMMA 3.9. Let $\{\mu_n\}$ be a sequence in $M^+(Y)$, and let $\{u_n\}$ be a sequence in $\mathbb{B}(Y)$ such that $\mu_n \rightarrow \mu$ weakly, and $u_n \rightarrow u$ uniformly.

(a) If u is in $\mathbb{B}(Y)$ and it is l.s.c., then

$$\liminf_{n \rightarrow \infty} \int_Y u_n \, d\mu_n \geq \int_Y u \, d\mu.$$

(b) If u is in $\mathbb{B}(Y)$ and it is u.s.c., then

$$\limsup_{n \rightarrow \infty} \int_Y u_n \, d\mu_n \leq \int_Y u \, d\mu.$$

Proof. (a) As u is (bounded and) l.s.c., the inequality $u_n \geq u - \|u_n - u\|$ yields

$$\liminf_{n \rightarrow \infty} \int_Y u_n \, d\mu_n \geq \liminf_{n \rightarrow \infty} \left[\int_Y u \, d\mu_n - \|u_n - u\| \mu_n(Y) \right] \geq \int_Y u \, d\mu.$$

(b) If u is u.s.c., then $-u$ is l.s.c. Hence part (b) follows from (a). $\square$

LEMMA 3.10. Let M_0 be a bounded subset of $M^+(Y)$. Then M_0 is relatively compact if either

(a) Y is compact or

(b) there is a nonnegative and inf-compact function $u : Y \rightarrow \mathbb{R}$ and a constant k such that $\int_Y u \, d\mu \leq k$ for all $\mu \in M_0$.

Under (a), Lemma 3.10 is well known; under (b), it follows from Prohorov's theorem (see, e.g., [7] or [13]).

The following lemma states the connection between the capacity problems GC_n and Theorem 2.3.

LEMMA 3.11. Either set of Assumptions 3.1 and 3.3, or 3.2 and 3.4, implies that GC_n satisfies Assumption 2.1.

Proof. We will first verify parts (b) and (c) of Assumption 2.1, and then part (a).

Verification of Assumption 2.1(b). For the unbounded case (Assumptions 3.2 and 3.4) Assumption 2.1(b) follows from Lemma 3.8(c); for the bounded case (Assumptions 3.1 and 3.3) it follows from Lemma 3.9(a).

Verification of Assumption 2.1(c): Unbounded case. Let $\mu \in M^+(Y)$ be a feasible solution for GC_∞ , i.e.,

$$(3.5) \quad \int_Y g_\infty(x, y) \, \mu(dy) \geq b(x) \quad \forall x \in X.$$

On the other hand, from Assumption 3.4(b) it follows that

$$\int_Y g_n(\cdot, y)\mu(dy) \rightarrow \int_Y g_\infty(\cdot, y)\mu(dy) \quad \text{uniformly,}$$

and so, choosing $\epsilon > 0$ with $\epsilon < \frac{\eta}{2}$, there exists $N \in \mathbb{N}$ such that, for each $n \geq N$,

$$\int_Y g_n(x, y)\mu(dy) \geq \int_Y g_\infty(x, y)\mu(dy) - \epsilon \quad \forall x \in X.$$

Hence, from (3.5), for each $n \geq N$,

$$(3.6) \quad \int_Y g_n(x, y)\mu(dy) \geq b(x) - \epsilon \quad \forall x \in X.$$

Now, let $\nu \in M(Y)$ and $\eta > 0$ be as in Assumption 3.4(a), i.e.,

$$(3.7) \quad \int_Y g_\infty(x, y)\nu(dy) \geq b(x) + \eta \quad \forall x \in X.$$

Proceeding as above, Assumption 3.4(b) and (3.7) imply that (renaming N if necessary) for each $n \geq N$,

$$(3.8) \quad \int_Y g_n(x, y)\nu(dy) \geq b(x) + (\eta - \epsilon) \quad \forall x \in X.$$

Thus, taking $\gamma_\lambda := \lambda\nu + (1 - \lambda)\mu$ with $\lambda \in [0, 1]$, from (3.6) and (3.8) we get, for each $n \geq N$,

$$\int_Y g_n(x, y)\gamma_\lambda(dy) \geq b(x) + (\lambda\eta - \epsilon) \quad \forall x \in X.$$

Therefore, for each $n \geq N$ and $x \in X$,

$$\int_Y g_n(x, y)\gamma_\lambda(dy) \geq b(x) \quad \forall \lambda \in [\lambda_\epsilon, 1],$$

where $\lambda_\epsilon := \frac{\epsilon}{\eta} < \frac{1}{2}$. That is, for each $\lambda \in [\lambda_\epsilon, 1]$, γ_λ is in $\mathcal{F}_n$ for each $n \geq N$. In particular, $\gamma_{\lambda_\epsilon} \in \mathcal{F}_n$ for all $n \geq N$.

To construct the desired sequence in Assumption 2.1(c), let $\{\epsilon_i\}$ be a sequence of positive numbers such that $\epsilon_i \downarrow 0$. As in the previous paragraph, for each ϵ_i there exists a number N_i (which we can suppose to be increasing in i and that $N_i \rightarrow \infty$ as $i \rightarrow \infty$) such that

$$(3.9) \quad \gamma_i := \frac{\epsilon_i}{\eta}\nu + \left(1 - \frac{\epsilon_i}{\eta}\right)\mu$$

is in $\mathcal{F}_n$ for each $n \geq N_i$ ($\gamma_i = \gamma_{\lambda_{\epsilon_i}}$). On the other hand, as $\frac{\epsilon_i}{\eta} \rightarrow 0$ when $i \rightarrow \infty$,

$$(3.10) \quad \gamma_i \rightarrow \mu \quad \text{weakly as } i \rightarrow \infty.$$

Finally, letting

$$(3.11) \quad \mu_n := \gamma_i \quad \text{for each } N_i \leq n < N_{i+1},$$

we obtain that μ_n is in $\mathcal{F}_n$ for all $n \geq N_1$. Note that from (3.9)–(3.11) the sequence $\{\mu_n\}$ is of the form

$$\mu_n := \lambda'_n \nu + (1 - \lambda'_n) \mu$$

for a sequence of real numbers $\{\lambda'_n\} \subset (0, \frac{1}{2})$ such that $\lambda'_n \rightarrow 0$. Hence,

$$(3.12) \quad \mu_n \text{ is in } \mathcal{F}_n \quad \forall n \geq N_1,$$

$$(3.13) \quad \mu_n \rightarrow \mu \text{ weakly,}$$

and, as $c_n \downarrow c_\infty$ uniformly, the monotone convergence theorem yields that

$$(3.14) \quad \begin{aligned} \lim_{n \rightarrow \infty} \int_Y c_n d\mu_n &= \lim_{n \rightarrow \infty} \left(\lambda'_n \int_Y c_n d\nu + (1 - \lambda'_n) \int_Y c_n d\mu \right) \\ &= \int_Y c_\infty d\mu. \end{aligned}$$

This yields Assumption 2.1(b) in the unbounded case.

For the *bounded case* the proof is exactly the same as above, except that to get (3.14) we now use Assumption 3.3(c) and the dominated convergence theorem.

Verification of Assumption 2.1(a). Let $\mu \in M^+(Y)$ be in the outer limit $\mathbf{OL}\{\mathcal{M}_n\}$. Then there is a sequence $\{n_i\} \subset \mathbb{N}$ and $\mu_{n_i} \in \mathcal{M}_{n_i}$ for all i such that

$$\mu_{n_i} \rightarrow \mu.$$

Now, using the assumptions on the sequence $\{g_{n_i}\}$ (which are the same for the bounded and unbounded cases), from Lemma 3.9(b) and (3.1) we obtain

$$\begin{aligned} \int_Y g_\infty(x, y) \mu(dy) &\geq \limsup_{i \rightarrow \infty} \int_Y g_{n_i}(x, y) \mu_{n_i}(dy) \\ &\geq b(x) \quad \forall x \in X. \end{aligned}$$

Thus, to show that μ is in $\mathcal{F}_\infty$ it only remains to prove that $\int_Y c_\infty d\mu < \infty$. This clearly holds in the *bounded case*. Let us consider the *unbounded case*. By Assumption 3.2(a), there exists $\mu' \in \mathcal{F}_\infty$. Let $\{\mu'_n\}$ be as in (3.12)–(3.14); that is, for all n sufficiently large,

$$(3.15) \quad \mu'_n \in \mathcal{F}_n, \quad \mu'_n \rightarrow \mu' \text{ weakly, and } \lim_{n \rightarrow \infty} \int_Y c_n d\mu'_n = \int_Y c_\infty d\mu'.$$

Then

$$\begin{aligned} \int_Y c_\infty d\mu &\leq \liminf_{n \rightarrow \infty} \int_Y c_n d\mu_n \\ &\leq \liminf_{n \rightarrow \infty} \int_Y c_n d\mu'_n = \int_Y c_\infty d\mu' < \infty. \end{aligned}$$

Hence μ is in $\mathcal{F}_\infty$, which completes the proof of the lemma. $\square$

Proof of Theorem 3.6. (a) For the solvability of GC_n in the *bounded case* see [3]; for the *unbounded case* see [11].

(b) First we will see that any sequence of optimal solutions $\mu_n \in \mathcal{M}_n$ is relatively compact. Let $\mu' \in M^+(Y)$ be a feasible solution for GC_∞ , which exists in both cases,

bounded and unbounded, by Assumptions 3.1(a) and 3.2(a), respectively. Let $\{\mu'_n\}$ be as in the proof of Lemma 3.11 for μ' (see (3.12)–(3.14)); that is, $\{\mu'_n\}$ is as in (3.15). For the *bounded case*, c_∞ is strictly bounded away from zero and $c_n \rightarrow c_\infty$ uniformly; hence, there exist $\delta_\infty > 0$ and $N \in \mathbb{N}$ such that, for each $n \geq N$, $c_n(y) \geq \delta_\infty$ for all $y \in Y$. Thus

$$\begin{aligned} \sup_{n \geq N} \delta_\infty \mu_n(Y) &\leq \sup_{n \geq N} \int_Y c_n d\mu_n \\ &\leq \sup_{n \geq N} \int_Y c_n d\mu'_n < \infty. \end{aligned}$$

Thus, the sequence $\{\mu_n\}_{n \geq N}$ is bounded, and by Lemma 3.10(a), $\{\mu_n\}$ is relatively compact. On the other hand, in the *unbounded case*, as $c_n \downarrow c_\infty$, it holds that

$$\begin{aligned} \sup_n \int_Y c_\infty d\mu_n &\leq \sup_n \int_Y c_n d\mu_n \\ (3.16) \qquad \qquad &\leq \sup_n \int_Y c_n d\mu'_n < \infty. \end{aligned}$$

Then, as c_∞ is strictly bounded away from zero, the sequence $\{\mu_n\}$ is bounded. Therefore, by Lemma 3.10(b) the sequence $\{\mu_n\}$ is relatively compact.

Now note that, by Lemma 3.11, the results of Theorem 2.3(a) hold for GC_n . Hence, to complete the proof we need to prove only the convergence (3.2) of the optimal values of GC_n . For each $n \in \mathbb{N}$, let $\mu_n \in M^+(Y)$ be an optimal solution for GC_n . Since $\{\mu_n\}$ is relatively compact, from Theorem 2.3(a) any weak accumulation point of $\{\mu_n\}$ is an optimal solution for GC_∞ . To prove (3.2), let $\{\min \text{GC}_m\}$ be a subsequence of the bounded sequence of real numbers $\{\min \text{GC}_n\}$; see (3.16). As $\{\mu_m\}$ is relatively compact, there exist a subsequence $\{\mu_{m_i}\}$ of $\{\mu_m\}$ and a measure μ in $M^+(Y)$ such that

$$\mu_{m_i} \rightarrow \mu \text{ weakly,}$$

which together with (2.5) yields that

$$\min \text{GC}_{m_i} = \int_Y c_{m_i}(y) \mu_{m_i}(dy) \longrightarrow \int_Y c_\infty(y) \mu(dy) = \min \text{GC}_\infty.$$

Since the subsequence $\{\min \text{GC}_m\}$ of $\{\min \text{GC}_n\}$ was arbitrary, (3.2) follows.

(c) For each $n \in \mathbb{N}$, choose an arbitrary optimal solution μ_n for GC_n ; μ_n exists, by part (a). Suppose that $\{\mu_n\}$ does not converge to the unique optimal solution μ for GC_∞ . By part (b), $\{\mu_n\}$ is relatively compact. Therefore, there exists a subsequence $\{\mu_{n_i}\}$ of $\{\mu_n\}$ that converges weakly to some measure $\mu' \in M^+(Y)$ with $\mu' \neq \mu$. This contradicts the uniqueness of μ because, by part (b), μ' is also an optimal solution for GC_∞ . $\square$

Remark 3.12. Suppose that $c_n \rightarrow c_\infty$ uniformly in Y . Then the following holds.

- If c_∞ is bounded away from zero, then so is c_n for all n sufficiently large. (This fact was used in the proof of Theorem 3.6(b).) Hence, in part (b) of Theorem 3.6 it suffices to require (only) that c_∞ is bounded away from zero, for both cases, bounded and unbounded.
- If the sequence $\{c_n\}$ is uniformly bounded away from zero (that is, there exists $\delta > 0$ such that, for each $n \in \mathbb{N}$, $c_n(y) \geq \delta$ for all $y \in Y$), then also c_∞ is bounded away from zero.

On the other hand, if $c_n \rightarrow c_\infty$ uniformly and $g_n \rightarrow g_\infty$ uniformly, then the following holds.

- If the Slater condition holds for GC_∞ (see Assumption 3.3(a)), then GC_n also satisfies the Slater condition for all n large enough. (This was shown to be the case in the proof of Lemma 3.11; see (3.8).) It follows that, for each $n \geq N$, GC_n is consistent, i.e., $\mathcal{F}_n \neq \emptyset$. Then Assumption 3.3 implies part (a) of Assumptions 3.1 and 3.2 for each $n \geq N$. Hence, in part (b) of Theorem 3.6, Assumptions 3.1(a) and 3.2(a) are not required.
- If the Slater condition uniformly holds for the sequence $\{\text{GC}_n\}$ (that is, for each $n \in \mathbb{N}$, there exist $\mu_n \in \mathcal{F}_\infty$ and $\eta > 0$ such that

$$\int_Y g_n(x, y) \mu_n(dy) \geq b(x) + \eta \quad \forall x \in X),$$

then the Slater condition also holds for GC_∞ .

REFERENCES

- [1] E. ALTMAN, *Constrained Markov Decision Processes*, Chapman & Hall/CRC, Boca Raton, FL, 1999.
- [2] E. J. ANDERSON, A. S. LEWIS, AND S. Y. WU, *The capacity problem*, Optimization, 20 (1989), pp. 725–742.
- [3] E. J. ANDERSON AND P. NASH, *Linear Programming in Infinite-Dimensional Spaces*, Wiley, Chichester, UK, 1987.
- [4] H. ATTOUCH, *Variational Convergence for Functions and Operators*, Pitman, Boston, 1984.
- [5] K. BACK, *Convergence of Lagrange multipliers and dual variables for convex optimization problems*, Math. Oper. Res., 13 (1988), pp. 74–79.
- [6] A. BALAYADI, Y. SONNTAG, AND C. ZALINESCU, *Stability of constrained optimization problems*, Nonlinear Anal. Theory Methods Appl., 28 (1997), pp. 1395–1409.
- [7] P. BILLINGSLEY, *Convergence of Probability Measures*, Wiley, New York, 1968.
- [8] G. DAL MASO, *An Introduction to Γ -Convergence*, Birkhäuser, Boston, 1993.
- [9] G. B. DANTZIG, J. FOLKMAN, AND N. SHAPIRO, *On the continuity of the minimum set of a continuous function*, J. Math. Anal. Appl., 17 (1967), pp. 519–548.
- [10] A. L. DONTCHEV AND T. ZOLEZZI, *Well-Posed Optimization Problems*, Lecture Notes in Math. 1543, Springer-Verlag, Berlin, 1993.
- [11] J. R. GABRIEL AND O. HERNÁNDEZ-LERMA, *Strong duality of the general capacity problem in metric spaces*, Math. Methods Oper. Res., 53 (2001), pp. 25–34.
- [12] M. A. GOBERNA AND M. A. LÓPEZ, *Linear Semi-infinite Optimization*, Wiley, New York, 1998.
- [13] O. HERNÁNDEZ-LERMA AND J. B. LASSERRE, *Further Topics on Discrete-Time Markov Control Processes*, Springer-Verlag, New York, 1999.
- [14] R. HETTICH AND K. O. KORTANEK, *Semi-infinite programming: Theory, methods, and applications*, SIAM Rev., 35 (1993), pp. 380–429.
- [15] P. KANNIAPPAN AND M. A. SUNDARAM, *Uniform convergence of convex optimization problems*, J. Math. Anal. Appl., 96 (1983), pp. 1–12.
- [16] K. KURATOWSKI, *Topology I*, Academic Press, New York, 1966.
- [17] I. E. SCHOCHETMAN, *Convergence of selections with applications in optimization*, J. Math. Anal. Appl., 155 (1991), pp. 278–292.
- [18] I. E. SCHOCHETMAN, *Pointwise versions of the maximum theorem with applications in optimization*, Appl. Math. Lett., 3 (1990), pp. 89–92.
- [19] M. M. TIDBALL, A. LOMBARDI, O. POURTALLIER, AND E. ALTMAN, *Continuity of optimal values and solutions for control of Markov chains with constraints*, SIAM J. Control Optim., 38 (2000), pp. 1204–1222.
- [20] A. M. VERSHIK AND V. TELMEL'T, *Some questions concerning the approximation of the optimal values of infinite-dimensional problems in linear programming*, Siberian Math. J., 9 (1968), pp. 591–601.

FARKAS-TYPE RESULTS WITH CONJUGATE FUNCTIONS*

RADU IOAN BOT[†] AND GERT WANKA[†]

Abstract. We present some new Farkas-type results for inequality systems involving a finite as well as an infinite number of convex constraints. For this, we use two kinds of conjugate dual problems, namely an extended Fenchel-type dual problem and the recently introduced Fenchel–Lagrange dual problem. For the latter, which is a “combination” of the classical Fenchel and Lagrange duals, the strong duality is established.

Key words. Farkas-type results, conjugate duality, finitely and infinitely many convex constraints

AMS subject classifications. 49N15, 90C25, 90C46

DOI. 10.1137/030602332

1. Introduction. The Farkas lemma [6] states that a linear inequality $c^T x \leq 0$ is a consequence of a linear system $Ax \leq 0$ if and only if c is a nonpositive linear combination of elements of A . This result has played an important role in the development of linear programming and optimization theory. During the last two decades a number of Farkas-type results have been given in the literature with applications to more general nonlinear programming problems and nonsmooth optimization problems (see [22], [11], [9], [10], [15], [8], [13], [18], [7], [14], [16], [12]).

In this paper we present some new Farkas-type results for inequality systems involving finitely as well as infinitely many convex constraints. The approach is based on the theory of conjugate duality for convex optimization problems. Two dual problems play an important role in our investigations, namely an extended Fenchel-type dual problem (see, for example, [17]) and the so-called Fenchel–Lagrange dual problem. The last one is a “combination” of the classical Fenchel and Lagrange dual problems and has been introduced and extensively studied recently by the authors of this paper (cf. [19], [2], [20], [21], [1], etc.). The construction of the Fenchel–Lagrange dual is described here in detail since we use a conjugacy approach which is based on perturbation theory (cf. [4]). Then we introduce a constraint qualification whose fulfillment is sufficient in order to guarantee strong duality. The strong duality assertion is also proved.

The results we are going to present here underline the connections that exist between Farkas-type results and alternative theorems and, on the other hand, the theory of duality. Furthermore, we bring some generalizations to some recently published results due to Jeyakumar (cf. [12]).

The paper is organized as follows. In section 2 we present definitions and preliminary results that will be used later in the paper. In section 3 we introduce a convex optimization problem and construct its Fenchel–Lagrange dual. Then we prove the existence of strong duality between these two problems. Section 4 provides some Farkas-type results, obtained by using the Fenchel–Lagrange dual problem, for inequality systems involving a finite number of convex constraints. Finally, in section 5 we give some Farkas-type results for inequality systems involving an infinite number

*Received by the editors December 9, 2003; accepted for publication (in revised form) June 2, 2004; published electronically February 19, 2005.

<http://www.siam.org/journals/siopt/15-2/60233.html>

[†]Faculty of Mathematics, Chemnitz University of Technology, D-09107 Chemnitz, Germany (radu.bot@mathematik.tu-chemnitz.de, gert.wanka@mathematik.tu-chemnitz.de).

of convex constraints. There we use an extended version of the classical Fenchel dual problem.

2. Notation and preliminaries. In this section we describe the notations we use throughout this paper and present preliminary results. All vectors will be column vectors. A column vector will be transposed to a row vector by an upper index T . The inner product of two vectors $x = (x_1, \dots, x_n)^T$ and $y = (y_1, \dots, y_n)^T$ in the n -dimensional real space $\mathbb{R}^n$ will be denoted by $x^T y = \sum_{i=1}^n x_i y_i$. For a set $X \subseteq \mathbb{R}^n$ we shall denote the *closure*, the *convex hull*, and the *relative interior* of X by $\text{cl}(X)$, $\text{co}(X)$, and $\text{ri}(X)$, respectively. Similarly, we shall denote the *cone* and the *convex cone* generated by the set X by $\text{cone}(X) = \bigcup_{\lambda \geq 0} \lambda X$, respectively, $\text{coneco}(X) = \bigcup_{\lambda \geq 0} \lambda \text{co}(X)$.

Furthermore, for the set $X \subseteq \mathbb{R}^n$, the *indicator function* $\delta_X : \mathbb{R}^n \rightarrow \overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ is defined by

$$\delta_X(x) = \begin{cases} 0 & \text{if } x \in X, \\ +\infty & \text{otherwise,} \end{cases}$$

while the *support function* $\sigma_X : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ is defined by

$$\sigma_X(p) = \sup_{x \in X} p^T x.$$

Considering now a function $f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$, we denote by

$$\text{dom}(f) = \{x \in \mathbb{R}^n : f(x) < +\infty\}$$

its *effective domain*, by

$$\text{epi}(f) = \{(x, r) \in \mathbb{R}^n \times \mathbb{R} : f(x) \leq r\}$$

its *epigraph*, and by $\text{cl}(f)$ its *closure*, namely the function whose epigraph is the closure of $\text{epi}(f)$ in $\mathbb{R}^{n+1}$. We say that f is proper if $\text{dom}(f) \neq \emptyset$ and $f(x) \neq -\infty \forall x \in \mathbb{R}^n$.

When X is a nonempty subset of $\mathbb{R}^n$ we define for f the so-called *conjugate relative to the set X*

$$f_X^* : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}, \quad f_X^*(p) = \sup_{x \in X} \{p^T x - f(x)\}.$$

When X equals the whole space $\mathbb{R}^n$, the conjugate relative to the set X becomes the classical *conjugate function of f* (the Fenchel–Moreau conjugate)

$$f^* : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}, \quad f^*(p) = \sup_{x \in \mathbb{R}^n} \{p^T x - f(x)\}.$$

Assuming that $f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ is a proper function, we have the following elementary result for its conjugate:

$$(2.1) \quad (\alpha f)^*(p) = \alpha f^*\left(\frac{1}{\alpha} p\right) \quad \forall p \in \mathbb{R}^n \quad \forall \alpha > 0.$$

Two results, which play an important role in this paper, follow. But first the following definition is necessary.

DEFINITION 2.1. Let the functions $f_1, \dots, f_m : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ be given. The function $f_1 \square \dots \square f_m : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$, defined by

$$f_1 \square \dots \square f_m(x) := \inf \left\{ \sum_{i=1}^m f_i(x_i) : \sum_{i=1}^m x_i = x \right\},$$

is called the infimal convolution function of $f_1, \dots, f_m$.

THEOREM 2.2 (cf. Theorem 16.4 in [17]). Let $f_1, \dots, f_m : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ be proper convex functions. Then

$$(\text{cl}(f_1) + \dots + \text{cl}(f_m))^* = \text{cl}(f_1^* \square \dots \square f_m^*).$$

If the sets $\text{ri}(\text{dom}(f_i))$, $i = 1, \dots, m$, have a point in common, then

$$\left(\sum_{i=1}^m f_i \right)^*(p) = (f_1^* \square \dots \square f_m^*)(p) = \inf \left\{ \sum_{i=1}^m f_i^*(p_i) : \sum_{i=1}^m p_i = p \right\},$$

where for each $p \in \mathbb{R}^n$ the infimum is attained.

A direct consequence of this theorem comes next.

COROLLARY 2.3. Let $f_1, \dots, f_m : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ be proper convex functions. Then

$$(2.2) \quad \text{epi} \left(\left(\sum_{i=1}^m \text{cl}(f_i) \right)^* \right) = \text{cl} \left(\sum_{i=1}^m \text{epi}(f_i^*) \right).$$

If the sets $\text{ri}(\text{dom}(f_i))$, $i = 1, \dots, m$, have a point in common, then

$$(2.3) \quad \text{epi} \left(\left(\sum_{i=1}^m f_i \right)^* \right) = \sum_{i=1}^m \text{epi}(f_i^*).$$

3. The Fenchel–Lagrange dual problem. In this section we assume that X is a nonempty convex subset of $\mathbb{R}^n$, $f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ is a proper convex function, and $g = (g_1, \dots, g_m)^T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a vector-valued function, with g_i also convex for $i = 1, \dots, m$. Using them we introduce the following convex optimization problem, further called the primal problem:

$$(P) \quad \inf_{x \in C} f(x),$$

$$C = \{x \in X : g(x) \leq 0\}.$$

As usual, $g(x) \leq 0$ means $g_i(x) \leq 0 \ \forall i = 1, \dots, m$.

In order to determine the so-called Fenchel–Lagrange dual problem of (P) , we need to introduce the perturbation function (cf. [19], [2], [20], [21], [1])

$$\Phi_{FL} : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \overline{\mathbb{R}}, \quad \Phi_{FL}(x, y, z) = \begin{cases} f(x+y) & \text{if } x \in X, \ g(x) \leq z, \\ +\infty & \text{otherwise,} \end{cases}$$

with the perturbation variables y and z . Following the path of the perturbation method described in [4] and [17], the next step is to calculate the conjugate function to Φ . Let us proceed with the definition for $\Phi_{FL}^* : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \overline{\mathbb{R}}$,

$$\begin{aligned} \Phi_{FL}^*(x^*, p, q) &= \sup_{\substack{x, y \in \mathbb{R}^n, \\ z \in \mathbb{R}^m}} \{x^{*T}x + p^T y + q^T z - \Phi(x, y, z)\} \\ &= \sup_{\substack{x \in X, y \in \mathbb{R}^n, \\ g(x) \leq z}} \{x^{*T}x + p^T y + q^T z - f(x+y)\}. \end{aligned}$$

In order to ease the calculations let us introduce the new variables r instead of y and s instead of z by

$$r := x + y, \quad s := z - g(x),$$

and therefore the supremum above can be separated into a sum of three suprema:

$$\begin{aligned} \Phi_{FL}^*(x^*, p, q) &= \sup_{s \in \mathbb{R}_+^m} q^T s + \sup_{r \in \mathbb{R}^n} \{p^T r - f(r)\} + \sup_{x \in X} \{(x^* - p)^T x + q^T g(x)\} \\ &= \begin{cases} f^*(p) - \inf_{x \in X} \{(p - x^*)^T x - q^T g(x)\} & \text{if } q \in -\mathbb{R}_+^m, \\ +\infty & \text{otherwise,} \end{cases} \end{aligned}$$

where $\mathbb{R}_+^m = \{z : z \in \mathbb{R}^m, 0 \leq z\}$.

According to [4] the dual problem to (P) is

$$(D_{FL}) \sup_{\substack{p \in \mathbb{R}^n, \\ q \in \mathbb{R}_+^m}} \{-\Phi_{FL}^*(0, p, q)\},$$

which becomes in our case, after changing the sign of the variable q ,

$$(D_{FL}) \sup_{\substack{p \in \mathbb{R}^n, \\ q \in \mathbb{R}_+^m}} \{-f^*(p) + \inf_{x \in X} [p^T x + q^T g(x)]\}$$

or, equivalently,

$$(3.1) \quad (D_{FL}) \sup_{\substack{p \in \mathbb{R}^n, \\ q \in \mathbb{R}_+^m}} \{-f^*(p) - (q^T g)_X^*(-p)\}.$$

The dual problem (D_{FL}) has been introduced by Wanka and Boş in [19] and has been extensively studied in [2], [20], [21], [1], etc. We call it the Fenchel–Lagrange dual problem because, as one may observe, it is a “combination” of the well-known Lagrange and Fenchel dual problems.

It is obvious from the construction of the dual that the weak duality assertion between (P) and (D_{FL}) , i.e., the value of the primal objective function at any feasible point is greater than or equal to the value of the dual objective function at any dual feasible point, always holds. By strong duality we understand the situation in which the optimal objective values of the primal and dual are equal and (D_{FL}) has an optimal solution. Unlike weak duality, strong duality can fail in the general case. To avoid this undesired situation, we introduce a constraint qualification that guarantees the validity of strong duality in case this constraint qualification is fulfilled. First let us divide the index set $\{1, \dots, m\}$ into two subsets,

$$L := \{i \in \{1, \dots, m\} : g_i : \mathbb{R}^n \rightarrow \mathbb{R} \text{ is an affine function}\}$$

and $N := \{1, \dots, m\} \setminus L$. The constraint qualification follows:

$$(CQ) \exists x' \in \text{ri}(X) \cap \text{ri}(\text{dom}(f)) : \begin{cases} g_i(x') \leq 0, & i \in L, \\ g_i(x') < 0, & i \in N. \end{cases}$$

We are ready now to formulate the strong duality assertion. Before that let us denote by $v(P)$ and $v(D_{FL})$ the optimal objective values of the primal and dual problem, respectively.

THEOREM 3.1. *Assume that $v(P) > -\infty$. Provided that the constraint qualification (CQ) is fulfilled, there is strong duality between problems (P) and (D_{FL}) ; i.e., their optimal objective values are equal and the dual problem has an optimal solution.*

Proof. We can write the problem (P) equivalently:

$$(P) \quad \inf_{\substack{x \in X \cap \text{dom}(f), \\ g(x) \leq 0}} f(x).$$

By Theorem 6.5 in [17], (CQ) yields

$$x' \in \text{ri}(X \cap \text{dom}(f)) = \text{ri}(X) \cap \text{ri}(\text{dom}(f)).$$

Theorem 5.7 in [5] states under the present hypotheses the existence of a $\bar{q} \geq 0$ such that the Lagrange duality holds, namely,

$$v(P) = \max_{q \geq 0} \inf_{x \in X \cap \text{dom}(f)} [f(x) + q^T g(x)] = \inf_{x \in X \cap \text{dom}(f)} [f(x) + \bar{q}^T g(x)].$$

Defining

$$h : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}, \quad h(x) = \begin{cases} \bar{q}^T g(x) & \text{if } x \in X, \\ +\infty & \text{if } x \notin X, \end{cases}$$

we can rewrite the right-hand side term in the following way:

$$v(P) = \inf_{x \in \mathbb{R}^n} [f(x) + h(x)].$$

Because $\text{ri}(\text{dom}(f)) \cap \text{ri}(\text{dom}(h)) = \text{ri}(\text{dom}(f)) \cap \text{ri}(X) \neq \emptyset$, by Theorem 31.1 (Fenchel's duality theorem) in [17], there exists a $\bar{p} \in \mathbb{R}^n$ such that this infimum is equal to

$$\begin{aligned} v(P) &= \max_{p \in \mathbb{R}^n} [-f^*(p) - h^*(-p)] = -f^*(\bar{p}) - h^*(-\bar{p}) \\ &= -f^*(\bar{p}) - \sup_{x \in \mathbb{R}^n} \{-\bar{p}^T x - h(x)\} \\ &= -f^*(\bar{p}) - \sup_{x \in X} \{-\bar{p}^T x - \bar{q}^T g(x)\} \\ (3.2) \quad &= -f^*(\bar{p}) - (\bar{q}^T g)_X^*(-\bar{p}). \end{aligned}$$

In the right-hand term of (3.2) one may recognize the objective function of (D_{FL}) at $(\bar{p}, \bar{q})$. From weak duality it follows that the supremum of (D_{FL}) is attained at $(\bar{p}, \bar{q})$, which is therefore an optimal solution of the dual problem. $\square$

Remark 3.2. Let us notice that in the proof above we have first proved that under the fulfillment of (CQ) there holds strong duality between the primal problem and its Lagrange dual problem. Then we proved the existence of strong duality between the inner infimum of the Lagrange dual at its outer maximum and its Fenchel dual problem, the last one proving to be exactly the Fenchel–Lagrange dual problem we introduced earlier.

4. Some new Farkas-type results for finitely many constraints. In the following we give some new Farkas-type results for inequality systems involving finitely many convex constraints. The main theorem yields a new dual characterization for this kind of result by using the duality concept introduced above. Moreover, it generalizes some recently published results due to Jeyakumar in [12].

Let X be again a nonempty convex subset of $\mathbb{R}^n$ and $I = \{1, \dots, m\}$ be an index set. As in the previous section, let $f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ be a proper and convex function and $g_i : \mathbb{R}^n \rightarrow \mathbb{R}$, $i \in I$, be real-valued convex functions. By $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ we denote the vector-valued function defined by $g(x) := (g_1(x), \dots, g_m(x))^T \forall x \in \mathbb{R}^n$. We formulate now the main result of this section.

THEOREM 4.1. *Let the constraint qualification (CQ) be fulfilled. Then the following statements are equivalent:*

- (i) $x \in X$, $g_i(x) \leq 0 \forall i \in I \Rightarrow f(x) \geq 0$.
- (ii) *There exist $p \in \mathbb{R}^n$ and $q \in \mathbb{R}^m$, $q \geq 0$, such that*

$$f^*(p) + (q^T g)_X^*(-p) \leq 0.$$

Proof. (ii) $\Rightarrow$ (i). Choose $p \in \mathbb{R}^n$ and $q \in \mathbb{R}^m$, $q \geq 0$, such that $f^*(p) + (q^T g)_X^*(-p) \leq 0$ or, equivalently, $-f^*(p) - (q^T g)_X^*(-p) \geq 0$. The optimal objective value $v(D_{FL})$ of the optimization problem

$$(D_{FL}) \sup_{\substack{p \in \mathbb{R}^n \\ q \geq 0}} \{-f^*(p) - (q^T g)_X^*(-p)\}$$

is greater than or equal to zero, which implies that the optimal objective value $v(P)$ of the problem (P) must be greater than or equal to zero, too. We recall that by weak duality the following inequality is true: $v(P) \geq v(D_{FL})$. Therefore $\forall x \in X$ such that $g_i(x) \leq 0 \forall i \in I$, we have $f(x) \geq 0$, and so (i) is fulfilled.

(i) $\Rightarrow$ (ii). Assuming now that (i) is true, it follows that the optimal objective value of the problem (P) is greater than or equal to zero. On the other hand, the constraint qualification (CQ) being fulfilled, we obtain by Theorem 3.1 that there exists (p, q) , $p \in \mathbb{R}^n$, $q \in \mathbb{R}^m$, $q \geq 0$, an optimal solution to (D_{FL}) , such that

$$v(P) = v(D_{FL}) = -f^*(p) - (q^T g)_X^*(-p) \geq 0.$$

This proves the validity of (ii). $\square$

Remark 4.2. For the implication (ii) $\Rightarrow$ (i) the constraint qualification (CQ) is not necessary.

As an immediate consequence of Theorem 4.1 we get the following theorem of the alternative.

COROLLARY 4.3. *Let the constraint qualification (CQ) be fulfilled. Then either the inequality system*

$$(I) \ x \in X, g_i(x) \leq 0 \forall i \in I, f(x) < 0$$

has a solution or the system

$$(II) \ f^*(p) + (q^T g)_X^*(-p) \leq 0, p \in \mathbb{R}^n, q \geq 0$$

has a solution but never both.

The next assertion offers an alternative formulation for the statement (ii) in Theorem 4.1.

PROPOSITION 4.4. *The statement (ii) in Theorem 4.1 is equivalent to*

$$(4.1) \quad 0 \in \text{epi}(f^*) + \text{coneco} \left(\bigcup_{i \in I} \text{epi}(g_i^*) \right) + \text{epi}(\sigma_X).$$

Proof. (ii) $\Rightarrow$ (4.1). Let $p \in \mathbb{R}^n$ and $q \in \mathbb{R}^m$, $q \geq 0$, be such that $f^*(p) + (q^T g)_X^*(-p) \leq 0$.

At first we assume that $q = 0$. This implies that $f^*(p) + \sup_{x \in X} (-p^T x) \leq 0$. Therefore $\sigma_X(-p) = \sup_{x \in X} (-p^T x) \leq -f^*(p)$, and so $(-p, -f^*(p)) \in \text{epi}(\sigma_X)$. This leads to the following relation:

$$\begin{aligned} 0 &= (p, f^*(p)) + (-p, -f^*(p)) \in \text{epi}(f^*) + \text{epi}(\sigma_X) \\ &\subseteq \text{epi}(f^*) + \text{coneco} \left(\bigcup_{i \in I} \text{epi}(g_i^*) \right) + \text{epi}(\sigma_X). \end{aligned}$$

When $q \neq 0$ the set of indices $I_q := \{i \in I : q_i \neq 0\}$ is not empty. On the other hand, because $f^*(p) + (q^T g)_X^*(-p) \leq 0$, there exists an $r \in \mathbb{R}$ such that

$$f^*(p) \leq r \leq -(q^T g)_X^*(-p),$$

which yields

$$(4.2) \quad (p, r) \in \text{epi}(f^*) \quad \text{and} \quad (-p, -r) \in \text{epi}((q^T g)_X^*).$$

From the definition of the conjugate function relative to the set X we have

$$\begin{aligned} (q^T g)_X^*(p) &= \sup_{x \in X} \{p^T x - q^T g(x)\} = \sup_{x \in \mathbb{R}^n} \{p^T x - (q^T g(x) + \delta_X(x))\} \\ &= (q^T g + \delta_X)^*(p) \quad \forall p \in \mathbb{R}^n, \end{aligned}$$

where δ_X is the indicator function of X . Obviously, $(q^T g)_X^* = (q^T g + \delta_X)^*$. Moreover, the fact that X is a convex set implies that the relative interior of $X = \text{dom}(\delta_X)$ is not empty. By Corollary 2.3 we have further

$$(4.3) \quad (-p, -r) \in \text{epi}((q^T g + \delta_X)^*) = \text{epi}((q^T g)^*) + \text{epi}(\delta_X^*).$$

By (2.1) and (2.3) we get

$$\text{epi}((q^T g)^*) = \text{epi} \left(\left(\sum_{i \in I_q} q_i g_i \right)^* \right) = \sum_{i \in I_q} \text{epi}((q_i g_i)^*) = \sum_{i \in I_q} q_i \text{epi}(g_i^*).$$

Therefore

$$\begin{aligned} \text{epi}((q^T g)^*) &= \sum_{i \in I_q} q_i \text{epi}(g_i^*) = \left(\sum_{i \in I_q} q_i \right) \sum_{i \in I_q} \frac{q_i}{\sum_{i \in I_q} q_i} \text{epi}(g_i^*) \\ &\subseteq \left(\sum_{i \in I_q} q_i \right) \text{co} \left(\bigcup_{i \in I} \text{epi}(g_i^*) \right) \subseteq \text{coneco} \left(\bigcup_{i \in I} \text{epi}(g_i^*) \right). \end{aligned}$$

So, by (4.2) and (4.3) and because $\delta_X^* = \sigma_X$, we obtain

$$0 \in \text{epi}(f^*) + \text{coneco} \left(\bigcup_{i \in I} \text{epi}(g_i^*) \right) + \text{epi}(\sigma_X).$$

(4.1) $\Rightarrow$ (ii). Now let be $p \in \mathbb{R}^n$ and $r \in \mathbb{R}$ such that $(p, r) \in \text{epi}(f^*)$ and $(-p, -r) \in \text{coneco}(\bigcup_{i \in I} \text{epi}(g_i^*)) + \text{epi}(\sigma_X)$. Further, there exists $\lambda \geq 0$ such that $(-p, -r) \in \lambda \text{co}(\bigcup_{i \in I} \text{epi}(g_i^*)) + \text{epi}(\sigma_X)$.

If $\lambda = 0$, then $(-p, -r) \in \text{epi}(\sigma_X) = \text{epi}(\delta_X^*)$ and, by taking $q \in \mathbb{R}^m$, $q := 0$, we get $(q^T g)_X^*(-p) = (0)_X^*(-p) = \sigma_X(-p) \leq -r$. It follows that $f^*(p) + (q^T g)_X^*(-p) \leq r - r = 0$.

Assuming now that $\lambda > 0$, there exist some $\mu_i \geq 0$, $i = 1, \dots, m$, $\sum_{i=1}^m \mu_i = 1$, such that $(-p, -r) \in \lambda \sum_{i=1}^m \mu_i \text{epi}(g_i^*) + \text{epi}(\sigma_X)$. Let us consider now the following set of indices $I_\mu := \{i \in I : \mu_i \neq 0\}$. The relation above can be written as

$$\begin{aligned} (-p, -r) &\in \sum_{i \in I_\mu} \lambda \mu_i \text{epi}(g_i^*) + \text{epi}(\delta_X^*) = \sum_{i \in I_\mu} \text{epi}((\lambda \mu_i g_i)^*) + \text{epi}(\delta_X^*) \\ &= \text{epi} \left(\left(\sum_{i \in I_\mu} \lambda \mu_i g_i \right)^* \right) + \text{epi}(\delta_X^*) = \text{epi} \left(\left(\sum_{i=1}^m \lambda \mu_i g_i \right)^* \right) + \text{epi}(\delta_X^*) \\ &= \text{epi} \left(\left(\sum_{i=1}^m \lambda \mu_i g_i + \delta_X \right)^* \right) = \text{epi}((q^T g)_X^*) \end{aligned}$$

for $q := (\lambda \mu_1, \dots, \lambda \mu_m)^T$. Therefore we found a pair $(p, q) \in \mathbb{R}^n \times \mathbb{R}_+^m$ such that $f^*(p) + (q^T g)_X^*(-p) \leq r - r = 0$. $\square$

COROLLARY 4.5. *Let $u \in \mathbb{R}^n$, let $\alpha \in \mathbb{R}$, and let us assume that there exists $x' \in \text{ri}(X)$ such that $g_i(x') \leq 0 \ \forall i \in L$ and $g_i(x') < 0 \ \forall i \in N$. Then the following statements are equivalent:*

- (i) $x \in X$, $g_i(x) \leq 0 \ \forall i \in I \Rightarrow u^T x \leq \alpha$.
- (ii) There exists $q \in \mathbb{R}^m$, $q \geq 0$, such that $(q^T g)_X^*(u) \leq \alpha$.

Proof. Considering $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $f(x) = \alpha - u^T x$, we get

$$f^*(p) = \begin{cases} -\alpha & \text{if } p = -u, \\ +\infty & \text{otherwise.} \end{cases}$$

The equivalence follows from Theorem 4.1. $\square$

Remark 4.6. For $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $f(x) = \alpha - u^T x$, the epigraph of f^* is nothing but the set $\{(-u, -\alpha)\} + \{0\} \times \mathbb{R}_+$. By Proposition 4.4, the statement (ii) in Corollary 4.5 can be equivalently written as

$$(u, \alpha) \in \text{coneco} \left(\bigcup_{i \in I} \text{epi}(g_i^*) \right) + \text{epi}(\sigma_X) + \{0\} \times \mathbb{R}_+,$$

and, because $\text{epi}(\sigma_X) + \{0\} \times \mathbb{R}_+ = \text{epi}(\sigma_X)$, we get

$$(u, \alpha) \in \text{coneco} \left(\bigcup_{i \in I} \text{epi}(g_i^*) \right) + \text{epi}(\sigma_X).$$

Remark 4.7. If we reconsider Corollary 4.5 for $X = \mathbb{R}^n$, the following statements turn out to be equivalent:

- (i) $x \in \mathbb{R}^n$, $g_i(x) \leq 0 \ \forall i \in I \Rightarrow u^T x \leq \alpha$.
- (ii) There exists $q \in \mathbb{R}^m$, $q \geq 0$, such that $(q^T g)^*(u) \leq \alpha$.

Because $\text{epi}(\sigma_{\mathbb{R}^n}) = \{0\} \times \mathbb{R}_+$, both (i) and (ii) are equivalent to

$$(4.4) \quad (u, \alpha) \in \text{coneco} \left(\bigcup_{i \in I} \text{epi}(g_i^*) \right) + \{0\} \times \mathbb{R}_+.$$

Assuming now that $u \neq 0$, relation (4.4) can be further equivalently written as

$$(u, \alpha) \in \text{coneco} \left(\bigcup_{i \in I} \text{epi}(g_i^*) \right),$$

which is exactly the result obtained in [12]. Thus our results extend those of Jeyakumar, because the equivalences hold here under much weaker conditions, namely, even if the set $\{x \in \mathbb{R}^n : g_i(x) < 0 \ \forall i \in I\}$ is empty.

Readers interested in the applicability of the Farkas-type results presented in this section are referred to the forthcoming paper [3]. There we characterize the containment of a nonempty polyhedral set in an arbitrary polyhedral set, in a reverse-polyhedral set and in a reverse-convex set determined by convex quadratic constraints. Moreover, some famous theorems of the alternative are rediscovered as special cases of our results.

5. Some new Farkas-type results for infinitely many constraints. In this section we consider Farkas-type results for inequality systems involving infinitely many convex constraints. To achieve them we use an extended Fenchel-type dual problem for which strong duality holds, provided that a regularity condition is fulfilled. Under the same hypotheses we give then equivalent formulations by using the epigraphs of the conjugates of the functions involved. Similar results obtained by using epigraphs are stated next, this time under closedness assumptions.

Let I be an arbitrary index set, let $g_i : \mathbb{R}^n \rightarrow \mathbb{R}$ be convex functions for $i \in I$, and let G denote the set $\{x \in \mathbb{R}^n : g_i(x) \leq 0 \ \forall i \in I\}$. For $X \subseteq \mathbb{R}^n$ a nonempty convex set and $f : \mathbb{R}^n \rightarrow \mathbb{R}$ a proper convex function, let us consider the primal optimization problem

$$(P^\infty) \quad \inf_{x \in C} f(x),$$

$$C := X \cap G = \{x \in X : g_i(x) \leq 0 \ \forall i \in I\}.$$

Moreover, let us assume that the set C is not empty.

By using the indicator functions δ_G and δ_X of the sets G and X , respectively, we may formulate (P^∞) as follows:

$$(P^\infty) \quad \inf_{x \in \mathbb{R}^n} (f(x) + \delta_G(x) + \delta_X(x)) = -(f + \delta_G + \delta_X)^*(0).$$

If $\text{ri}(\text{dom}(f)) \cap \text{ri}(G) \cap \text{ri}(X) \neq \emptyset$, then, by Theorem 2.2, it follows that the extended Fenchel-type dual problem to (P^∞) ,

$$(D_F^\infty) \quad \sup_{p_1, p_2 \in \mathbb{R}^n} \{-f^*(p_1 + p_2) - \delta_G^*(-p_1) - \delta_X^*(-p_2)\},$$

has an optimal solution and the optimal objective values of (P^∞) and (D_F^∞) coincide, i.e., $v(P^\infty) = v(D_F^\infty)$.

Remark 5.1. The extended Fenchel-type dual problem (D_F^∞) can also be obtained using the approach described in section 3 by considering as perturbation function

$$\Phi_F : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \overline{\mathbb{R}}, \quad \Phi_F(x, y) = \begin{cases} f(x + y) & \text{if } x + z \in G, x \in X, \\ +\infty & \text{otherwise,} \end{cases}$$

with the perturbation variables y and z . The relations between the classical Fenchel, Lagrange and the Fenchel–Lagrange dual problems have been investigated in [19] and [2].

By using the extended Fenchel-type dual problem we can prove the following Farkas-type result.

THEOREM 5.2. *Assume that $\text{ri}(\text{dom}(f)) \cap \text{ri}(G) \cap \text{ri}(X) \neq \emptyset$. Then the following statements are equivalent:*

(i) $x \in X, g_i(x) \leq 0 \ \forall i \in I \Rightarrow f(x) \geq 0$.

(ii) *There exist $p_1, p_2 \in \mathbb{R}^n$ such that $f^*(p_1 + p_2) + \delta_G^*(-p_1) + \delta_X^*(-p_2) \leq 0$.*

Proof. (ii) $\Rightarrow$ (i). Let $p_1, p_2 \in \mathbb{R}^n$ such that $f^*(p_1 + p_2) + \delta_G^*(-p_1) + \delta_X^*(-p_2) \leq 0$. The optimal objective value $v(D_F^\infty)$ of the extended Fenchel-type dual problem

$$(D_F^\infty) \sup_{p_1, p_2 \in \mathbb{R}^n} \{-f^*(p_1 + p_2) - \delta_G^*(-p_1) - \delta_X^*(-p_2)\}$$

is greater than or equal to zero. By weak duality we have that the optimal objective value $v(P^\infty)$ of the primal problem (P^∞) must also be greater than or equal to zero. This implies that $\forall x \in C, f(x) \geq 0$, and so (i) is fulfilled.

(i) $\Rightarrow$ (ii). Assuming now that (i) is true it follows that $v(P^\infty) \geq 0$. The regularity condition being fulfilled, Theorem 2.2 guarantees the existence of (p_1, p_2) , an optimal solution to (D_F^∞) , such that

$$v(P^\infty) = -(f + \delta_G + \delta_X)^*(0) = -f^*(p_1 + p_2) - \delta_G^*(-p_1) - \delta_X^*(-p_2) = v(D_F^\infty) \geq 0.$$

Therefore (i) is also fulfilled. $\square$

Remark 5.3. For the implication (ii) $\Rightarrow$ (i) the condition $\text{ri}(\text{dom}(f)) \cap \text{ri}(G) \cap \text{ri}(X) \neq \emptyset$ is not necessary.

Let us reformulate now Theorem 5.2 as a theorem of the alternative.

COROLLARY 5.4. *Assume that the regularity condition in Theorem 5.2 is fulfilled. Then either the inequality system*

$$(I) \ x \in X, g_i(x) \leq 0 \ \forall i \in I, f(x) < 0$$

has a solution or the system

$$(II) \ f^*(p_1 + p_2) + \delta_G^*(-p_1) + \delta_X^*(-p_2) \leq 0, p_1 \in \mathbb{R}^n, p_2 \in \mathbb{R}^n$$

has a solution but never both.

The next result provides an equivalent characterization of the statement (ii) in Theorem 5.2.

PROPOSITION 5.5. *The statement (ii) in Theorem 5.2 is equivalent to*

$$(5.1) \quad 0 \in \text{epi}(f^*) + \text{cl} \left(\text{coneco} \left(\bigcup_{i \in I} \text{epi}(g_i^*) \right) \right) + \text{epi}(\sigma_X).$$

Proof. The statement (ii) in Theorem 5.2 can equivalently be written as follows.

There exist $p_1, p_2 \in \mathbb{R}^n$, $r \in \mathbb{R}$, and $s \in \mathbb{R}$ such that $f^*(p_1 + p_2) \leq -r - s$, $\delta_G^*(-p_1) \leq r$, and $\delta_X^*(-p_2) \leq s \Leftrightarrow$ there exist $p_1, p_2 \in \mathbb{R}^n$, $r \in \mathbb{R}$, and $s \in \mathbb{R}$ such that $(p_1 + p_2, -r - s) \in \text{epi}(f^*)$, $(-p_1, r) \in \text{epi}(\delta_G^*)$, and $(-p_2, s) \in \text{epi}(\delta_X^*) \Leftrightarrow$

$$0 \in \text{epi}(f^*) + \text{epi}(\delta_G^*) + \text{epi}(\delta_X^*) = \text{epi}(f^*) + \text{epi}(\sigma_G) + \text{epi}(\sigma_X).$$

Using that $\text{epi}(\sigma_G) = \text{cl}(\text{coneco}(\bigcup_{i \in I} \text{epi}(g_i^*)))$ (see, for example, [14], [16], [12]), we get (5.1). $\square$

The next theorem provides a similar characterization to (5.1), this time under some closedness assumptions.

THEOREM 5.6. *Alongside the initial hypotheses we assume that X is a closed set and f is a lower semicontinuous function. Then the following statements are equivalent:*

- (i) $x \in X$, $g_i(x) \leq 0 \ \forall i \in I \Rightarrow f(x) \geq 0$.
- (ii) $0 \in \text{cl}(\text{epi}(f^*) + \text{coneco}(\bigcup_{i \in I} \text{epi}(g_i^*)) + \text{epi}(\sigma_X))$.

Proof. As we have seen, statement (i) is rewritable as

$$\inf_{x \in \mathbb{R}^n} (f(x) + \delta_G(x) + \delta_X(x)) = -(f + \delta_G + \delta_X)^*(0) \geq 0$$

or, equivalently,

$$0 \in \text{epi}((f + \delta_G + \delta_X)^*).$$

The sets X and $G = \bigcap_{i \in I} \{x \in \mathbb{R}^n : g_i(x) \leq 0\}$ are closed and therefore the functions δ_G and δ_X are lower semicontinuous. Thus by (2.2) in Corollary 2.3 we have

$$\begin{aligned} \text{epi}((f + \delta_G + \delta_X)^*) &= \text{epi}((\text{cl}(f) + \text{cl}(\delta_G) + \text{cl}(\delta_X))^*) \\ &= \text{cl}(\text{epi}(f^*) + \text{epi}(\delta_G^*) + \text{epi}(\delta_X^*)) \\ &= \text{cl}(\text{epi}(f^*) + \text{epi}(\sigma_G) + \text{epi}(\sigma_X)) \\ (5.2) \quad &= \text{cl} \left(\text{epi}(f^*) + \text{cl} \left(\text{coneco} \left(\bigcup_{i \in I} \text{epi}(g_i^*) \right) \right) + \text{epi}(\sigma_X) \right). \end{aligned}$$

Since $\text{cl}(S_1 + \text{cl}(S_2)) = \text{cl}(S_1 + S_2)$ for arbitrary sets S_1 and S_2 in $\mathbb{R}^n$, we get

$$\begin{aligned} 0 &\in \text{cl} \left(\text{epi}(f^*) + \text{cl} \left(\text{coneco} \left(\bigcup_{i \in I} \text{epi}(g_i^*) \right) \right) + \text{epi}(\sigma_X) \right) \\ &= \text{cl} \left(\text{epi}(f^*) + \text{coneco} \left(\bigcup_{i \in I} \text{epi}(g_i^*) \right) + \text{epi}(\sigma_X) \right). \quad \square \end{aligned}$$

Remark 5.7. If $X = \mathbb{R}^n$ and $f : \mathbb{R}^n \rightarrow \mathbb{R}$, the assumptions in both Theorems 5.2 and 5.6 are valid; namely,

- (i) $\text{ri}(\text{dom}(f)) \cap \text{ri}(G) \cap \text{ri}(X) \neq \emptyset$,
- (ii) X is closed and f is lower semicontinuous.

Having

$$\text{epi}(f^*) + \text{epi}(\sigma_{\mathbb{R}^n}) = \text{epi}(f^*) + \{0\} \times \mathbb{R}_+ = \text{epi}(f^*),$$

the statement

$$(i) \ x \in \mathbb{R}^n, g_i(x) \leq 0 \ \forall i \in I \Rightarrow f(x) \geq 0$$

turns out to be equivalent to

$$0 \in \text{cl} \left(\text{epi}(f^*) + \text{coneco} \left(\bigcup_{i \in I} \text{epi}(g_i^*) \right) \right) = \text{epi}(f^*) + \text{cl} \left(\text{coneco} \left(\bigcup_{i \in I} \text{epi}(g_i^*) \right) \right),$$

which is nothing else than Theorem 4.1 in [12].

The next theorem provides a Farkas-type result involving a difference of convex functions.

THEOREM 5.8. *Let $h : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ be a proper convex function with the property that C is a subset of $\text{ri}(\text{dom}(h))$. Assume that $\text{ri}(\text{dom}(f)) \cap \text{ri}(G) \cap \text{ri}(X) \neq \emptyset$. Then the following statements are equivalent:*

- (i) $x \in X, g_i(x) \leq 0 \ \forall i \in I \Rightarrow h(x) - f(x) \leq 0$.
- (ii) $\forall u \in \mathbb{R}^n$ there exists $p_1^u, p_2^u \in \mathbb{R}^n$ such that $f^*(u + p_1^u + p_2^u) + \delta_G^*(-p_1^u) + \delta_X^*(-p_2^u) \leq h^*(u)$.

Proof. By Theorem 7.4 and Theorem 12.2 in [17] we have that $\forall x \in C, h(x) = \text{cl}(h)(x) = h^{**}(x) = \sup_{u \in \mathbb{R}^n} \{u^T x - h^*(u)\}$. This implies that (i) is rewritable as

$$(5.3) \quad \forall u \in \mathbb{R}^n \ \forall x \in C \quad f(x) - u^T x + h^*(u) \geq 0.$$

Applying Theorem 5.2 we obtain, equivalently, that $\forall u \in \mathbb{R}^n$ there exist $p_1^u, p_2^u \in \mathbb{R}^n$ (even if $h^*(u) = +\infty$) such that

$$\begin{aligned} \sup_{x \in \mathbb{R}^n} \{ (p_1^u + p_2^u)^T x - (f(x) - u^T x + h^*(u)) \} + \delta_G^*(-p_1^u) + \delta_X^*(-p_2^u) &\leq 0 \\ \Leftrightarrow f^*(u + p_1^u + p_2^u) + \delta_G^*(-p_1^u) + \delta_X^*(-p_2^u) &\leq h^*(u). \quad \square \end{aligned}$$

PROPOSITION 5.9. *The statement (ii) in Theorem 5.8 is equivalent to*

$$(5.4) \quad \text{epi}(h^*) \subseteq \text{epi}(f^*) + \text{cl} \left(\text{coneco} \left(\bigcup_{i \in I} \text{epi}(g_i^*) \right) \right) + \text{epi}(\sigma_X).$$

Proof. (ii) $\Rightarrow$ (5.4). Let $(u, r) \in \text{epi}(h^*) \Leftrightarrow h^*(u) \leq r$. By (ii) in Theorem 5.8 there exist $p_1^u, p_2^u \in \mathbb{R}^n$ such that

$$r \geq h^*(u) \geq f^*(u + p_1^u + p_2^u) + \delta_G^*(-p_1^u) + \delta_X^*(-p_2^u).$$

Therefore

$$\begin{aligned} (u, r) &= (u + p_1^u + p_2^u, r - \delta_G^*(-p_1^u) - \delta_X^*(-p_2^u)) + (-p_1^u, \delta_G^*(-p_1^u)) + (-p_2^u, \delta_X^*(-p_2^u)) \\ &\in \text{epi}(f^*) + \text{epi}(\delta_G^*) + \text{epi}(\delta_X^*) = \text{epi}(f^*) + \text{cl} \left(\text{coneco} \left(\bigcup_{i \in I} \text{epi}(g_i^*) \right) \right) + \text{epi}(\sigma_X), \end{aligned}$$

and so (5.4) is verified.

(5.4) $\Rightarrow$ (ii). Let $u \in \mathbb{R}^n$ be fixed. If $h^*(u) = +\infty$, then (ii) is true. Assuming that $h^*(u) < +\infty$, we have $(u, h^*(u)) \in \text{epi}(h^*)$, and so there exist $(a, s) \in \text{epi}(f^*)$,

$(b, t) \in \text{cl}(\text{coneco}(\bigcup_{i \in I} \text{epi}(g_i^*))) = \text{epi}(\delta_G^*)$ and $(c, z) \in \text{epi}(\sigma_X)$ such that $a + b + c = u$ and $s + t + z = h^*(u)$. Taking $p_1^u := -b$ and $p_2^u := -c$, we get

$$f^*(u + p_1^u + p_2^u) + \delta_G^*(-p_1^u) + \delta_X^*(-p_2^u) = f^*(a) + \delta_G^*(b) + \delta_X^*(c) \leq s + t + z = h^*(u). \quad \square$$

We consider now in the hypotheses of Theorem 5.8 instead of the regularity condition $\text{ri}(\text{dom}(f)) \cap \text{ri}(G) \cap \text{ri}(X) \neq \emptyset$, the assumptions that X is a closed set and f is a lower semicontinuous function. This leads us to the following result.

THEOREM 5.10. *Let $h : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ be a proper convex function with the property that C is a subset of $\text{ri}(\text{dom}(h))$. Assume that X is a closed set and f is a lower semicontinuous function. Then the following statements are equivalent:*

- (i) $x \in X, g_i(x) \leq 0 \forall i \in I \Rightarrow h(x) - f(x) \leq 0$.
- (ii) $\text{epi}(h^*) \subseteq \text{cl}(\text{epi}(f^*) + \text{coneco}(\bigcup_{i \in I} \text{epi}(g_i^*)) + \text{epi}(\sigma_X))$.

Proof. As we have seen in the proof of Theorem 5.8, the statement (i) can be equivalently written as (cf. (5.3))

$$\begin{aligned} \forall u \in \mathbb{R}^n \forall x \in \mathbb{R}^n \quad & f(x) + \delta_G(x) + \delta_X(x) - u^T x + h^*(u) \geq 0 \\ \Leftrightarrow \forall u \in \mathbb{R}^n \forall x \in \mathbb{R}^n \quad & h^*(u) \geq u^T x - (f + \delta_G + \delta_X)(x) \\ \Leftrightarrow \forall u \in \mathbb{R}^n \quad & h^*(u) \geq (f + \delta_G + \delta_X)^*(u) \Leftrightarrow \text{epi}(h^*) \subseteq \text{epi}((f + \delta_G + \delta_X)^*). \end{aligned}$$

The desired conclusion follows using that (cf. (5.2))

$$\text{epi}((f + \delta_G + \delta_X)^*) = \text{cl} \left(\text{epi}(f^*) + \text{coneco} \left(\bigcup_{i \in I} \text{epi}(g_i^*) \right) + \text{epi}(\sigma_X) \right). \quad \square$$

Remark 5.11. For $X = \mathbb{R}^n$ and $f : \mathbb{R}^n \rightarrow \mathbb{R}$ the assumptions in Theorems 5.8 and 5.10 are again valid. If $G \subseteq \text{ri}(\text{dom}(h))$, the statement

- (i) $x \in \mathbb{R}^n, g_i(x) \leq 0 \forall i \in I \Rightarrow h(x) - f(x) \leq 0$

becomes equivalent to

$$\begin{aligned} \text{epi}(h^*) &\subseteq \text{cl} \left(\text{epi}(f^*) + \text{coneco} \left(\bigcup_{i \in I} \text{epi}(g_i^*) \right) \right) \\ &= \text{epi}(f^*) + \text{cl} \left(\text{coneco} \left(\bigcup_{i \in I} \text{epi}(g_i^*) \right) \right). \end{aligned}$$

In case $h : \mathbb{R}^n \rightarrow \mathbb{R}$, this result is nothing else than Theorem 4.2 in [12] (see also Theorems 3.1 and 3.2 in [16]).

For the last part of this section we consider under the same hypotheses as for Theorem 5.8 that $f : \mathbb{R}^n \rightarrow \mathbb{R}, f(x) = 0 \forall x \in \mathbb{R}^n$. The above-mentioned theorem leads us to the following result.

THEOREM 5.12. *Let $h : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ be a proper convex function with the property that $C \subseteq \text{ri}(\text{dom}(h))$. Assume that $\text{ri}(G) \cap \text{ri}(X) \neq \emptyset$. Then the following statements are equivalent:*

- (i) $x \in X, g_i(x) \leq 0 \forall i \in I \Rightarrow h(x) \leq 0$.
- (ii) $\forall u \in \mathbb{R}^n$ there exists $p^u \in \mathbb{R}^n$ such that $h^*(u) \geq \delta_G^*(u + p^u) + \delta_X^*(-p^u)$.

Because $\text{epi}(f^*) = \{0\} \times \mathbb{R}_+$, the statement (i) in Theorem 5.12 can be equivalently written as (cf. (5.4))

$$(5.5) \quad \text{epi}(h^*) \subseteq \text{cl} \left(\text{coneco} \left(\bigcup_{i \in I} \text{epi}(g_i^*) \right) \right) + \text{epi}(\sigma_X).$$

By the same argument, Theorem 5.10 can be formulated as follows.

THEOREM 5.13. *Let $h : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ be a proper convex function with the property that C is a subset of $\text{ri}(\text{dom}(h))$. Assume that X is a closed set. Then the following statements are equivalent:*

- (i) $x \in X, g_i(x) \leq 0 \ \forall i \in I \Rightarrow h(x) \leq 0$.
- (ii) $\text{epi}(h^*) \subseteq \text{cl}(\text{coneco}(\bigcup_{i \in I} \text{epi}(g_i^*)) + \text{epi}(\sigma_X))$.

For $X = \mathbb{R}^n$, we have $\text{epi}(\sigma_X) = \{0\} \times \mathbb{R}_+$ and therefore (5.5) and (ii) in Theorem 5.13 become

$$(5.6) \quad \text{epi}(h^*) \subseteq \text{cl} \left(\text{coneco} \left(\bigcup_{i \in I} \text{epi}(g_i^*) \right) \right).$$

In case $h : \mathbb{R}^n \rightarrow \mathbb{R}$, (5.6) is nothing else than Corollary 4.3 in [12].

6. Conclusion. In this paper we present some new Farkas-type results for inequality systems involving finitely as well as infinitely many convex constraints. Our approach, which employs the conjugates of the functions involved, is based on the theory of duality for convex optimization problems. An important role is played by an extended Fenchel-type dual problem as well as by the Fenchel–Lagrange dual problem. The last one has been introduced and extensively studied in the last years by the authors of this paper. The results we formulate and prove here generalize some recently published results due to Jeyakumar in [12]. Moreover, they underline the connections that exist between Farkas-type results and alternative theorems and, on the other hand, the theory of duality.

REFERENCES

- [1] R. I. BOŢ, S. M. GRAD, AND G. WANKA, *Fenchel-Lagrange versus Geometric Duality in Convex Optimization*, Preprint 2/2005, Chemnitz University of Technology, Faculty of Mathematics, Chemnitz, Germany, 2005.
- [2] R. I. BOŢ, G. KASSAY, AND G. WANKA, *Strong duality for generalized convex optimization problems*, J. Optim. Theory Appl., to appear.
- [3] R. I. BOŢ AND G. WANKA, *Farkas-Type Results for Max-Functions and Applications*, Preprint 16/2004, Chemnitz University of Technology, Faculty of Mathematics, Chemnitz, Germany, 2004.
- [4] I. EKELAND AND R. TEMAM, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.
- [5] K. H. ELSTER, R. REINHARDT, M. SCHÄUBLE, AND G. DONATH, *Einführung in die Nichtlineare Optimierung*, Teubner Verlag, Leipzig, 1977.
- [6] J. FARKAS, *Theorie der einfachen Ungleichungen*, J. Reine Angew. Math., 124 (1901), pp. 1–27.
- [7] B. M. GLOVER, Y. ISHIZUKA, V. JEYAKUMAR, AND H. D. TUAN, *Complete characterizations of global optimality for problems involving the pointwise minimum of sublinear functions*, SIAM J. Optim., 6 (1996), pp. 362–372.
- [8] B. M. GLOVER, V. JEYAKUMAR, AND W. OETTLI, *A Farkas lemma for difference sublinear systems and quasidifferentiable programming*, Math. Programming, 63 (1994), pp. 109–125.
- [9] J. GWINNER, *Results of Farkas type*, Numer. Funct. Anal. Optim., 9 (1987), pp. 471–520.
- [10] J. GWINNER, *Corrigendum and addendum to “Results of Farkas type,”* Numer. Funct. Anal. Optim., 10 (1989), pp. 415–418.

- [11] C. W. HA, *On systems of convex inequalities*, J. Math. Anal. Appl., 68 (1979), pp. 25–34.
- [12] V. JEYAKUMAR, *Characterizing set containments involving infinite convex constraints and reverse-convex constraints*, SIAM J. Optim., 13 (2003), pp. 947–959.
- [13] V. JEYAKUMAR AND B. M. GLOVER, *Nonlinear extensions of Farkas’ lemma with applications to global optimization and least squares*, Math. Oper. Res., 20 (1995), pp. 818–837.
- [14] V. JEYAKUMAR AND B. M. GLOVER, *Characterizing global optimality for DC optimization problems under convex inequality constraints*, J. Global Optim., 8 (1996), pp. 171–187.
- [15] V. JEYAKUMAR AND J. GWINNER, *Inequality systems and optimization*, J. Math. Anal. Appl., 159 (1991), pp. 51–71.
- [16] V. JEYAKUMAR, A. M. RUBINOV, B. M. GLOVER, AND Y. ISHIZUKA, *Inequality systems and global optimization*, J. Math. Anal. Appl., 202 (1996), pp. 900–919.
- [17] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [18] A. M. RUBINOV, B. M. GLOVER, AND V. JEYAKUMAR, *A general approach to dual characterizations of solvability of inequality systems with applications*, J. Convex Anal., 2 (1995), pp. 309–344.
- [19] G. WANKA AND R. I. BOT, *On the relations between different dual problems in convex mathematical programming*, in Operations Research Proceedings 2001, P. Chamoni, R. Leisten, A. Martin, J. Minnemann, and H. Stadtler, eds., Springer-Verlag, Berlin, 2002, pp. 255–262.
- [20] G. WANKA AND R. I. BOT, *A new duality approach for multiobjective convex optimization problems*, J. Nonlinear Convex Anal., 3 (2002), pp. 41–57.
- [21] G. WANKA AND R. I. BOT, *Multiobjective duality for convex ratios*, J. Math. Anal. Appl., 275 (2002), pp. 354–368.
- [22] C. ZĂLINESCU, *A generalization of the Farkas lemma with applications to convex programming*, J. Math. Anal. Appl., 66 (1978), pp. 651–678.

PROXIMAL POINT ALGORITHM CONTROLLED BY A SLOWLY VANISHING TERM: APPLICATIONS TO HIERARCHICAL MINIMIZATION*

ALEXANDRE CABOT†

Abstract. Let $\Phi_0 : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ be a closed convex function and $\Phi_1 : \mathbb{R}^n \rightarrow \mathbb{R}$ be a finite convex function that are bounded from below. Our goal is to build an algorithm which first minimizes the map Φ_0 and secondly the map Φ_1 over the set $S_0 := \operatorname{argmin} \Phi_0$. For that purpose, we define the following proximal-type algorithm:

$$(\mathcal{A}_1) \quad -(x_{n+1} - x_n)/\lambda_n \in \partial_{\eta_n}(\Phi_0 + \varepsilon_n \Phi_1)(x_{n+1}),$$

where (λ_n) is a positive step sequence, (η_n) is a summable error sequence, and (ε_n) is a control sequence tending toward 0; ∂_η denotes the η -approximate subdifferential. When (ε_n) is a slow control, i.e., $\sum_{n=0}^{+\infty} \varepsilon_n = +\infty$, we prove that, under adequate conditions, the sequence (x_n) defined by $(\mathcal{A}_1)$ tends toward an element of $S_1 := \operatorname{argmin}_{S_0} \Phi_1$.

More generally, given finite convex functions $\Phi_2, \dots, \Phi_N : \mathbb{R}^n \rightarrow \mathbb{R}$, let us define the sets $(S_i)_{i \in \{1, \dots, N\}}$ by the recursive relation $S_i := \operatorname{argmin}_{S_{i-1}} \Phi_i$. We introduce an extension of algorithm $(\mathcal{A}_1)$ to minimize hierarchically each function Φ_i on the set S_{i-1} , for $i \in \{1, \dots, N\}$.

Key words. steepest descent system, proximal point method, hierarchical optimization, convex minimization, slow control

AMS subject classifications. 37N40, 49M30, 65K10

DOI. 10.1137/S105262340343467X

1. Introduction. Let $\Phi_1 : \mathbb{R}^n \rightarrow \mathbb{R}$ be a smooth convex function that we wish to minimize over the convex set S_0 . A powerful method consists in following the orbits of a discrete or continuous dynamical system, hopefully converging toward some element of $\operatorname{argmin}_{S_0} \Phi_1$. It is classical to apply the steepest descent system to the function $\Phi_1 + \delta_{S_0}$ (δ_{S_0} is the indicator function of S_0), thus leading to

$$\dot{x}(t) + \nabla \Phi_1(x(t)) \in -N_{S_0}(x(t)), \quad t \geq 0,$$

where $N_{S_0}(x(t))$ is the normal cone to S_0 at the point $x(t)$. It can be shown that this differential inclusion falls into the framework of gradient-projection methods (see Brézis [12]). Antipin [3] has initiated another continuous gradient-projection system, where the constraint S_0 is integrated through the projection operator P_{S_0} . In another direction, Cabot [13] has considered in a recent paper the following continuous dynamical system:

$$(\text{SDC}) \quad \dot{x}(t) + \nabla \Phi_0(x(t)) + \varepsilon(t) \nabla \Phi_1(x(t)) = 0, \quad t \geq 0,$$

where $\Phi_0 : \mathbb{R}^n \rightarrow \mathbb{R}$ is a smooth convex function satisfying $\operatorname{argmin} \Phi_0 = S_0$ and $\varepsilon : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is a control parameter tending to 0 when $t \rightarrow +\infty$. In the (SDC) system, the information on the constraint S_0 is contained in the function Φ_0 . The main difference from the previous methods lies in the fact that we do not have to handle nonsmooth operators, like the normal cone N_{S_0} or the projection P_{S_0} . However, the

*Received by the editors September 12, 2003; accepted for publication (in revised form) May 13, 2004; published electronically February 19, 2005.

<http://www.siam.org/journals/siopt/15-2/43467.html>

†Département de Mathématiques, Université de Limoges, 123 avenue Albert Thomas, 87060 Limoges Cedex, France (alexandre.cabot@unilim.fr).

difficulty comes from the choice of the control parameter ε . If the map ε tends to 0 too quickly, the potential Φ_1 cannot sufficiently influence the trajectory $x(\cdot)$, so that the minimization of Φ_1 may not occur. We feel the interest of a “slow control,” and it is shown in [13] that the adequate condition on ε is $\int_0^{+\infty} \varepsilon(t) dt = +\infty$. The notion of slow control has already been pointed out by several authors, essentially for continuous dynamical systems. See, for example, Cominetti [16] and Attouch and Cominetti [5], where the slow control is aimed at stabilizing a continuous gradient-like system toward a peculiar equilibrium. In the same direction, Attouch and Czarnecki [6], Cabot and Czarnecki [15], and Cabot [14] apply the notion of slow control to stabilize second-order-in-time systems.

Coming back to the (SDC) system and keeping in mind numerical applications, it is natural to deal with a discretized version of (SDC). In this paper, we will be especially interested in the following implicit discretization of (SDC):

$$-(x_{n+1} - x_n)/\lambda_n = \nabla(\Phi_0 + \varepsilon_n \Phi_1)(x_{n+1}), \quad n \in \mathbb{N},$$

where λ_n is the step length at iteration n and ε_n is the value of $\varepsilon(\cdot)$ at time $t_n := \sum_{k=0}^{n-1} \lambda_k$. If the closed convex functions Φ_0 and Φ_1 take their values in $\mathbb{R} \cup \{+\infty\}$ without regularity assumptions, one can easily adapt the previous algorithm as follows:

$$-(x_{n+1} - x_n)/\lambda_n \in \partial(\Phi_0 + \varepsilon_n \Phi_1)(x_{n+1}),$$

where ∂ denotes the subdifferential in the sense of convex analysis. Notice that this algorithm falls into the field of proximal point methods proposed in [24, 25] and inspired by [26]. Such methods have been intensively studied over the last few decades, and there is a significant amount of results concerning this type of algorithm, ranging from abstract convergence theorems to applications in nonlinear programming (see, for example, [1, 2, 7, 8, 10, 17, 19, 20, 21, 22, 23, 27, 28, 29, 31]). In the previous algorithm, the iterate x_{n+1} is uniquely determined by $x_{n+1} = J_{\lambda_n}^{A_n}(x_n)$, where A_n is the maximal monotone operator $\partial(\Phi_0 + \varepsilon_n \Phi_1)$ and $J_{\lambda}^A := (I + \lambda A)^{-1} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is the resolvent of A of parameter λ . In order to deal with numerical applications, it is convenient to authorize at each iteration n an error η_n in the evaluation of the subdifferential. More precisely, denoting by ∂_η the η -approximate subdifferential, we are led to the following algorithm:

$$(\mathcal{A}_1) \quad -(x_{n+1} - x_n)/\lambda_n \in \partial_{\eta_n}(\Phi_0 + \varepsilon_n \Phi_1)(x_{n+1}).$$

The sequence (η_n) of errors is assumed to be summable so as to remain close to the exact subdifferential. We show that, under the slow control criterion, i.e., $\sum_{n=0}^{+\infty} \varepsilon_n = +\infty$, each sequence generated by $(\mathcal{A}_1)$ tends to minimize Φ_1 over $\operatorname{argmin} \Phi_0$ in a sense that will be made precise throughout the paper.

The next stage consists in building an algorithm which is able to minimize hierarchically several functions over their successive argmin sets. More precisely, consider finite convex functions $\Phi_2, \dots, \Phi_N$ (with $N \geq 2$) and define their successive argmin sets by $S_0 := \operatorname{argmin} \Phi_0$ and $S_i := \operatorname{argmin}_{S_{i-1}} \Phi_i$ for $i \in \{1, \dots, N\}$. We introduce in the paper the algorithm $(\mathcal{A}_N)$, given by

$$(\mathcal{A}_N) \quad -(x_{n+1} - x_n)/\lambda_n \in \partial_{\eta_n} \left(\Phi_0 + \varepsilon_n \Phi_1 + \varepsilon_n^{(2)} \Phi_2 + \dots + \varepsilon_n^{(N)} \Phi_N \right) (x_{n+1}),$$

where the choice of the sequences $(\varepsilon_n^{(2)}), \dots, (\varepsilon_n^{(N)})$ depends on the functions $\Phi_0, \dots, \Phi_N$. We prove that, under adequate conditions, the algorithm $(\mathcal{A}_N)$ tends to minimize

hierarchically each function Φ_i on the set S_{i-1} , for $i \in \{1, \dots, N\}$. The introduction of the algorithms $(\mathcal{A}_i)_{i \in \{1, \dots, N\}}$ seems to be a new and promising tool in hierarchical minimization problems.

The paper is organized as follows. In section 2, we recall general features about proximal algorithms. We also study the case of a fast parametrization (ε_n) and we obtain the weak convergence of the algorithm $(\mathcal{A}_1)$ toward some element of $S_0 = \operatorname{argmin} \Phi_0$. Section 3 is devoted to the case of a slow control. In the finite-dimensional setting, we prove that the distance of the iterate x_n to the set $S_1 = \operatorname{argmin}_{S_0} \Phi_1$ tends toward 0. We give sufficient conditions ensuring that the sequence (x_n) converges toward some element of S_1 . Finally, in section 4, we generalize the previous results by considering the algorithm $(\mathcal{A}_N)$, which is shown to minimize hierarchically the respective functions $\Phi_1, \dots, \Phi_N$ over the respective sets $S_0, \dots, S_{N-1}$. For pedagogical reasons, the first paragraph of section 4 starts with the case $N = 2$.

2. General results. Case of a fastly vanishing term. In this section, we consider a Hilbert space H endowed with scalar product and corresponding norm, respectively, denoted by $\langle \cdot, \cdot \rangle$ and $|\cdot|$. Let $\Phi_0 : H \rightarrow \mathbb{R} \cup \{+\infty\}$ be a closed convex function and $\Phi_1 : H \rightarrow \mathbb{R}$ a finite convex function that are bounded from below. We are also given nonnegative sequences (ε_n) , (λ_n) , and (η_n) . Denoting by ∂_η the η -approximate subdifferential, we consider the following algorithm:

$$(\mathcal{A}_1) \quad -\frac{x_{n+1} - x_n}{\lambda_n} \in \partial_{\eta_n}(\Phi_0 + \varepsilon_n \Phi_1)(x_{n+1}).$$

Setting

$$(2.1) \quad \Psi_n := \Phi_0 - \inf \Phi_0 + \varepsilon_n(\Phi_1 - \inf \Phi_1) \geq 0,$$

algorithm $(\mathcal{A}_1)$ can be rewritten as the following diagonal proximal iteration:

$$(\text{DProx}) \quad -\frac{x_{n+1} - x_n}{\lambda_n} \in \partial_{\eta_n} \Psi_n(x_{n+1}),$$

where the sequence (Ψ_n) converges pointwise to Φ_0 as soon as $\lim_{n \rightarrow +\infty} \varepsilon_n = 0$. In the context of convex programming, this idea of coupling the proximal method with an approximate scheme has been used by Auslender, Crouzeix, and Fedit [8], Kaplan [19], Mouallif [27], and Mouallif and Tossings [28]. Later, a more systematic study of diagonal proximal methods was developed by Alart and Lemaire [1], Bahraoui and Lemaire [10], Cominetti [17], and Lemaire [20, 21, 22, 23]. In the literature, we may find other inexact diagonal processes, such as

$$(\text{DProx}^*) \quad x_{n+1} \in \eta_n\text{-argmin} \left(\frac{|\cdot - x_n|^2}{2\lambda_n} + \Psi_n \right),$$

where η -argmin denotes the η -approximate set of optimal minimizers. The approximate algorithm (DProx^*) has been introduced by Auslender [7] in the case $\Psi_n \equiv \Psi$ and the relations with (DProx) are explained in [7, Proof of Theorem 1.1].

Throughout this paper, we will assume the following hypotheses on the sequences (ε_n) , (λ_n) , (η_n) :

- $(\mathcal{H}_\varepsilon)$ The sequence (ε_n) is nonincreasing and $\lim_{n \rightarrow +\infty} \varepsilon_n = 0$.
- $(\mathcal{H}_\lambda)$ There exist $\underline{\lambda} > 0$ and $\bar{\lambda} > 0$ such that $\underline{\lambda} \leq \lambda_n \leq \bar{\lambda}$ for every $n \in \mathbb{N}$.
- $(\mathcal{H}_\eta)$ The sequence (η_n) is summable, i.e., $\sum_{n=0}^{+\infty} \eta_n < +\infty$.

The summability condition $(\mathcal{H}_\eta)$ means that the authorized error η_n in the computation of the subdifferential is small enough so as to remain close to the exact subdifferential. The following proposition states the main general features about the algorithm $(\mathcal{A}_1)$.

PROPOSITION 2.1. *Let H be a Hilbert space endowed with the norm $|\cdot|$, $\Phi_0 : H \rightarrow \mathbb{R} \cup \{+\infty\}$ a closed convex function, and $\Phi_1 : H \rightarrow \mathbb{R}$ a finite convex function. Assume that the functions Φ_0 and Φ_1 are bounded from below. We are given nonnegative sequences (ε_n) , (λ_n) , (η_n) satisfying, respectively, $(\mathcal{H}_\varepsilon)$, $(\mathcal{H}_\lambda)$, $(\mathcal{H}_\eta)$. Any sequence (x_n) generated by the algorithm $(\mathcal{A}_1)$ satisfies (i) $\lim_{n \rightarrow +\infty} \Phi_0(x_n) = \inf \Phi_0$. (ii) If the set $\operatorname{argmin} \Phi_0$ is nonempty and if $x_{n_k} \rightharpoonup \bar{x}$ weakly in H , then $\bar{x} \in \operatorname{argmin} \Phi_0$.*

This result is an immediate consequence of [23, Proposition 3.2] in view of the monotonicity of the sequence (Ψ_n) defined by (2.1). Proposition 2.1 is also closely related to [1, Theorem 3.1] and [21, Corollary 3.1].

Our purpose is now to specify the convergence properties of (x_n) . When $\varepsilon_n = 0$ for every $n \geq 0$, the $(\mathcal{A}_1)$ algorithm reduces to the standard proximal point method applied to Φ_0 . The sequence (x_n) generated by $(\mathcal{A}_1)$ is then known to weakly converge toward a minimum of Φ_0 : the arguments of the proof rely on the Opial lemma [30]. This result can be generalized when the sequence (ε_n) tends to zero fast enough. The key condition is $\sum_{n=0}^{+\infty} \varepsilon_n < +\infty$, and any sequence (ε_n) satisfying such a criterion will be referred to as a fast control (or sequence). Let us now state the following proposition.

PROPOSITION 2.2. *Under the hypotheses of Proposition 2.1, assume moreover that $\operatorname{argmin} \Phi_0 \neq \emptyset$ and that $\sum_{n=0}^{+\infty} \varepsilon_n < +\infty$. Then any sequence (x_n) defined by the algorithm $(\mathcal{A}_1)$ weakly converges toward some $x_\infty \in \operatorname{argmin} \Phi_0$.*

We will let the reader check that this result is implied by [23, Corollary 4.2]. Weak convergence of the sequences generated by (DProx) or (DProx*) has been established with varying degrees of generality in [1, 8, 17, 23, 27, 28]. A common characteristic of these results is that the approximate functions Ψ_n are expected to converge to their limit sufficiently fast.

It is also interesting to notice the analogy between the discrete case and the continuous one. In the continuous case, it has been proved in [13] that the trajectories of (SDC) weakly converge toward a minimum of Φ_0 under the condition $\int_0^{+\infty} \varepsilon(t) dt < +\infty$. This result is exactly the continuous version of Proposition 2.2. Conversely, it is shown in [13] that the assumption $\int_0^{+\infty} \varepsilon(t) dt = +\infty$ allows us to rescale the (SDC) system conveniently, then giving rise to the minimization of Φ_1 over the set $\operatorname{argmin} \Phi_0$. The same phenomenon occurs in the discrete case, as will be shown by Theorem 3.1.

3. Slow control: Minimization of Φ_1 over $\operatorname{argmin} \Phi_0$. From now on, H is a finite-dimensional Hilbert space, i.e., $H = \mathbb{R}^n$.

3.1. Convergence of the distance to the set $\operatorname{argmin}_{S_0} \Phi_1$. When (ε_n) is a fast control, Proposition 2.2 shows the convergence of the algorithm, but the limit does not depend explicitly on Φ_1 . The potential Φ_1 plays no crucial role because the sequence (ε_n) vanishes too quickly, from which comes the idea of introducing a slow control satisfying $\sum_{n=0}^{+\infty} \varepsilon_n = +\infty$. Assuming that $\Phi_1 = |\cdot|^2/2$, the algorithm $(\mathcal{A}_1)$ combines the proximal point method and the Tikhonov regularization [33]. Convergence under slow parametrization has been proved in this case by Moudafi [29], and the limit equals the element of minimal norm of $S_0 = \operatorname{argmin} \Phi_0$. This result presents interesting similarities with the corresponding results of Attouch and Cominetti [5] and Attouch and Czarnecki [6] in the continuous case. Coming back to the general

case, Cominetti [17] has established the convergence of $(\mathcal{A}_1)$ under slow parametrization when $\Phi_0 + \varepsilon \Phi_1$ is strongly convex for every $\varepsilon > 0$ (in fact, the results of [17] concern a more general approximation scheme). Our approach here is different and does not rely on the existence of an optimal path $(x(\varepsilon))_{\varepsilon > 0}$. The following theorem states that, under the condition $\sum_{n=0}^{+\infty} \varepsilon_n = +\infty$, the algorithm $(\mathcal{A}_1)$ tends to minimize Φ_1 over the set $\operatorname{argmin} \Phi_0$, without strong convexity assumptions on Φ_0 and Φ_1 .

THEOREM 3.1. *Let $\Phi_0 : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ be a closed convex function and $\Phi_1 : \mathbb{R}^n \rightarrow \mathbb{R}$ be a finite convex function. Assume that the functions Φ_0 and Φ_1 are bounded from below, that the set $S_0 := \operatorname{argmin} \Phi_0$ is nonempty, and that the set $S_1 := \operatorname{argmin}_{S_0} \Phi_1$ is nonempty and bounded. We consider a sequence (ε_n) satisfying $(\mathcal{H}_\varepsilon)$ and $\sum_{n=0}^{+\infty} \varepsilon_n = +\infty$. We are also given nonnegative sequences (λ_n) , (η_n) verifying, respectively, $(\mathcal{H}_\lambda)$ and $(\mathcal{H}_\eta)$. Then any sequence (x_n) generated by the algorithm $(\mathcal{A}_1)$ satisfies:*

- (i) $\lim_{n \rightarrow +\infty} d(x_n, S_1) = 0$,
- (ii) $\lim_{n \rightarrow +\infty} (\Phi_0(x_n), \Phi_1(x_n)) = (\min \Phi_0, \min_{S_0} \Phi_1)$, where $d(\cdot, S_1)$ stands for the distance function to the set S_1 .

Proof. (i) The proof relies on the study of the sequence (h_n) defined by

$$h_n = \frac{1}{2} d(x_n, S_1)^2.$$

Denoting by P_{S_1} the projection operator onto the convex set S_1 , we have

$$\begin{aligned} h_k &= \frac{1}{2} |x_k - P_{S_1}(x_k)|^2 \\ &= \frac{1}{2} |x_k - P_{S_1}(x_k) - (x_{k+1} - P_{S_1}(x_{k+1}))|^2 + \frac{1}{2} |x_{k+1} - P_{S_1}(x_{k+1})|^2 \\ &\quad + \langle x_k - P_{S_1}(x_k) - (x_{k+1} - P_{S_1}(x_{k+1})), x_{k+1} - P_{S_1}(x_{k+1}) \rangle \\ &\geq h_{k+1} + \langle x_k - x_{k+1}, x_{k+1} - P_{S_1}(x_{k+1}) \rangle \\ &\quad + \langle P_{S_1}(x_{k+1}) - P_{S_1}(x_k), x_{k+1} - P_{S_1}(x_{k+1}) \rangle. \end{aligned}$$

Since $P_{S_1}(x_k) \in S_1$, we classically have $\langle P_{S_1}(x_{k+1}) - P_{S_1}(x_k), x_{k+1} - P_{S_1}(x_{k+1}) \rangle \geq 0$ and finally

$$(3.1) \quad h_{k+1} - h_k \leq \langle x_{k+1} - x_k, x_{k+1} - P_{S_1}(x_{k+1}) \rangle.$$

From the fact that $-(x_{k+1} - x_k)/\lambda_k \in \partial_{\eta_k} (\Phi_0 + \varepsilon_k \Phi_1)(x_{k+1})$, we infer

$$\begin{aligned} &\langle -(x_{k+1} - x_k)/\lambda_k, x_{k+1} - P_{S_1}(x_{k+1}) \rangle \\ (3.2) \quad &\geq \Phi_0(x_{k+1}) - \min \Phi_0 + \varepsilon_k (\Phi_1(x_{k+1}) - \min_{S_0} \Phi_1) - \eta_k. \end{aligned}$$

By combining (3.1) and (3.2), we are led to

$$h_{k+1} - h_k + \lambda_k \left(\Phi_0(x_{k+1}) - \min \Phi_0 + \varepsilon_k (\Phi_1(x_{k+1}) - \min_{S_0} \Phi_1) \right) \leq \bar{\lambda} \eta_k$$

and therefore

$$(3.3) \quad h_{k+1} - h_k + \lambda_k \varepsilon_k (\Phi_1(x_{k+1}) - \min_{S_0} \Phi_1) \leq \bar{\lambda} \eta_k.$$

We now distinguish the following cases:

- (a) There exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$, $\Phi_1(x_n) > \min_{S_0} \Phi_1$.
- (b) For all $n_0 \in \mathbb{N}$, there exists $n \geq n_0$, $\Phi_1(x_n) \leq \min_{S_0} \Phi_1$.

Case (a). We assume that there exists $n_0 \in \mathbb{N}$ such that, for every $n \geq n_0$, $\Phi_1(x_n) > \min_{S_0} \Phi_1$. From (3.3), we deduce that, for every $k \geq n_0$, $h_{k+1} - h_k \leq \bar{\lambda} \eta_k$. Therefore, we have $(h_{k+1} - h_k)^+ \leq \bar{\lambda} \eta_k$, which in view of the following lemma implies that

$$(3.4) \quad \lim_{n \rightarrow +\infty} h_n \text{ exists.}$$

LEMMA 3.2. *Let (h_n) be a sequence of nonnegative reals satisfying $\sum_{k=0}^{+\infty} (h_{k+1} - h_k)^+ < +\infty$, where $[t]^+ := \max\{t, 0\}$ denotes the positive part of t . Then $\lim_{n \rightarrow +\infty} h_n$ exists.*

The proof of Lemma 3.2 is elementary and left to the reader. Coming back to the proof of Theorem 3.1, we infer that the sequence $(d(x_n, S_1))_{n \in \mathbb{N}}$ is bounded, and since S_1 is bounded, we conclude that (x_n) is bounded too. The end of the proof consists in showing that this limit equals zero. Let us add inequalities (3.3) from $k = n_0$ to $n - 1$; we find

$$h_n - h_{n_0} + \lambda \sum_{k=n_0}^{n-1} \varepsilon_k (\Phi_1(x_{k+1}) - \min_{S_0} \Phi_1) \leq \bar{\lambda} \sum_{k=n_0}^{n-1} \eta_k.$$

Since $h_n \geq 0$ and $\sum_{k=n_0}^{+\infty} \eta_k < +\infty$, we deduce that

$$(3.5) \quad \sum_{k=n_0}^{+\infty} \varepsilon_k (\Phi_1(x_{k+1}) - \min_{S_0} \Phi_1) < +\infty.$$

From (3.5), it is immediate that $\liminf_{n \rightarrow +\infty} \Phi_1(x_n) = \min_{S_0} \Phi_1$ (indeed, the assumption $\liminf_{n \rightarrow +\infty} \Phi_1(x_n) > \min_{S_0} \Phi_1$ would lead to a contradiction with the fact that $\sum_{k=0}^{+\infty} \varepsilon_k = +\infty$). Consider a subsequence of (x_n) , still denoted (x_n) such that $\lim_{n \rightarrow +\infty} \Phi_1(x_n) = \min_{S_0} \Phi_1$. Since the sequence (x_n) is bounded, we can extract a converging subsequence (x_{n_k}) of (x_n) : there exists $\bar{x} \in \mathbb{R}^n$ such that $\lim_{k \rightarrow +\infty} x_{n_k} = \bar{x}$. In view of Proposition 2.1(ii), we have $\bar{x} \in \operatorname{argmin} \Phi_0 = S_0$. The continuity of the finite convex function Φ_1 implies that

$$\Phi_1(\bar{x}) = \lim_{k \rightarrow +\infty} \Phi_1(x_{n_k}) = \lim_{n \rightarrow +\infty} \Phi_1(x_n) = \min_{S_0} \Phi_1,$$

and hence $\bar{x} \in \operatorname{argmin}_{S_0} \Phi_1 = S_1$. On the other hand, we have

$$\lim_{k \rightarrow +\infty} h_{n_k} = \lim_{k \rightarrow +\infty} \frac{1}{2} d(x_{n_k}, S_1)^2 = \frac{1}{2} d(\bar{x}, S_1)^2 = 0,$$

which means that 0 is a limit point of (h_n) . Since the sequence (h_n) is convergent, we conclude that $\lim_{n \rightarrow +\infty} h_n = 0$. (Notice that our arguments are no more valid in infinite-dimensional spaces. The strong convergence of the sequence (x_{n_k}) toward $\bar{x}$ would be replaced by the weak convergence. Since the distance function $x \mapsto d(x, S_1)^2$ is lower semicontinuous for the weak topology, we obtain that $\liminf_{k \rightarrow +\infty} \frac{1}{2} d(x_{n_k}, S_1)^2 \geq \frac{1}{2} d(\bar{x}, S_1)^2 = 0$, but this inequality is insufficient to conclude.)

Case (b). We assume that for every $n_0 \in \mathbb{N}$, there exists $n \geq n_0$ such that $\Phi_1(x_n) \leq \min_{S_0} \Phi_1$. Let us consider the sequence (τ_n) defined by

$$\tau_n = \max\{k \in \mathbb{N}, \quad k \leq n \quad \text{and} \quad \Phi_1(x_k) \leq \min_{S_0} \Phi_1\}.$$

From the above assumption, it is immediate that (τ_n) is defined for n large enough and that $\lim_{n \rightarrow +\infty} \tau_n = +\infty$. Suppose now that $\tau_n \leq n - 1$. From the definition of τ_n and formula (3.3), we have

$$\forall k \in \{\tau_n, n - 1\}, \quad h_{k+1} - h_k \leq \bar{\lambda} \eta_k.$$

By adding these $(n - \tau_n)$ inequalities, we obtain

$$h_n - h_{\tau_n} \leq \bar{\lambda} \sum_{k=\tau_n}^{n-1} \eta_k$$

and therefore

$$(3.6) \quad h_n \leq h_{\tau_n} + \bar{\lambda} \sum_{k=\tau_n}^{+\infty} \eta_k.$$

Notice that, if $\tau_n = n$, then we have $h_{\tau_n} = h_n$ so that relation (3.6) is always true. If we are able to prove that $\lim_{n \rightarrow +\infty} h_{\tau_n} = 0$, then inequality (3.6) combined with the fact that $\lim_{n \rightarrow +\infty} \sum_{k=\tau_n}^{+\infty} \eta_k = 0$, will immediately imply that $\lim_{n \rightarrow +\infty} h_n = 0$.

Let us first prove that the sequence (h_{τ_n}) is bounded. From Proposition 2.1(i), there exists $M_0 \in \mathbb{R}$ such that, for every $n \in \mathbb{N}$, $\Phi_0(x_n) \leq M_0$, i.e.,

$$\{x_n, n \in \mathbb{N}\} \subset [\Phi_0 \leq M_0].$$

From the definition of τ_n , the point x_{τ_n} belongs to $[\Phi_1 \leq \min_{S_0} \Phi_1]$ so that

$$x_{\tau_n} \in [\Phi_0 \leq M_0] \cap [\Phi_1 \leq \min_{S_0} \Phi_1],$$

as soon as x_{τ_n} is defined. In view of the following lemma, the set $[\Phi_0 \leq M_0] \cap [\Phi_1 \leq \min_{S_0} \Phi_1]$ is bounded, which implies that the sequence (h_{τ_n}) is bounded.

LEMMA 3.3. *Assume that the functions Φ_0 and Φ_1 satisfy the assumptions of Theorem 3.1. Then, for every $(M_0, M_1) \in \mathbb{R}^2$, the set $[\Phi_0 \leq M_0] \cap [\Phi_1 \leq M_1]$ is bounded. The proof of Lemma 3.3 is postponed to the appendix.*

Let us now show that $\lim_{n \rightarrow +\infty} h_{\tau_n} = 0$. Consider a convergent subsequence of the bounded sequence (x_{τ_n}) , still denoted by (x_{τ_n}) : there exists $\bar{x} \in \mathbb{R}^n$ such that $\lim_{n \rightarrow +\infty} x_{\tau_n} = \bar{x}$. The set $[\Phi_1 \leq \min_{S_0} \Phi_1]$ is closed as a sublevel set of the continuous function Φ_1 . From the definition of τ_n , we have $x_{\tau_n} \in [\Phi_1 \leq \min_{S_0} \Phi_1]$ for every $n \in \mathbb{N}$, thus implying $\bar{x} \in [\Phi_1 \leq \min_{S_0} \Phi_1]$. On the other hand, from Proposition 2.1(ii), we have $\bar{x} \in S_0 = \operatorname{argmin} \Phi_0$. We immediately deduce that $\bar{x} \in S_1 = \operatorname{argmin}_{S_0} \Phi_1$, which means that every limit point of the sequence (x_{τ_n}) belongs to S_1 . It is then clear that 0 is the unique limit point of the bounded sequence $(d(x_{\tau_n}, S_1))$. We easily conclude that $\lim_{n \rightarrow +\infty} h_{\tau_n} = \lim_{n \rightarrow +\infty} \frac{1}{2} d(x_{\tau_n}, S_1)^2 = 0$.

(ii) Since $\lim_{n \rightarrow +\infty} \Phi_0(x_n) = \min_{S_0} \Phi_0$ from Proposition 2.1(i), we just have to prove that $\lim_{n \rightarrow +\infty} \Phi_1(x_n) = \min_{S_0} \Phi_1$. The sequence $(\Phi_1(x_n))$ is the image of the bounded sequence (x_n) by the continuous function Φ_1 , and hence $(\Phi_1(x_n))$ is bounded. Let $(\Phi_1(x_{n_k}))$ be a converging subsequence of $(\Phi_1(x_n))$. Since (x_{n_k}) is bounded, there is a subsequence of (x_{n_k}) , still denoted by (x_{n_k}) , which converges to $\bar{x} \in \mathbb{R}^n$. From (i), we have $\lim_{k \rightarrow +\infty} d(x_{n_k}, S_1) = 0$, and hence $\bar{x} \in S_1$. From the continuity of Φ_1 , we deduce that $\lim_{k \rightarrow +\infty} \Phi_1(x_{n_k}) = \min_{S_0} \Phi_1$. Since $\min_{S_0} \Phi_1$ is the limit of every converging subsequence of $(\Phi_1(x_n))$, we conclude that $\lim_{n \rightarrow +\infty} \Phi_1(x_n) = \min_{S_0} \Phi_1$. $\square$

Remark 3.1. The main idea of the proof of Theorem 3.1 consists in the study of the sequence $(d(x_n, S_1)^2)_n$ by the distinction of cases (a) and (b). In the context of control and stabilization of nonlinear oscillators, this type of proof has been initiated by Attouch and Czarnecki [6]. It has been used in [13] to derive convergence properties of the (SDC) system (see the introduction). In fact, Theorem 3.1 can also be recovered¹ by a repeated application of [2, Theorem 3]. This last result is itself a discrete version of a former result due to Baillon and Cominetti [11, Theorem 2.1].

Remark 3.2. We point out that, in Theorem 3.1, the involved spaces are finite dimensional. Removing this restriction would extend the applicability of the approach to algorithms coming from the discretization of PDE models. It is out of the scope of the paper but this certainly indicates a matter for future research.

3.2. Convergence of the sequence (x_n) . From Theorem 3.1, the distance of the sequence (x_n) (generated by $(\mathcal{A}_1)$) to the set S_1 tends to 0. In particular, if S_1 is reduced to a singleton, then the sequence (x_n) converges. However, in the slow case we do not have any general result of convergence for (x_n) like Proposition 2.2 in the fast case. The convergence of the sequence (x_n) can be obtained by strengthening the assumptions on (ε_n) . Before stating the result, let us introduce the sequence (ω_n) defined by

$$(3.7) \quad \omega_n := \inf_{x \in \mathbb{R}^n} [\Phi_0(x) - \min \Phi_0 + \varepsilon_n (\Phi_1(x) - \min_{S_0} \Phi_1)].$$

The sequence (ω_n) is minorized by

$$\omega_n \geq -\varepsilon_n (\min_{S_0} \Phi_1 - \inf \Phi_1).$$

The following proposition establishes that the negative part of the sequence (ω_n) is negligible with respect to ε_n when $n \rightarrow +\infty$.

PROPOSITION 3.4. *Assume that $\Phi_0 : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ and $\Phi_1 : \mathbb{R}^n \rightarrow \mathbb{R}$ satisfy the hypotheses of Theorem 3.1. Consider a sequence (ε_n) such that $\lim_{n \rightarrow +\infty} \varepsilon_n = 0$. Then the sequence (ω_n) defined by (3.7) satisfies $\lim_{n \rightarrow +\infty} \omega_n^- / \varepsilon_n = 0$.*

Proof. Let us argue by contradiction and assume that $\liminf_{n \rightarrow +\infty} \omega_n / \varepsilon_n < 0$. Then there exist $\eta > 0$ and a sequence (n_k) tending toward $+\infty$ such that

$$\forall k \in \mathbb{N}, \quad \inf_{x \in \mathbb{R}^n} [\Phi_0(x) - \min \Phi_0 + \varepsilon_{n_k} (\Phi_1(x) - \min_{S_0} \Phi_1)] \leq -\eta \varepsilon_{n_k}.$$

Therefore, there exists a sequence (x_{n_k}) in $\mathbb{R}^n$ such that

$$(3.8) \quad \forall k \in \mathbb{N}, \quad \Phi_0(x_{n_k}) - \min \Phi_0 + \varepsilon_{n_k} (\Phi_1(x_{n_k}) - \min_{S_0} \Phi_1) \leq -\frac{\eta}{2} \varepsilon_{n_k}.$$

Noticing that $\Phi_1(x_{n_k}) \geq \inf \Phi_1$ and taking the upper limit when $k \rightarrow +\infty$, we find

$$(3.9) \quad \limsup_{k \rightarrow +\infty} \Phi_0(x_{n_k}) \leq \min \Phi_0.$$

Since, on the other hand, $\Phi_0(x_{n_k}) \geq \min \Phi_0$, we infer from (3.8) that

$$(3.10) \quad \limsup_{k \rightarrow +\infty} \Phi_1(x_{n_k}) \leq \min_{S_0} \Phi_1 - \frac{\eta}{2}.$$

¹The author is indebted to R. Cominetti and to one of the referees for this pertinent remark.

From (3.9) and (3.10), it is clear that for every $M_0 > \min \Phi_0$, we have

$$x_{n_k} \in [\Phi_0 \leq M_0] \cap [\Phi_1 \leq \min_{S_0} \Phi_1]$$

for k large enough. Since the set S_1 is bounded, we conclude in view of Lemma 3.3 that the sequence (x_{n_k}) is bounded. Therefore, there exist $\bar{x} \in \mathbb{R}^n$ and a subsequence of (x_{n_k}) , still denoted by (x_{n_k}) , such that $\lim_{k \rightarrow +\infty} x_{n_k} = \bar{x}$. From the closedness of Φ_0 and inequality (3.9), we deduce that

$$\Phi_0(\bar{x}) \leq \liminf_{k \rightarrow +\infty} \Phi_0(x_{n_k}) \leq \limsup_{k \rightarrow +\infty} \Phi_0(x_{n_k}) \leq \min \Phi_0,$$

i.e., $\bar{x} \in S_0$. Similarly, the continuity of Φ_1 combined with inequality (3.10) yields

$$\Phi_1(\bar{x}) = \lim_{k \rightarrow +\infty} \Phi_1(x_{n_k}) \leq \min_{S_0} \Phi_1 - \frac{\eta}{2},$$

a contradiction with $\bar{x} \in S_0$. $\square$

Let us now give examples for which it is possible to compute explicitly the sequence (ω_n) (or at least a lower bound for (ω_n)).

PROPOSITION 3.5. *Assume that $\Phi_0 : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ and $\Phi_1 : \mathbb{R}^n \rightarrow \mathbb{R}$ satisfy the hypotheses of Theorem 3.1.*

- (i) *If $\operatorname{argmin} \Phi_0 \cap \operatorname{argmin} \Phi_1 \neq \emptyset$, then $\omega_n \geq 0$ for every $n \in \mathbb{N}$.*
- (ii) *Assume that there exist $a > 0$, $b > 0$, and $p \geq 1$ such that*
 - (a) $\Phi_0 - \min \Phi_0 \geq a d(\cdot, S_0)^p$,
 - (b) $\Phi_1 - \min_{S_0} \Phi_1 \geq -b d(\cdot, [\Phi_1 \geq \min_{S_0} \Phi_1])$.

Then there exist $\alpha \geq 0$ and $q > 1$ such that $\omega_n \geq -\alpha \varepsilon_n^q$ for n large enough (when $p > 1$ the exponent q is the conjugate of p , i.e., $q = 1/(1 - 1/p)$).

Proof. (i) Notice that the assumption $\operatorname{argmin} \Phi_0 \cap \operatorname{argmin} \Phi_1 \neq \emptyset$ implies that $\operatorname{argmin}_{S_0} \Phi_1 = \operatorname{argmin} \Phi_0 \cap \operatorname{argmin} \Phi_1$ and $\min_{S_0} \Phi_1 = \min \Phi_1$. As a consequence,

$$\Phi_0 - \min \Phi_0 + \varepsilon_n (\Phi_1 - \min_{S_0} \Phi_1) = \Phi_0 - \min \Phi_0 + \varepsilon_n (\Phi_1 - \min \Phi_1) \geq 0$$

and $\omega_n \geq 0$ for every $n \in \mathbb{N}$.

(ii) Since the set S_0 is included in the set $[\Phi_1 \geq \min_{S_0} \Phi_1]$, we have $d(\cdot, S_0) \geq d(\cdot, [\Phi_1 \geq \min_{S_0} \Phi_1])$, which combined with the assumption on Φ_1 , implies

$$\Phi_1 - \min_{S_0} \Phi_1 \geq -b d(\cdot, S_0).$$

Taking into account assumption (a), we deduce from the previous inequality that, for every $x \in \mathbb{R}^n$,

$$\Phi_0(x) - \min \Phi_0 + \varepsilon_n (\Phi_1(x) - \min_{S_0} \Phi_1) \geq a d(x, S_0)^p - b \varepsilon_n d(x, S_0).$$

First assume that $p = 1$. It is then immediate that we have, for n large enough,

$$\Phi_0(x) - \min \Phi_0 + \varepsilon_n (\Phi_1(x) - \min_{S_0} \Phi_1) \geq 0,$$

so that $\omega_n \geq 0$. Now assume that $p > 1$. An elementary computation then shows that

$$a d(x, S_0)^p - b \varepsilon_n d(x, S_0) \geq -b \left(\frac{b}{pa} \right)^{\frac{1}{p-1}} \left(\frac{p-1}{p} \right) \varepsilon_n^{\frac{p}{p-1}},$$

so that the expected inequality holds with $\alpha = b \left(\frac{b}{pa} \right)^{\frac{1}{p-1}} \left(\frac{p-1}{p} \right)$ and $q = \frac{p}{p-1}$. $\square$

Remark 3.3. Given a convex set S_0 and a convex function $\Phi_1 : \mathbb{R}^n \rightarrow \mathbb{R}$, let us consider the following constrained minimization problem:

$$(\mathcal{P}) \quad \min\{\Phi_1(x), \quad x \in S_0\}.$$

As recalled in the introduction, we associate with $(\mathcal{P})$ the algorithm $(\mathcal{A}_1)$ via the choice of a function $\Phi_0 : \mathbb{R}^n \rightarrow \mathbb{R}$ satisfying $\operatorname{argmin} \Phi_0 = S_0$. For theoretical purposes, it is convenient to choose $\Phi_0 := d(\cdot, S_0)$ or $\Phi_0 := d(\cdot, S_0)^2$. For these functions, condition (a) of Proposition 3.5 trivially holds.

Remark 3.4. The Lipschitz-type condition (b) on Φ_1 is verified in many situations. Indeed, let us denote by M the quantity

$$M := \sup \{|\xi| : \xi \in \partial\Phi_1(x), \Phi_1(x) = \min_{S_0} \Phi_1\}.$$

Supposing that $M < +\infty$, then for any $x \in [\Phi_1 \leq \min_{S_0} \Phi_1]$, one has

$$\Phi_1(x) - \min_{S_0} \Phi_1 \geq \langle \partial\Phi_1(y), x - y \rangle \geq -M d(x, [\Phi_1 \geq \min_{S_0} \Phi_1]),$$

where y is the projection of x on $[\Phi_1 = \min_{S_0} \Phi_1]$.

Let us come back to the convergence of the sequence (x_n) associated with the algorithm $(\mathcal{A}_1)$. Proposition 3.4 shows the existence of a gap between the sequences (ε_n) and (ω_n^-) . The next proposition shows how to exploit this property to obtain the convergence of the sequence (x_n) generated by $(\mathcal{A}_1)$.

PROPOSITION 3.6. *Under the hypotheses of Theorem 3.1, let us consider the sequence (ω_n) defined by (3.7). We assume that $\sum_{n=0}^{+\infty} \varepsilon_n = +\infty$ and $\sum_{n=0}^{+\infty} \omega_n^- < +\infty$. Then, any sequence (x_n) generated by the algorithm $(\mathcal{A}_1)$ converges to some $\bar{x} \in S_1$.*

Proof. Given any $z \in S_1$, let us define the sequence (g_n) by

$$g_n = \frac{1}{2} |x_n - z|^2.$$

Decomposing $x_k - z$ as $x_k - z = x_k - x_{k+1} + x_{k+1} - z$, we obtain for every $k \in \mathbb{N}$

$$g_k = \frac{1}{2} |x_k - x_{k+1}|^2 + g_{k+1} + \langle x_k - x_{k+1}, x_{k+1} - z \rangle,$$

and hence

$$\begin{aligned} g_{k+1} &= g_k - \frac{1}{2} |x_{k+1} - x_k|^2 + \langle x_{k+1} - x_k, x_{k+1} - z \rangle \\ &\leq g_k - \lambda_k \langle -(x_{k+1} - x_k)/\lambda_k, x_{k+1} - z \rangle. \end{aligned}$$

Since $-(x_{k+1} - x_k)/\lambda_k \in \partial_{\eta_k}(\Phi_0 + \varepsilon_k \Phi_1)(x_{k+1})$, the previous inequality gives

$$g_{k+1} \leq g_k - \lambda_k \left(\Phi_0(x_{k+1}) + \varepsilon_k \Phi_1(x_{k+1}) - \Phi_0(z) - \varepsilon_k \Phi_1(z) - \eta_k \right).$$

As a consequence, we have

$$g_{k+1} - g_k + \lambda_k \left(\Phi_0(x_{k+1}) - \min \Phi_0 + \varepsilon_k (\Phi_1(x_{k+1}) - \min_{S_0} \Phi_1) \right) \leq \bar{\lambda} \eta_k,$$

and hence

$$g_{k+1} - g_k - \lambda_k \omega_k^- \leq g_{k+1} - g_k + \lambda_k \omega_k \leq \bar{\lambda} \eta_k.$$

This implies that

$$(g_{k+1} - g_k)^+ \leq \bar{\lambda} (\eta_k + \omega_k^-).$$

By using the fact that $\sum_{k=0}^{+\infty} \eta_k < +\infty$ and $\sum_{k=0}^{+\infty} \omega_k^- < +\infty$, we obtain in view of Lemma 3.2 that $\lim_{n \rightarrow +\infty} g_n$ exists, and hence

$$(3.11) \quad \lim_{n \rightarrow +\infty} |x_n - z| \quad \text{exists for any } z \in S_1.$$

Since the sequence (x_n) is bounded, we can extract a converging subsequence: there exist $\bar{x} \in \mathbb{R}^n$ and (x_{n_k}) such that $\lim_{k \rightarrow +\infty} x_{n_k} = \bar{x}$. From Theorem 3.1, we have $\bar{x} \in S_1$. Taking $z = \bar{x}$ in (3.11), we deduce that $\lim_{n \rightarrow +\infty} |x_n - \bar{x}|$ exists, which combined with $\lim_{k \rightarrow +\infty} |x_{n_k} - \bar{x}| = 0$, finally yields $\lim_{n \rightarrow +\infty} |x_n - \bar{x}| = 0$. $\square$

Let us denote by l^1 the set of nonnegative sequences (a_n) satisfying $\sum_{n=0}^{+\infty} a_n < +\infty$. The assumptions of Proposition 3.6 can be rewritten as $(\varepsilon_n) \notin l^1$ and $(\omega_n^-) \in l^1$. The question of the convergence is open when $(\omega_n^-) \notin l^1$, which corresponds to the “very slow case.” If $\omega_n \geq 0$ for every $n \in \mathbb{N}$, the required condition reduces to $(\varepsilon_n) \notin l^1$.

4. Generalization: Toward hierarchical minimization.

4.1. Algorithm $(\mathcal{A}_2)$: Minimization of Φ_2 over $\operatorname{argmin}_{S_0} \Phi_1$. Given three convex functions Φ_0 , Φ_1 , and Φ_2 , it is natural in view of Theorem 3.1 to try to minimize Φ_2 over the set $\operatorname{argmin}_{S_0} \Phi_1$. For that purpose, we define the algorithm $(\mathcal{A}_2)$:

$$(\mathcal{A}_2) \quad -\frac{x_{n+1} - x_n}{\lambda_n} \in \partial_{\eta_n} \left(\Phi_0 + \varepsilon_n \Phi_1 + \varepsilon_n^{(2)} \Phi_2 \right) (x_{n+1}),$$

where the choice of the sequence $(\varepsilon_n^{(2)})$ is specified in the next theorem.

THEOREM 4.1. *Let $\Phi_0 : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ be a closed convex function and $\Phi_1, \Phi_2 : \mathbb{R}^n \rightarrow \mathbb{R}$ finite convex functions. Assume that the functions Φ_0, Φ_1, Φ_2 are bounded from below, that the set $S_0 := \operatorname{argmin}_{S_0} \Phi_0$ is nonempty, and that the sets $S_1 := \operatorname{argmin}_{S_0} \Phi_1$, $S_2 := \operatorname{argmin}_{S_1} \Phi_2$ are nonempty and bounded. We are given a sequence (ε_n) satisfying $(\mathcal{H}_\varepsilon)$ and a sequence $(\varepsilon_n^{(2)})$ such that*

$$(4.1) \quad \lim_{n \rightarrow +\infty} \varepsilon_n^{(2)} / \varepsilon_n = 0 \quad \text{and} \quad \lim_{n \rightarrow +\infty} \omega_n^- / \varepsilon_n^{(2)} = 0,$$

where the sequence (ω_n) is defined by (3.7). We assume moreover that $\sum_{n=0}^{+\infty} \varepsilon_n^{(2)} = +\infty$. We are also given nonnegative sequences $(\lambda_n), (\eta_n)$ verifying, respectively, $(\mathcal{H}_\lambda)$ and $(\mathcal{H}_\eta)$. Any sequence (x_n) generated by the algorithm $(\mathcal{A}_2)$ satisfies

- (i) $\lim_{n \rightarrow +\infty} d(x_n, S_2) = 0$,
- (ii) $\lim_{n \rightarrow +\infty} (\Phi_0(x_n), \Phi_1(x_n), \Phi_2(x_n)) = (\min \Phi_0, \min_{S_0} \Phi_1, \min_{S_1} \Phi_2)$.

Proof. The reader is referred to the proof of Theorem 4.4, which is a generalization of Theorem 4.1 to the case $N \geq 2$. $\square$

As for the algorithm $(\mathcal{A}_1)$, we now study the convergence of the sequence (x_n) generated by $(\mathcal{A}_2)$. Let us introduce the sequence $(\omega_n^{(2)})$ defined by

$$(4.2) \quad \omega_n^{(2)} := \inf_{x \in \mathbb{R}^n} [\Phi_0(x) - \min \Phi_0 + \varepsilon_n (\Phi_1(x) - \min_{S_0} \Phi_1) + \varepsilon_n^{(2)} (\Phi_2(x) - \min_{S_1} \Phi_2)].$$

The sequence $(\omega_n^{(2)})$ is minorized by

$$\omega_n^{(2)} \geq -\varepsilon_n (\min_{S_0} \Phi_1 - \inf \Phi_1) - \varepsilon_n^{(2)} (\min_{S_1} \Phi_2 - \inf \Phi_2).$$

Let us notice that if $\operatorname{argmin} \Phi_0 \cap \operatorname{argmin} \Phi_1 \cap \operatorname{argmin} \Phi_2 \neq \emptyset$, then we have $\omega_n \geq 0$ for every $n \in \mathbb{N}$. Moreover, for every choice of $(\varepsilon_n^{(2)})$, we also have $\omega_n^{(2)} \geq 0$ for every $n \in \mathbb{N}$. The following proposition provides us with another example for which the sequence $(\omega_n^{(2)})$ can be explicitly evaluated.

PROPOSITION 4.2. *Assume that $\Phi_0 : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$, $\Phi_1, \Phi_2 : \mathbb{R}^n \rightarrow \mathbb{R}$ satisfy the hypotheses of Theorem 4.1. Assume that there exist $a_0, a_1, b_1, b_2, c \in \mathbb{R}_+^*$ and $p_0, p_1 \in [1, +\infty[$ such that*

$$\begin{cases} \text{(i)} & \Phi_0 - \min \Phi_0 \geq a_0 d(\cdot, S_0)^{p_0}, \\ \text{(ii)} & \Phi_1 - \min_{S_0} \Phi_1 \geq a_1 d(\cdot, [\Phi_1 \leq \min_{S_0} \Phi_1])^{p_1} - b_1 d(\cdot, [\Phi_1 \geq \min_{S_0} \Phi_1]), \\ \text{(iii)} & \Phi_2 - \min_{S_1} \Phi_2 \geq -b_2 d(\cdot, [\Phi_2 \geq \min_{S_1} \Phi_2]), \\ \text{(iv)} & d(\cdot, S_1) \leq c (d(\cdot, [\Phi_1 \leq \min_{S_0} \Phi_1]) + d(\cdot, S_0)). \end{cases}$$

Then the following assertions hold:

- (a) *There exist $m \geq 0$ and $q_0 > 1$ such that $\omega_n \geq -m \varepsilon_n^{q_0}$ for n large enough.*
- (b) *Take $\varepsilon_n^{(2)} = \varepsilon_n^r$ for some $r \in]1, q_0[$. Then there exist $M \geq 0$ and $q_1 > r$ such that $\omega_n^{(2)} \geq -M \varepsilon_n^{q_1}$ for n large enough.*

Proof. (a) Let us first minorize the quantity

$$A_{\varepsilon_n}(x) := \Phi_0(x) - \min \Phi_0 + \varepsilon_n (\Phi_1(x) - \min_{S_0} \Phi_1).$$

Since $S_0 \subset [\Phi_1 \geq \min_{S_0} \Phi_1]$, we have $d(\cdot, S_0) \geq d(\cdot, [\Phi_1 \geq \min_{S_0} \Phi_1])$, which, combined with (ii), implies

$$(4.3) \quad \Phi_1(x) - \min_{S_0} \Phi_1 \geq a_1 d(x, [\Phi_1 \leq \min_{S_0} \Phi_1])^{p_1} - b_1 d(x, S_0).$$

In view of (i), we deduce that $A_{\varepsilon_n}(x) \geq a_0 d(x, S_0)^{p_0} - b_1 \varepsilon_n d(x, S_0)$. The same computation as in the proof of Proposition 3.5 (ii) shows that there exist $m \geq 0$ and $q_0 > 1$ such that, for n large enough,

$$a_0 d(x, S_0)^{p_0} - b_1 \varepsilon_n d(x, S_0) \geq -m \varepsilon_n^{q_0}.$$

More precisely, if $p_0 > 1$ one can take $q_0 = p_0/(p_0 - 1)$ and if $p_0 = 1$, the previous inequality is satisfied with $m = 0$. We have established that $A_{\varepsilon_n}(x) \geq -m \varepsilon_n^{q_0}$, which ends the proof of (a).

(b) Let us now find a lower bound for the quantity

$$B_{\varepsilon_n}(x) := \Phi_0(x) - \min \Phi_0 + \varepsilon_n (\Phi_1(x) - \min_{S_0} \Phi_1) + \varepsilon_n^r (\Phi_2(x) - \min_{S_1} \Phi_2).$$

The inclusion $S_1 \subset [\Phi_2 \geq \min_{S_1} \Phi_2]$ implies $d(\cdot, S_1) \geq d(\cdot, [\Phi_2 \geq \min_{S_1} \Phi_2])$, and hence in view of (iii) and (iv),

$$(4.4) \quad \Phi_2(x) - \min_{S_1} \Phi_2 \geq -b_2 c (d(x, [\Phi_1 \leq \min_{S_0} \Phi_1]) + d(x, S_0)).$$

Combining (i), (4.3), and (4.4), we obtain

$$(4.5) \quad B_{\varepsilon_n}(x) \geq a_0 d(x, S_0)^{p_0} - b_1 \varepsilon_n d(x, S_0) - b_2 c \varepsilon_n^r d(x, S_0) \\ + a_1 \varepsilon_n d(x, [\Phi_1 \leq \min_{S_0} \Phi_1])^{p_1} - b_2 c \varepsilon_n^r d(x, [\Phi_1 \leq \min_{S_0} \Phi_1]).$$

Since the quantity ε_n^r is negligible with respect to ε_n when $n \rightarrow +\infty$, the same arguments as above show that there exists $m_2 \geq 0$ such that

$$(4.6) \quad a_0 d(x, S_0)^{p_0} - b_1 \varepsilon_n d(x, S_0) - b_2 c \varepsilon_n^r d(x, S_0) \geq -m_2 \varepsilon_n^{q_0}.$$

In the same way, we let the reader check that there exist $m_3 \geq 0$ and $q_1 > r$ such that

$$(4.7) \quad a_1 \varepsilon_n d(x, [\Phi_1 \leq \min_{S_0} \Phi_1])^{p_1} - b_2 c \varepsilon_n^r d(x, [\Phi_1 \leq \min_{S_0} \Phi_1]) \geq -m_3 \varepsilon_n^{q_1}.$$

(If $p_1 > 1$, then one can take $q_1 = (r-1)(\frac{p_1}{p_1-1}) + 1$ and if $p_1 = 1$, then inequality (4.7) is satisfied with $m_3 = 0$.) In view of (4.5), (4.6), and (4.7), we deduce that

$$B_{\varepsilon_n}(x) \geq -m_2 \varepsilon_n^{q_0} - m_3 \varepsilon_n^{q_1}.$$

Since $\min\{q_0, q_1\} > r$, the conclusion immediately follows. $\square$

Let us now comment on items (ii) and (iv) of Proposition 4.2 by means of the following remarks.

Remark 4.1. Assumption (ii) of Proposition 4.2 may be decomposed into two parts: the first one is

$$(4.8) \quad \forall x \in [\Phi_1 \geq \min_{S_0} \Phi_1], \quad \Phi_1(x) - \min_{S_0} \Phi_1 \geq a_1 d(x, [\Phi_1 \leq \min_{S_0} \Phi_1])^{p_1},$$

while the second one is

$$(4.9) \quad \forall x \in [\Phi_1 \leq \min_{S_0} \Phi_1], \quad \Phi_1(x) - \min_{S_0} \Phi_1 \geq -b_1 d(x, [\Phi_1 \geq \min_{S_0} \Phi_1]).$$

First assume that $\min_{S_0} \Phi_1 = \min \Phi_1$. In such a case, (4.8) reduces to

$$\forall x \in \mathbb{R}^n, \quad \Phi_1(x) - \min \Phi_1 \geq a_1 d(x, \operatorname{argmin} \Phi_1)^{p_1},$$

which is a classical assumption on Φ_1 . Conversely, assume that $\min_{S_0} \Phi_1 > \min \Phi_1$: we then have $0 \notin \partial \Phi_1(x)$ for every $x \in [\Phi_1 = \min_{S_0} \Phi_1]$. Let us denote by m the quantity

$$m := \inf \{|\xi| : \xi \in \partial \Phi_1(x), \Phi_1(x) = \min_{S_0} \Phi_1\}.$$

The theory of error bounds for closed convex functions (see, for example, [9]) shows that for every $x \in [\Phi_1 \geq \min_{S_0} \Phi_1]$,

$$\Phi_1(x) - \min_{S_0} \Phi_1 \geq m d(x, [\Phi_1 \leq \min_{S_0} \Phi_1]),$$

so that (4.8) holds with $p_1 := 1$ and $a_1 := m$ as soon as $m > 0$. For comments relative to the second part of (4.9), we refer the reader to Remark 3.4.

Remark 4.2. Let us recall that the set S_1 equals the intersection $S_0 \cap [\Phi_1 \leq \min_{S_0} \Phi_1]$. Assumption (iv) of Proposition 4.2 means that the intersecting sets S_0 and $[\Phi_1 \leq \min_{S_0} \Phi_1]$ are not tangent at the points of S_1 which belong to the boundary of $S_0 \cup [\Phi_1 \leq \min_{S_0} \Phi_1]$. In other words, the (possibly) inward corners of the set $S_0 \cup$

$[\Phi_1 \leq \min_{S_0} \Phi_1]$ must not be inward cusps. We let the reader check that assumption (iv) can be replaced by the following less stringent one: there exist $p_2, p_3 \in]0, 1]$ such that

$$(iv - bis) \quad d(., S_1) \leq c \left(d(., [\Phi_1 \leq \min_{S_0} \Phi_1])^{p_2} + d(., S_0)^{p_3} \right).$$

This last assumption allows a wider class of intersecting sets S_0 and $[\Phi_1 \leq \min_{S_0} \Phi_1]$.

Let us come back to the problem of the convergence of the algorithm $(\mathcal{A}_2)$. By using arguments which are similar to those of Proposition 3.4, it is easy to show that the quantity $[\omega_n^{(2)}]^-$ is negligible with respect to $\varepsilon_n^{(2)}$ when $n \rightarrow +\infty$. The following proposition shows how to exploit this property to deduce the convergence of the algorithm $(\mathcal{A}_2)$.

PROPOSITION 4.3. *Under the hypotheses of Theorem 4.1, let us consider the sequence $(\omega_n^{(2)})$ defined by (4.2). We assume that $\sum_{n=0}^{+\infty} \varepsilon_n^{(2)} = +\infty$ and $\sum_{n=0}^{+\infty} [\omega_n^{(2)}]^- < +\infty$. Then, any sequence (x_n) generated by the algorithm $(\mathcal{A}_2)$ converges to some $\bar{x} \in S_2$.*

Proof. The arguments are similar to the ones of Proposition 3.6. The reader is also referred to the proof of Proposition 4.5, which is a generalization of Proposition 4.3 to the case $N \geq 2$. $\square$

4.2. Algorithm $(\mathcal{A}_N)$ and hierarchical minimization. Let $\Phi_0 : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ be a closed convex function and $\Phi_1, \dots, \Phi_N : \mathbb{R}^n \rightarrow \mathbb{R}$ finite convex functions ($N \geq 1$). Assume that $\Phi_0, \Phi_1, \dots, \Phi_N$ are bounded from below and define the sets $(S_i)_{-1 \leq i \leq N}$ as follows: $S_{-1} := \mathbb{R}^n$ and for $i \in \{0, \dots, N\}$, $S_i := \operatorname{argmin}_{S_{i-1}} \Phi_i$. A challenging task consists in minimizing each function Φ_i on the set S_{i-1} for $i \in \{0, \dots, N\}$. This question of hierarchical minimization has been addressed by many authors (see, for example, Attouch [4], Cominetti and Courdurier [18]).

According to the previous results of the paper, the algorithm $(\mathcal{A}_1)$ (resp., $(\mathcal{A}_2)$) allows us to generate a sequence (x_n) which minimizes the function Φ_1 (resp., Φ_2) over the set S_0 (resp., S_1). It is then natural to define the algorithm $(\mathcal{A}_N)$ by

$$(\mathcal{A}_N) \quad -\frac{x_{n+1} - x_n}{\lambda_n} \in \partial_{\eta_n} \left(\Phi_0 + \varepsilon_n^{(1)} \Phi_1 + \varepsilon_n^{(2)} \Phi_2 + \dots + \varepsilon_n^{(N)} \Phi_N \right) (x_{n+1}),$$

where the sequences $(\varepsilon_n^{(1)}), \dots, (\varepsilon_n^{(N)})$ satisfy the construction suggested in the previous paragraph for $N = 2$. To unify the presentation, let us define the sequence $(\varepsilon_n^{(0)})$ by $\varepsilon_n^{(0)} = 1$ for every $n \in \mathbb{N}$. Assume that we have built the family of sequences $(\varepsilon_n^{(j)})_{1 \leq j \leq i}$ for some $i \in \{2, \dots, N\}$. For every $j \in \{1, \dots, i\}$, let us introduce the sequence $(\omega_n^{(j)})$ defined by

$$(4.10) \quad \omega_n^{(j)} := \inf_{x \in \mathbb{R}^n} \left[\sum_{l=0}^j \varepsilon_n^{(l)} (\Phi_l(x) - \min_{S_{l-1}} \Phi_l) \right].$$

Suppose that the family of sequences $(\varepsilon_n^{(j)})_{1 \leq j \leq i}$ satisfies the following property:

$$\forall j \in \{2, \dots, i\}, \quad \lim_{n \rightarrow +\infty} \varepsilon_n^{(j)} / \varepsilon_n^{(j-1)} = 0 \quad \text{and} \quad \lim_{n \rightarrow +\infty} [\omega_n^{(j-1)}]^- / \varepsilon_n^{(j)} = 0.$$

By imitating the proof of Proposition 3.4, we can easily show that the sequence $(\omega_n^{(i)})$ defined by (4.10) satisfies $\lim_{n \rightarrow +\infty} [\omega_n^{(i)}]^- / \varepsilon_n^{(i)} = 0$. The next step consists in choosing

a sequence $(\varepsilon_n^{(i+1)})$ satisfying

$$\lim_{n \rightarrow +\infty} \varepsilon_n^{(i+1)} / \varepsilon_n^{(i)} = 0 \quad \text{and} \quad \lim_{n \rightarrow +\infty} [\omega_n^{(i)}]^- / \varepsilon_n^{(i+1)} = 0.$$

By iterating the previous arguments, we obtain the existence of a family of sequences $(\varepsilon_n^{(i)})_{1 \leq i \leq N}$ fulfilling

$$\forall i \in \{2, \dots, N\}, \quad \lim_{n \rightarrow +\infty} \varepsilon_n^{(i)} / \varepsilon_n^{(i-1)} = 0 \quad \text{and} \quad \lim_{n \rightarrow +\infty} [\omega_n^{(i-1)}]^- / \varepsilon_n^{(i)} = 0.$$

Once the existence of such a family is established, the main result relative to the algorithm $(\mathcal{A}_N)$ is the following.

THEOREM 4.4. *Let $\Phi_0 : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ be a closed convex function and $\Phi_1, \dots, \Phi_N : \mathbb{R}^n \rightarrow \mathbb{R}$ finite convex functions. Assume that $\Phi_0, \Phi_1, \dots, \Phi_N$ are bounded from below and that the sets $(S_i)_{1 \leq i \leq N}$ defined as above are nonempty and bounded. We are given a sequence $(\varepsilon_n^{(1)})$ satisfying $(\mathcal{H}_\varepsilon)$ and a family of sequences $(\varepsilon_n^{(2)}), \dots, (\varepsilon_n^{(N)})$ satisfying*

$$\forall i \in \{2, \dots, N\}, \quad \lim_{n \rightarrow +\infty} \varepsilon_n^{(i)} / \varepsilon_n^{(i-1)} = 0 \quad \text{and} \quad \lim_{n \rightarrow +\infty} [\omega_n^{(i-1)}]^- / \varepsilon_n^{(i)} = 0,$$

where the sequences $(\omega_n^{(i)})_{1 \leq i \leq N}$ are defined by relation (4.10). The sequence $(\varepsilon_n^{(N)})$ is supposed to verify $\sum_{n=0}^{+\infty} \varepsilon_n^{(N)} = +\infty$. We are also given nonnegative sequences $(\lambda_n), (\eta_n)$ verifying, respectively, $(\mathcal{H}_\lambda)$ and $(\mathcal{H}_\eta)$. Any sequence (x_n) generated by the algorithm $(\mathcal{A}_N)$ fulfills:

- (i) $\lim_{n \rightarrow +\infty} d(x_n, S_N) = 0$,
- (ii) for all $i \in \{0, \dots, N\}$, $\lim_{n \rightarrow +\infty} \Phi_i(x_n) = \min_{S_{i-1}} \Phi_i$.

Proof. (i) The results of Proposition 2.1 for $(\mathcal{A}_1)$ can be immediately extended to $(\mathcal{A}_N)$. We then obtain that every limit point of the sequence (x_n) belongs to $\text{argmin} \Phi_0$. Let us argue by recurrence and let us denote by $(\mathcal{R}_i)$ the following recurrence assumption:

$$(\mathcal{R}_i) \quad \lim_{n \rightarrow +\infty} d(x_n, S_i) = 0.$$

We let the reader check that $(\mathcal{R}_1)$ is true. Let us prove the implication $(\mathcal{R}_{i-1}) \implies (\mathcal{R}_i)$ for every $i \in \{2, \dots, N\}$. For that purpose, we define the sequence $h_n^{(i)}$ by

$$h_n^{(i)} = \frac{1}{2} d(x_n, S_i)^2.$$

We let the reader check that

$$h_{k+1}^{(i)} - h_k^{(i)} + \lambda_k \sum_{j=0}^N \varepsilon_k^{(j)} (\Phi_j(x_{k+1}) - \Phi_j(P_{S_i}(x_{k+1}))) \leq \bar{\lambda} \eta_k.$$

Notice that, since $P_{S_i}(x_{k+1}) \in S_i$, we have

$$\forall j \in \{0, \dots, i\}, \quad \Phi_j(P_{S_i}(x_{k+1})) = \min_{S_{j-1}} \Phi_j,$$

so that the previous inequality can be rewritten as

$$(4.11) \quad \begin{aligned} h_{k+1}^{(i)} - h_k^{(i)} + \lambda_k \sum_{j=0}^i \varepsilon_k^{(j)} (\Phi_j(x_{k+1}) - \min_{S_{j-1}} \Phi_j) + \lambda_k \sum_{j=i+1}^N \varepsilon_k^{(j)} (\Phi_j(x_{k+1}) \\ - \Phi_j(P_{S_i}(x_{k+1}))) \leq \bar{\lambda} \eta_k. \end{aligned}$$

Each function Φ_j is minorized by $\inf \Phi_j$ on $\mathbb{R}^n$ and majorized by $\sup_{S_i} \Phi_j$ on the compact set S_i . Setting $m_j := \sup_{S_i} \Phi_j - \inf \Phi_j \geq 0$, we deduce that, for every $k \geq 0$,

$$\Phi_j(x_{k+1}) - \Phi_j(P_{S_i}(x_{k+1})) \geq -m_j.$$

Since $\lim_{n \rightarrow +\infty} \varepsilon_n^{(j)} / \varepsilon_n^{(j-1)} = 0$ when $n \rightarrow +\infty$ for every $j \in \{2, \dots, N\}$, there exists $M \in \mathbb{R}_+$ such that for k large enough,

$$(4.12) \quad \sum_{j=i+1}^N \varepsilon_k^{(j)} (\Phi_j(x_{k+1}) - \Phi_j(P_{S_i}(x_{k+1}))) \geq -M \varepsilon_k^{(i+1)}.$$

On the other hand, from the definition of the sequence $(\omega_n^{(i-1)})$, we have

$$(4.13) \quad \sum_{j=0}^{i-1} \varepsilon_k^{(j)} (\Phi_j(x_{k+1}) - \min_{S_{j-1}} \Phi_j) \geq \omega_k^{(i-1)} \geq -[\omega_k^{(i-1)}]^-.$$

By combining (4.11), (4.12), and (4.13), we deduce

$$h_{k+1}^{(i)} - h_k^{(i)} + \lambda_k \varepsilon_k^{(i)} \left(\Phi_i(x_{k+1}) - \min_{S_{i-1}} \Phi_i - M \varepsilon_k^{(i+1)} / \varepsilon_k^{(i)} - [\omega_k^{(i-1)}]^- / \varepsilon_k^{(i)} \right) \leq \bar{\lambda} \eta_k.$$

The rest of the proof consists in distinguishing the following cases:

(a_i) There exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$, $\Phi_i(x_{n+1}) > \min_{S_{i-1}} \Phi_i + M \varepsilon_n^{(i+1)} / \varepsilon_n^{(i)} + [\omega_n^{(i-1)}]^- / \varepsilon_n^{(i)}$.

(b_i) For all $n \in \mathbb{N}$, there exists $n \geq n_0$, $\Phi_i(x_{n+1}) \leq \min_{S_{i-1}} \Phi_i + M \varepsilon_n^{(i+1)} / \varepsilon_n^{(i)} + [\omega_n^{(i-1)}]^- / \varepsilon_n^{(i)}$.

The arguments of the proof of Theorem 3.1(i) still apply here insofar as the quantities $\varepsilon_n^{(i+1)} / \varepsilon_n^{(i)}$ and $[\omega_n^{(i-1)}]^- / \varepsilon_n^{(i)}$ tend to 0 when $n \rightarrow +\infty$. As a consequence, we obtain that $\lim_{n \rightarrow +\infty} h_n^{(i)} = \lim_{n \rightarrow +\infty} \frac{1}{2} d(x_n, S_i)^2 = 0$ and $(\mathcal{R}_i)$ is satisfied. The details are left to the reader.

(ii) is an immediate consequence of (i). For further details, we refer the reader to the proof of Theorem 3.1(ii). $\square$

Under the hypotheses of Theorem 4.4, it is easy to verify that the negative part of the sequence $(\omega_n^{(N)})$ is negligible with respect to $\varepsilon_n^{(N)}$ when $n \rightarrow +\infty$. The next proposition shows how to exploit this property to obtain the convergence of the sequence (x_n) generated by $(\mathcal{A}_N)$.

PROPOSITION 4.5. *Under the hypotheses of Theorem 4.4, assume that $\sum_{n=0}^{+\infty} \varepsilon_n^{(N)} = +\infty$ and $\sum_{n=0}^{+\infty} [\omega_n^{(N)}]^- < +\infty$. Then, any sequence (x_n) generated by the algorithm $(\mathcal{A}_N)$ converges to some $\bar{x} \in S_N$.*

Proof. Given any $z \in S_N$, let us define the sequence (g_n) by $g_n = \frac{1}{2} |x_n - z|^2$. An elementary computation shows that

$$g_{k+1} - g_k + \lambda_k \sum_{j=0}^N \varepsilon_k^{(j)} (\Phi_j(x_{k+1}) - \min_{S_{j-1}} \Phi_j) \leq \bar{\lambda} \eta_k.$$

From the definition of $(\omega_n^{(N)})$, we deduce

$$g_{k+1} - g_k - \lambda_k [\omega_k^{(N)}]^- \leq g_{k+1} - g_k + \lambda_k \omega_k^{(N)} \leq \bar{\lambda} \eta_k,$$

and hence

$$(g_{k+1} - g_k)^+ \leq \bar{\lambda}(\eta_k + [\omega_k^{(N)}]^-).$$

The sequel of the proof is similar to the one of Proposition 3.6, and the reader is referred to it. $\square$

5. Appendix: Proof of Lemma 3.3. Without loss of generality, we can assume that $M_0 \geq \min \Phi_0$ and $M_1 \geq \min_{S_0} \Phi_1$. We then have

$$S_1 \subset [\Phi_0 \leq M_0] \cap [\Phi_1 \leq M_1],$$

so that the set $[\Phi_0 \leq M_0] \cap [\Phi_1 \leq M_1]$ is nonempty. It is then well known (see, for example, [32]) that the closed convex sets $[\Phi_0 \leq M_0]$ and $[\Phi_1 \leq M_1]$ satisfy the following equality:

$$([\Phi_0 \leq M_0] \cap [\Phi_1 \leq M_1])^\infty = [\Phi_0 \leq M_0]^\infty \cap [\Phi_1 \leq M_1]^\infty,$$

where C^∞ is the horizon cone of C . Denoting by f^∞ the horizon function of f , we have $[\Phi_0 \leq M_0]^\infty = [\Phi_0^\infty \leq 0]$ and $[\Phi_1 \leq M_1]^\infty = [\Phi_1^\infty \leq 0]$. As a consequence, the following equivalences hold:

$$\begin{aligned} [\Phi_0 \leq M_0] \cap [\Phi_1 \leq M_1] \quad \text{bounded} &\iff ([\Phi_0 \leq M_0] \cap [\Phi_1 \leq M_1])^\infty = \{0\} \\ &\iff [\Phi_0^\infty \leq 0] \cap [\Phi_1^\infty \leq 0] = \{0\}. \end{aligned}$$

It is then clear that the boundedness of the set $[\Phi_0 \leq M_0] \cap [\Phi_1 \leq M_1]$ does not depend on the values M_0 and M_1 . Taking $M_0 = \min \Phi_0$ and $M_1 = \min_{S_0} \Phi_1$, the previous set reduces to S_1 which is bounded by assumption. We conclude that the set $[\Phi_0 \leq M_0] \cap [\Phi_1 \leq M_1]$ is bounded for any value of M_0 and M_1 .

Acknowledgments. The author thanks Professor Cominetti for pertinent remarks and fruitful discussions about the paper. The author also expresses his gratitude to the referees for their careful reading of the paper. Their valuable suggestions and critical comments made numerous improvements throughout.

REFERENCES

- [1] P. ALART AND B. LEMAIRE, *Penalization in nonclassical convex programming via variational convergence*, Math. Programming, 51 (1991), pp. 307–331.
- [2] F. ALVAREZ AND R. COMINETTI, *Primal and dual convergence of a proximal point exponential penalty method for linear programming*, Math. Programming, 93 (2002), pp. 87–96.
- [3] A. S. ANTIPIN, *Minimization of convex functions on convex sets by means of differential equations*, Differential Equations, 30 (1994), pp. 1365–1375.
- [4] H. ATTOUCH, *Viscosity solutions of minimization problems*, SIAM J. Optim., 6 (1996), pp. 769–806.
- [5] H. ATTOUCH AND R. COMINETTI, *A dynamical approach to convex minimization coupling approximation with the steepest descent method*, J. Differential Equations, 128 (1996), pp. 519–540.
- [6] H. ATTOUCH AND M.-O. CZARNECKI, *Asymptotic control and stabilization of nonlinear oscillators with non isolated equilibria*, J. Differential Equations, 179 (2002), pp. 278–310.
- [7] A. AUSLENDER, *Numerical methods for nondifferentiable convex optimization*, Math. Programming Stud., 30 (1987), pp. 102–126.
- [8] A. AUSLENDER, J.-P. CROUZEIX, AND P. FEDIT, *Penalty-proximal methods in convex programming*, J. Optim. Theory Appl., 55 (1987), pp. 1–21.
- [9] D. AZÉ AND J. N. CORVELLEC, *Characterizations of error bounds for lower semicontinuous functions on metric spaces*, ESAIM Control Optim. Calc. Var., 10 (2004), pp. 409–425.

- [10] M. A. BAHRAOUI AND B. LEMAIRE, *Convergence of diagonally stationary sequences in convex optimization*, Set-Valued Anal., 2 (1994), pp. 49–61.
- [11] J.-B. BAILLON AND R. COMINETTI, *A convergence result for non-autonomous subgradient evolution equations and its application to the steepest descent exponential penalty trajectory in linear programming*, J. Funct. Anal., 187 (2001), pp. 263–273.
- [12] H. BRÉZIS, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, North Holland Math. Stud. 5, North-Holland, Amsterdam, 1973.
- [13] A. CABOT, *The steepest descent dynamical system with control. Applications to constrained minimization*, ESAIM Control Optim. Calc. Var., 10 (2004), pp. 243–258.
- [14] A. CABOT, *Inertial gradient-like dynamical system controlled by a stabilizing term*, J. Optim. Theory Appl., 120 (2004), pp. 275–303.
- [15] A. CABOT AND M.-O. CZARNECKI, *Asymptotic control of pairs of oscillators coupled by a repulsion, with nonisolated equilibria I: The regular case*, SIAM J. Control Optim., 41 (2002), pp. 1254–1280.
- [16] R. COMINETTI, *Asymptotic convergence of the steepest descent method for the exponential penalty in linear programming*, J. Convex Anal., 2 (1995), pp. 145–152.
- [17] R. COMINETTI, *Coupling the proximal point algorithm with approximation methods*, J. Optim. Theory Appl., 95 (1997), pp. 581–600.
- [18] R. COMINETTI AND M. COURDURIER, *Coupling general penalty schemes for convex programming with the steepest descent and the proximal point algorithm*, SIAM J. Optim., 13 (2002), pp. 745–765.
- [19] A. KAPLAN, *A convex programming method with internal regularization*, Dokl. Akad. Nauk, 241 (1978), pp. 22–25.
- [20] B. LEMAIRE, *Coupling optimization methods and variational convergence*, in Trends in Mathematical Optimization, International Ser. Numer. Math. 84, Birkhäuser, Basel, 1988, pp. 163–179.
- [21] B. LEMAIRE, *About the convergence of the proximal method*, in Advances in Optimization, Lecture Notes in Econom. and Math. Systems 382, Springer, Berlin, 1992, pp. 39–51.
- [22] B. LEMAIRE, *Bounded diagonally stationary sequences in convex optimization*, J. Convex Anal., 1 (1994), pp. 75–86.
- [23] B. LEMAIRE, *On the convergence of some iterative methods for convex minimization*, in Recent Developments in Optimization, Lecture Notes in Econom. and Math. Systems 429, Springer, Berlin, 1995, pp. 252–268.
- [24] B. MARTINET, *Régularisation d'inéquations variationnelles par approximations successives*, Rev. Française Informat. Recherche Opérationnelle, 4 (1970), pp. 154–159.
- [25] B. MARTINET, *Détermination approchée d'un point fixe d'une application pseudo-contractante*, C. R. Acad. Sci. Paris Ser. A-B, 274 (1972), pp. 163–165.
- [26] J. J. MOREAU, *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.
- [27] K. MOUALLIF, *Sur la convergence d'une méthode associant pénalisation et régularisation*, Bull. Soc. Roy. Sci. Liège, 56 (1987), pp. 175–180.
- [28] K. MOUALLIF AND P. TOSSINGS, *Une méthode de pénalisation exponentielle associée à une régularisation proximale*, Bull. Soc. Roy. Sci. Liège, 56 (1987), pp. 181–192.
- [29] A. MOUDAFI, *Coupling proximal algorithm and Tikhonov method*, Nonlinear Times Digest, 1 (1994), pp. 203–209.
- [30] Z. OPIAL, *Weak convergence of the sequence of successive approximations for nonexpansive mappings*, Bull. Amer. Math. Soc., 73 (1967), pp. 591–597.
- [31] R. T. ROCKAFELLAR, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
- [32] R. T. ROCKAFELLAR AND R. WETS, *Variational Analysis*, Springer, Berlin, 1998.
- [33] A. N. TIKHONOV AND V. Y. ARSENEV, *Méthodes de résolution de problèmes mal posés*, MIR, Moscow, 1976.

A FAST AUTOMATIC ALGORITHM FOR IMAGE DENOISING BY A REGULARIZATION METHOD*

P. BARONE[†] AND M. CARAMIA[†]

Dedicated to the memory of our friend and mentor Stefano Pallottino

Abstract. A new fast iterative algorithm is proposed to denoise images affected by Gaussian additive noise, by solving a constrained optimization problem. A family of functionals parameterized by a few hyperparameters already proposed in the literature to solve denoising problems is considered. The algorithm estimates jointly the image and the hyperparameters, therefore providing an automatic method. Each member of the family is made up of a coherence with the data term and a term enforcing a roughness penalty and preserving jump discontinuities. Assuming we know the noise variance, an adequacy constraint is also considered. The algorithm computes a member of the family and a minimizer of it which satisfies the constraint. A convergence proof is provided. We then consider a heuristic version of the algorithm which gives restorations of comparable quality whose computational complexity is a linear function of the pixels number. Experimental results on synthetic and real data are presented. Moreover, numerical comparisons with several fast denoising methods are provided.

Key words. image processing, constrained optimization, adaptive regularization

AMS subject classifications. 47A52, 64K05, 90C25, 90C26

DOI. 10.1137/S1052623402406789

Introduction. In this paper a new fast algorithm to solve a typical problem of image analysis is proposed. An image can be considered as a discretization of a function defined on a compact set $D \subset \mathbb{R}^2$ taking values in $\mathbb{R}^+$. We assume to have a degraded image obtained by adding to the true image a realization of a Gaussian white noise with known standard deviation σ . The problem consists of retrieving the original image from the measured one. Despite this simple setup the problem can be quite difficult, depending on the shape of the true image as discussed, e.g., in [8]. For example, to retrieve the jump discontinuities, which are usually the most important features of the true image, nonquadratic regularization functionals should be used [4, 5, 10, 11]. In fact, a simple Tikhonov regularization approach consisting of minimizing the L_2 norm of the difference between the true and the measured images with a quadratic smoothness constraint is unable to cope with this situation. Moreover, it has been proved by Nikolova [21] that stable restoration of images that exactly satisfy locally a linear difference equation—such as, e.g., piecewise constant images—can be obtained only for a specific class of regularization term that is not included in the class considered in the following.

However, it is a well-established fact that a nonquadratic convex [2, 5, 13] or nonconvex [4, 5, 10] regularization term, even if not belonging to the Nikolova class, can provide satisfactorily good results as long as the correct order of the linear difference equation that controls the local smoothness of the true image is chosen [11]. These choices cause severe computational problems, especially in the nonconvex case because of the large dimension of the solution space.

*Received by the editors May 2, 2002; accepted for publication (in revised form) May 12, 2004; published electronically February 19, 2005.

<http://www.siam.org/journals/siopt/15-2/40678.html>

[†]Istituto per le Applicazioni del Calcolo “M. Picone”—C.N.R. Viale del Policlinico 137, 00161 Rome, Italy (barone@iac.rm.cnr.it, caramia@iac.rm.cnr.it).

A further crucial problem has to be addressed. The hyperparameter weighting the relative importance of the coherence to data term and the regularization term has to be chosen carefully in order to get reasonable results. This is a known difficult problem that is usually left to the reader. An extensive literature exists on the subject (e.g., see [9, 11, 14, 25, 29]), where several different principles are used to build up an estimation procedure based either on a model for the true image or directly on the measured data. The most convincing approach is probably the full Bayesian approach, where the image and the hyperparameters are estimated jointly. Unfortunately, almost all of these methods are computationally demanding, requiring, e.g., the solution of nonlinear systems of equations and/or numerical integration in large dimension spaces. Therefore, they are not suitable if computational speed is an important issue, as in real time applications. Suboptimal solutions have been proposed (e.g., see [7]), where several image estimates are computed for different values of the hyperparameter and the best one is chosen by performing an adequacy test, consisting of checking that the residuals distribution is close to the noise distribution.

In what follows, we go a step further in this direction with the aim to speed up the computations, by proposing an iterative algorithm that estimates the hyperparameter adaptively during the image estimation process. The idea is the following. Let us assume to know a threshold under which the discrepancy between the residuals standard deviation and the noise standard deviation is small enough to produce a realistic image estimate. Consider then the family of regularized functionals described above parameterized by the hyperparameter. We then look for a hyperparameter value such that a (local) minimizer of the corresponding functional produces residuals such that the discrepancy between their standard deviation and the noise standard deviation is less than the given threshold.

In what follows, we prove that under mild assumptions it is possible to devise a convergent iterative algorithm that solves this problem. Moreover, the proof holds in a more general setup than that described above. In fact the algorithm can be applied to n -dimensional arrays and more than one regularization term; therefore, more than one hyperparameter can be considered. For simplifying the proofs all results will be given for the simplest 2-dimensional model. However, some experiments are reported for a real 3-dimensional case. We also show experimentally that further speedup can be achieved by relaxing the condition of getting a (local) minimum of the current functional still verifying the adequacy condition. It turns out that this relaxed version of the method can be interpreted as a generalization of the nonlinear diffusion approach to image denoising [23, 31], where a coherence to the data term is taken into account in the iteration, adaptively weighted with respect to the nonlinear diffusion term.

Further pursuing this connection, following [3], the relaxed version can also be interpreted as a generalization of a robust estimation approach in which a coherence to the data term is included in the optimization criterion and the regularization parameter is adaptively estimated. Finally, we notice that the relaxed version of the method generalizes the method proposed in [28] by adaptively estimating the regularization parameter.

The paper is organized as follows. In section 1 the optimization problem is stated. In section 2 a prototypical algorithm for a convex regularization term is discussed. In section 3 the convergence of this algorithm is proved and the nonconvex case is discussed. In section 4 a heuristic variant of the previous algorithms that produces suboptimal solutions and requires a number of flops of the order of the pixels number is introduced. Moreover, the connection with nonlinear diffusion is discussed.

Finally, in section 5 some experimental results are shown both on synthetic and real 3-dimensional data representing time series of images. Moreover, numerical comparisons with several fast denoising methods are provided.

1. Problem statement and basic properties. In order to simplify notations, we assume that the measured image corresponds to a square matrix which can be vectorized, giving rise to a vector $y \in \mathbb{R}^n$. With the same convention, let us denote by $x \in \mathbb{R}^n$ the image to be retrieved. The measure equation is then

$$(1) \quad y = x + \varepsilon,$$

where ε is a vector of independent Gaussian variables with mean zero and variance σ^2 . Among the standard approaches to make inference about x , the regularization technique is based on functionals defined by

$$F(x; \Theta, \eta) = q(x) + \Theta \Psi(x, \eta), \quad 0 \leq \Theta < +\infty, \quad 0 < \eta < +\infty,$$

where

$$q(x) = (x - y)^T (x - y)$$

and

$$\Psi(x, \eta) = \sum_{\langle i, j \rangle} \phi \left(\frac{x_i - x_j}{\eta} \right)$$

is a regularization functional where $\langle i, j \rangle$ indicates sets of neighbor pixels. In what follows, we will consider only first order neighbors, i.e., if pixel i has coordinates (p, q) in the associated square matrix, its neighbors have coordinates $(p - 1, q)$, $(p + 1, q)$, $(p, q - 1)$, $(p, q + 1)$. The function ϕ can be chosen in several ways as discussed, e.g., in [5]. Specifically, we require that

1. $\phi(t) \geq 0 \quad \forall t$;
2. $\phi(t) = \phi(-t)$;
3. $\phi(t)$ twice differentiable;
4. $\phi'(t) \geq 0 \quad \forall t \geq 0$;
5. $\phi'(t)/t$ continuous and strictly decreasing on $[0, \infty)$;
6. $\lim_{t \rightarrow \infty} \frac{\phi'(t)}{t} = 0$;
7. $\lim_{t \rightarrow 0} \frac{\phi'(t)}{t} = 1$.

The hyperparameters Θ and η have to be estimated too. Usually this last problem is left to the user and attention is concentrated either on the problem of minimizing $F(x; \Theta, \eta)$ if a deterministic approach is adopted or on the problem of sampling from the Gibbs distribution $P(x; \Theta, \eta) \propto e^{-\frac{F(x; \Theta, \eta)}{\mathcal{T}}}$ if a stochastic approach is used, where $\mathcal{T}$ is kept fixed or not, depending on which estimator is chosen. The parameter η is easier to estimate than Θ because of its interpretation as a scaling parameter [11].

Therefore, in what follows, we assume to know σ and η and consider the following problem. Find a nonnegative $\hat{\Theta} < \infty$ and a $\hat{x} \in \mathbb{R}^n$ such that

$$(2) \quad \begin{cases} \hat{x}(\hat{\Theta}) = \operatorname{argmin}_x F(x, \hat{\Theta}), \\ T(\hat{x}(\hat{\Theta})) \leq \tau, \end{cases}$$

where T is given by

$$(3) \quad T(x) = \frac{\left| \sqrt{\frac{q(x)}{n}} - \sigma \right|}{\sigma}$$

which represents the relative error on the estimation of the noise standard deviation pertinent to the restoration x . In this case a reasonable threshold τ will be a value close to 0. In what follows, to simplify the notations, when no confusion arises, we will use $\hat{x}$ instead of $\hat{x}(\Theta)$, for any fixed Θ .

Remark 1. Given a value $\hat{\Theta}$, the denoising problem cannot be reduced to finding

$$(4) \quad \begin{cases} \hat{x} = \operatorname{argmin}_x F(x, \hat{\Theta}), \\ T(x) \leq \tau \end{cases}$$

since $\hat{x}$ in (4) is not necessarily a minimizer of $F(x, \hat{\Theta})$. Thus, classical constrained optimization methods (e.g., see [1, 20]) are not suitable for our problem. Besides formulation (2), the denoising problem could be formulated as a bilevel optimization problem (e.g., see [30]) as follows:

$$(5) \quad \begin{cases} \hat{x} = \operatorname{argmin}_x F(x, \Theta), \\ \hat{\Theta} = \operatorname{argmin}_{\Theta \geq 0} T(\Theta, \hat{x}). \end{cases}$$

However, because of the complexity of bilevel optimization (strongly NP-hard [15]) and the goal of developing a fast method, we transformed the lower level program in (5) in a constraint for the upper program as shown in (2). Finally, we remark that, although the similarity with parametric optimization (e.g., see [18, 24]), in our model the constraint is not parameterized by Θ , which, moreover, is not used to optimize F .

We now show some properties of the solution of the unconstrained problem. The proof of the following proposition is given in the appendix.

PROPOSITION 1. *The necessary conditions (NC) which a minimum of $F(x, \Theta)$ must satisfy can be written as*

$$(6) \quad (I + \Theta W(x))x = y$$

and the Hessian matrix of $F(x, \Theta)$ is

$$(7) \quad H(x, \Theta) = 2I + \Theta V(x),$$

where $W(x)$ and $V(x)$ have the following properties:

1. $W(x)$ is a symmetric nonnegative definite matrix of rank $(n-1)$, and the null eigenvalue corresponds to the eigenvector $u = [1, \dots, 1]^T$;
2. $V(x)$ admits the factorization $V(x) = D^T G(x) D$, where the eigenvalues of $D^T D$ are $\tilde{\lambda}_j = 2[2 - \cos(\pi k/\sqrt{n}) - \cos(\pi h/\sqrt{n})]$; $j = k + \sqrt{n}h + 1$; $h, k = 0, \dots, \sqrt{n} - 1$.

Let us denote by $\lambda_j(x)$ the eigenvalues of $W(x)$. Then, $\forall \Theta \geq 0$, the eigenvalues of $I + \Theta W(x)$ are $1 + \Theta \lambda_j(x) > 0$. Hence, $I + \Theta W(x)$ is invertible $\forall \Theta \geq 0$ and $\forall x$, and the minimizers $\hat{x}$ satisfy the nonlinear system

$$(8) \quad \hat{x} = (I + \Theta W(\hat{x}))^{-1}y.$$

We then have the following lemma.

LEMMA 1. *The set $\hat{x} \in \operatorname{argmin}_x F(x, \Theta) \in \Pi = \{x : \sum_{h=1}^n x_h = \sum_{h=1}^n y_h\} \forall \Theta$.*

Proof. By (6),

$$u^T \hat{x} + u^T \Theta W(\hat{x}) = u^T y.$$

But $W(\hat{x})u = 0$ by property 1 of Proposition 1, and W is symmetric; hence, $u^T \hat{x} = u^T y$. $\square$

Without loss of generality, we can assume that $\sum_{h=1}^n y_h = 0$, and, therefore, $\sum_{h=1}^n \hat{x}_h = 0$.

LEMMA 2. *The set of global minimizers of F is nonempty and bounded. Moreover, there exists a compact convex set X which contains as internal points $\hat{x} \in \operatorname{argmin}_x F(x, \Theta) \forall \Theta, 0 \leq \Theta < +\infty$.*

Proof. First we prove that each member of the family of functionals $F(x, \Theta)$ admits a set of global minimizers nonempty and bounded. By assumptions 1 and 2 on $\phi(t)$, $\lim_{k \rightarrow \infty} F(x^{(k)}, \Theta) = +\infty$ whenever $\lim_{k \rightarrow \infty} \|x^{(k)}\| \rightarrow \pm\infty$ for all sequences $\{x^{(k)}\} \subset \mathbb{R}^n$. Therefore, by [22, Prop. 4.3.2], all level sets of F are bounded, and hence, by [22, Prop. 4.3.1], as F is continuous, the set of global minimizers of F is nonempty and bounded. Let be $\hat{x} \in \operatorname{argmin}_x F(x, \Theta)$. We want to show that $\|\hat{x}\|_\infty \leq \|y\|_\infty$, i.e., $y_{\min} \leq \hat{x}_i \leq y_{\max}$. Let us suppose that this is false and consider a new vector $\tilde{x}$ such that

$$\tilde{x} = \begin{cases} \hat{x}_i & \text{if } \hat{x}_i \in [y_{\min}, y_{\max}], \\ y_{\min} & \text{if } \hat{x}_i < y_{\min}, \\ y_{\max} & \text{if } \hat{x}_i > y_{\max}. \end{cases}$$

It is easy to show that $F(\tilde{x}_i, \Theta) < F(\hat{x}_i, \Theta)$, therefore contradicting the fact that $\hat{x}$ is a minimizer. Therefore, $\|x\|_\infty \leq \|y\|_\infty$ and then $\|x\|_2 \leq \sqrt{n}\|y\|_2$. Hence, the set $S = \{x \mid \|x\|_2 \leq \sqrt{n}\|y\|_2\}$ contains all the minimizers. Thus, the ball $B = \{x \mid \|x\|_2 \leq \sqrt{n}\|y\|_2 + a\}$ for some $0 < a < +\infty$ contains all the minimizers as internal points. But all the minimizers are such that $\sum_{i=1}^n x_i = 0$, and thus we can choose the compact convex set X as the intersection between B and the hyperplane $\sum_{i=1}^n x_i = 0$. $\square$

2. The algorithm. Given a value $\hat{\Theta}$, $\hat{x} = \operatorname{argmin}_{x \in X} F(x, \hat{\Theta})$ is a solution of (2) whether $\tau - T(\hat{x}) \geq 0$. However, we are not guaranteed that for a generic $\hat{\Theta}$ the minimizer $\hat{x}$ is such that $\tau - T(\hat{x}) \geq 0$. Therefore, we propose solving (2) by means of an iterative algorithm which achieves feasibility in a finite number of steps. We notice that to a solution of problem (2) perfectly consistent with the hypothesis on the noise $\tau = 0$ should correspond. It will be clear in what follows that this constraint could require a long computation time. On the opposite direction, if $\tau = 1$, the solution of problem (2) will be trivially $x = y$ and $\Theta = 0$. In order to solve problem (2) let us reformulate the constraint $T(x) \leq \tau$ as $\sigma(1 - \tau) \leq \xi(x) \leq \sigma(1 + \tau)$, where $\xi(x) = \sqrt{\frac{q(x)}{n}}$.

Let us assume that $F(x, \Theta)$ has a unique minimizer. We then observe and prove in Theorem 1 that there exists a $\tilde{\Theta}$ such that the minimizer of $F(x, \tilde{\Theta})$ produces a value of ξ larger than σ . Moreover, when $\Theta = 0$, the minimizer of $F(x, 0)$ is y to which $\xi(y) = 0$ corresponds. It is therefore reasonable to expect that when Θ moves from $\tilde{\Theta}$ to zero, the minimizer $\hat{x}$ of $F(x, \Theta)$ is such that the corresponding $\xi(\hat{x})$ moves along a curve crossing the region defined by the constraint. In order to solve the problem we need to control the decrease of Θ to ensure that at least one point of the curve

TABLE 1
The algorithm FADN.

Step 1. $p^{(0)} < \frac{\sigma^2 \tau^2 n d}{2MN}$; $\Theta^{(1)} = p^{(0)} \frac{q(x^{(0)})}{\Psi(x^{(0)})}$, where $x^{(0)} = u + \alpha e_{[n/2]}$ with $\alpha > 0$ such that $\xi(x^{(1)}) > \sigma$ and $x^{(1)} = \operatorname{argmin}_x F(x, \Theta^{(1)})$; $\delta > 1$; $i=1$;
Step 2. while $(\tau - T(x^{(i)})) < 0$ do
 $p^{(i)} = \frac{p^{(i-1)}}{\delta}$;
 $i = i + 1$;
 $\Theta^{(i)} = p^{(i-1)} \frac{q(x^{(i-1)})}{\Psi(x^{(i-1)})}$;
 $x^{(i)} = \operatorname{argmin}_x F(x, \Theta^{(i)})$;

belongs to the region defined by the constraint. The sketch of the algorithm, denoted by FADN (Fast Automatic Denoising), is given in Table 1.

The finite positive constants M, N, d in Step 1 are defined in section 3.1, while u is $[1, \dots, 1]^T$, and e_k is the k -column of the identity matrix. The existence of a constant α with the required property is stated in Theorem 2.

Remark 2. We notice that in order to properly define the ratios $\frac{q(x^{(i)})}{\Psi(x^{(i)})}$ we need to exclude the case in which the minimizer of $F(x, \Theta)$ is a vector $\hat{x}_c$ with equal components because $\Psi(\hat{x}_c) = 0$. By Lemma 1 the only constant minimizer is the null vector which can be obtained from NC conditions (6) only when $y = 0$. We need therefore to make the obvious assumption that $y \neq 0$.

The fact that $\Psi(x) > 0$ implies that $\Theta^{(i)} < +\infty$ and in particular there exists a $\Theta_{max} < +\infty$ such that $\Theta^{(i)} \leq \Theta_{max}$, as proved in the next lemma. This fact will be used to prove the convergence of the algorithm in the next section.

LEMMA 3. *There exists $\Theta_{max} < +\infty$ such that $\Theta^{(i)} \leq \Theta_{max} \forall i \geq 1$.*

Proof. For $i \geq 1$ we have

$$|\Theta^{(i+1)} - \Theta^{(i)}| = \left| p^{(i)} \frac{q(x^{(i)})}{\Psi(x^{(i)})} - p^{(i-1)} \frac{q(x^{(i-1)})}{\Psi(x^{(i-1)})} \right|.$$

We consider the two cases $p^{(i)} \frac{q(x^{(i)})}{\Psi(x^{(i)})} \leq p^{(i-1)} \frac{q(x^{(i-1)})}{\Psi(x^{(i-1)})}$ and $p^{(i)} \frac{q(x^{(i)})}{\Psi(x^{(i)})} > p^{(i-1)} \frac{q(x^{(i-1)})}{\Psi(x^{(i-1)})}$. In the first case we have that

$$\begin{aligned} \left| p^{(i)} \frac{q(x^{(i)})}{\Psi(x^{(i)})} - p^{(i-1)} \frac{q(x^{(i-1)})}{\Psi(x^{(i-1)})} \right| &= p^{(i-1)} \frac{q(x^{(i-1)})}{\Psi(x^{(i-1)})} - p^{(i)} \frac{q(x^{(i)})}{\Psi(x^{(i)})} \\ &\leq p^{(i-1)} \frac{q(x^{(i-1)})}{\Psi(x^{(i-1)})} \leq p^{(i-1)} \cdot N, \end{aligned}$$

where $N = \max_{x \in X} \left| \frac{q(x)}{\Psi(x)} \right| < \infty$. In the second case, we have that

$$\left| p^{(i)} \frac{q(x^{(i)})}{\Psi(x^{(i)})} - p^{(i-1)} \frac{q(x^{(i-1)})}{\Psi(x^{(i-1)})} \right| = p^{(i)} \frac{q(x^{(i)})}{\Psi(x^{(i)})} - p^{(i-1)} \frac{q(x^{(i-1)})}{\Psi(x^{(i-1)})} \leq p^{(i)} \frac{q(x^{(i)})}{\Psi(x^{(i)})} \leq p^{(i)} \cdot N.$$

This means that

$$\left| p^{(i)} \frac{q(x^{(i)})}{\Psi(x^{(i)})} - p^{(i-1)} \frac{q(x^{(i-1)})}{\Psi(x^{(i-1)})} \right| \leq \max\{p^{(i)} \cdot N, p^{(i-1)} \cdot N\} < p^{(i-1)} \cdot N = \frac{N \cdot p^{(0)}}{\delta^{i-1}}.$$

Now we assume to execute infinitely many iterations and compute an upper bound on $\lim_{i \rightarrow \infty} \Theta^{(i)}$. Since, in the worst case,

$$\Theta^{(i)} \leq \Theta^{(1)} + Np^{(0)} \sum_{j=0}^{i-2} \frac{1}{\delta^j} \quad \text{and} \quad \sum_{j=0}^{\infty} \delta^{-j} = \frac{\delta}{\delta-1},$$

we have $\Theta_{max} = \Theta^{(1)} + Np^{(0)} \frac{\delta}{\delta-1}$. $\square$

In order to prove that it is possible to initialize the algorithm, we need to make the assumption $\|y\|_2^2 > n\sigma^2$, which is supported by the following lemma, whose proof is in the appendix.

LEMMA 4. *If $y \sim N(x, \sigma^2 I_n)$ and $\lim_{n \rightarrow \infty} \|x\|_2^2 < \infty$, then $\forall \epsilon > 0$, there exists n_ϵ such that, $\forall n > n_\epsilon$, $P(\|y\|_2^2 > \sigma^2 n) \geq 1 - \epsilon$.*

We can now show that it is possible to correctly initialize the algorithm *FADN* by means of the following two theorems.

THEOREM 1. *Let be $\rho = \min_{x \in X} \min_{j=2, \dots, n} \lambda_j(x)$, where $\lambda_j(x)$ are the eigenvalues of $W(x)$, and let $\|y\|_2^2 > n\sigma^2$. Then $\tilde{\Theta} = \frac{\sqrt{n}\sigma}{\rho(\|y\|_2 - \sqrt{n}\sigma)}$ is such that $\xi(x^{(1)}) > \sigma$, where $x^{(1)} = \operatorname{argmin}_{x \in X} F(x, \tilde{\Theta})$.*

Proof. By (8) a minimizer $\hat{x}$ must satisfy $\hat{x} = (I + \Theta W(\hat{x}))^{-1}y$. Hence

$$q(\hat{x}) = \|(I + \Theta W(\hat{x}))^{-1} - I\|y\|_2^2 = \sum_{j=1}^n z_j(\hat{x})^2 \frac{\Theta^2 \lambda_j(\hat{x})^2}{(1 + \Theta \lambda_j(\hat{x}))^2},$$

where $z(\hat{x}) = U^T(\hat{x})y$, and, by Proposition 1, $U(\hat{x}) = [u, u_2(\hat{x}), \dots, u_n(\hat{x})]$ is the matrix of the eigenvectors of $W(\hat{x})$ corresponding to the eigenvalues $\{0, \lambda_2(\hat{x}), \dots, \lambda_n(\hat{x})\}$. We notice that $y^T U U^T y = \sum_{j=2}^n z_j^2$. Therefore, as the functions $p_j(\Theta, \lambda_j(\hat{x})) = \frac{\Theta^2 \lambda_j(\hat{x})^2}{(1 + \Theta \lambda_j(\hat{x}))^2}$ are monotonic increasing of both arguments taking values in $[0, 1)$, we have

$$q(\hat{x}) \geq \sum_{j=2}^n z_j(\hat{x})^2 p_j(\Theta, \rho) = \|y\|_2^2 \frac{\Theta^2 \rho^2}{(1 + \Theta \rho)^2}.$$

Thus $\tilde{\Theta} = \frac{\sqrt{n}\sigma}{\rho(\|y\|_2 - \sqrt{n}\sigma)}$ is such that $\frac{\tilde{\Theta}^2 \rho^2}{(1 + \tilde{\Theta} \rho)^2} = \frac{n\sigma^2}{\|y\|_2^2}$. But then if $x^{(1)} = \operatorname{argmin}_{x \in X} F(x, \tilde{\Theta})$, we have $\xi(x^{(1)}) \geq \sigma$. $\square$

THEOREM 2. *Let $x^{(0)} = u + \alpha e_{[n/2]}$. Given $\Theta^{(1)} > 0$, there exists $\alpha > 0$ such that $\Theta^{(1)} = p^{(0)} \frac{q(x^{(0)})}{\Psi(x^{(0)})}$.*

Proof. Let $h = [n/2]$, where $[\]$ indicates the integer part.

$$q(x^{(0)}) = \|y - u - \alpha e_h\|_2^2 = \sum_{i \neq h}^n (y_i - 1)^2 + (y_h - 1 - \alpha)^2.$$

But $\Psi(x^{(0)}) = 8 \cdot \phi(\alpha)$. Hence

$$\Theta^{(1)}(\alpha) = p^{(0)} \frac{\sum_{i \neq h} (y_i - 1)^2 + (y_h - 1 - \alpha)^2}{8 \cdot \phi(\alpha)}$$

and $\lim_{\alpha \rightarrow +\infty} \Theta^{(1)}(\alpha) = +\infty$ because of assumption 6 made on $\phi(t)$. Therefore, for every $0 < a < +\infty$, there exists α_a such that $\Theta^{(1)}(\alpha_a) = a$. $\square$

3. Convergence analysis.

3.1. The convex case. In this section we prove the convergence of the algorithm *FADN* to a solution of problem (2) when ϕ (and hence F) is a convex function.

LEMMA 5. *If ϕ is convex and $\lim_{i \rightarrow \infty} p^{(i)} = 0$, then $\{q(x^{(i)})\} \rightarrow 0$.*

Proof. As F is convex on the compact convex set X , the minimum $x^{(i+1)}$ is global. Therefore, since

$$F(x^{(i+1)}, \Theta^{(i+1)}) \leq F(x^{(i)}, \Theta^{(i+1)})$$

and $\Theta^{(i+1)}\Psi(x^{(i+1)}) \geq 0$ we have that

$$\begin{aligned} q(x^{(i+1)}) &\leq q(x^{(i+1)}) + \Theta^{(i+1)}\Psi(x^{(i+1)}) \leq q(x^{(i)}) + \Theta^{(i+1)}\Psi(x^{(i)}) \\ (9) \quad &= q(x^{(i)}) + p^{(i)}q(x^{(i)}) = q(x^{(i)})[1 + p^{(i)}]. \end{aligned}$$

Now define a new sequence $\tilde{q}$ as follows: $\tilde{q}(x^{(0)}) = q(x^{(0)})$ and

$$\tilde{q}(x^{(i+1)}) = \begin{cases} q(x^{(i+1)}) & \text{if } q(x^{(i+1)}) < q(x^{(i)}), \\ q(x^{(i)}) & \text{if } q(x^{(i+1)}) \geq q(x^{(i)}). \end{cases}$$

Let us consider the following two cases:

if $\tilde{q}(x^{(i+1)}) = q(x^{(i+1)})$, then trivially $|\tilde{q}(x^{(i+1)}) - q(x^{(i+1)})| = 0$;

if $\tilde{q}(x^{(i+1)}) = q(x^{(i)})$, then

$$|\tilde{q}(x^{(i+1)}) - q(x^{(i+1)})| = |q(x^{(i)}) - q(x^{(i+1)})| = q(x^{(i+1)}) - q(x^{(i)})$$

as $q(x^{(i+1)}) \geq q(x^{(i)})$.

By (9) it follows that $q(x^{(i+1)}) - q(x^{(i)}) \leq q(x^{(i)}) \cdot p^{(i)} \leq Q \cdot p^{(i)}$, where $Q = \max_{x \in X} q(x) < \infty$ because X is compact and $q(x)$ is continuous, and therefore $|\tilde{q}(x^{(i+1)}) - q(x^{(i+1)})| \leq q(x^{(i+1)}) - q(x^{(i)}) \leq Q \cdot p^{(i)}$.

Thus, since by hypothesis $\lim_{i \rightarrow \infty} p^{(i)} = 0$, we have

$$\lim_{i \rightarrow \infty} |\tilde{q}(x^{(i+1)}) - q(x^{(i+1)})| = 0.$$

As $\{\tilde{q}(x^{(i)})\}$ is bounded and monotone not increasing, it admits a lower bound ℓ , i.e., $\tilde{q}(x^{(i)}) \rightarrow \ell$. But then,

$$|q(x^{(i)}) - \ell| = |q(x^{(i)}) - \ell + \tilde{q}(x^{(i)}) - \tilde{q}(x^{(i)})| \leq |\tilde{q}(x^{(i)}) - \ell| + |q(x^{(i)}) - \tilde{q}(x^{(i)})|.$$

As the last two terms tend to zero, $q(x^{(i)}) \rightarrow \ell$. But then

$$\lim_{i \rightarrow \infty} \Theta^{(i+1)} = \lim_{i \rightarrow \infty} q(x^{(i)}) \frac{p^{(i)}}{\Psi(x^{(i)})} = \ell \lim_{i \rightarrow \infty} \frac{p^{(i)}}{\Psi(x^{(i)})},$$

which tends to zero as by hypothesis $\lim_{i \rightarrow \infty} p^{(i)} = 0$ and $\min_{x \in X} \Psi(x) > 0$ (see Remark 2). But if $\Theta = 0$, the problem reduces to find the minimum of the quadratic form $q(x)$ whose solution is $x = y$ and $q(y) = 0$. Therefore, as $\lim_{i \rightarrow \infty} \Theta^{(i)} = 0$, we have that $\lim_{i \rightarrow \infty} x^{(i)} = y$ and $\ell = 0$. $\square$

Before stating the next lemma we notice that there exist finite constants M and d such that $M = \max_{x \in X} \Psi(x)$ and $d = \min_{\hat{x} \in X, \Theta \leq \Theta_{max}} \lambda_1[H(\hat{x}, \Theta)]$, where $H(\hat{x}, \Theta)$ is the Hessian matrix of $F(x, \Theta)$ computed in $\hat{x}$ and $\lambda_1[H(x, \Theta)]$ is its minimum

eigenvalue. This follows from the compactness of X , the continuity of $\Psi(x)$ and $\lambda_1[H(x, \Theta)]$, and the fact that Θ is definitely less than a finite constant Θ_{max} as proved in Lemma 3.

LEMMA 6. *If $i \geq 1$ and $|\Theta^{(i+1)} - \Theta^{(i)}| \leq \epsilon$, then $|\sqrt{q(x^{(i+1)})} - \sqrt{q(x^{(i)})}| < \sqrt{\frac{2M\epsilon}{d}}$.*

Proof. By definition

$$F(x, \Theta^{(i)}) = q(x) + \Theta^{(i)}\Psi(x)$$

and

$$F(x, \Theta^{(i+1)}) = q(x) + \Theta^{(i+1)}\Psi(x).$$

By subtracting the two equations and by $|\Theta^{(i+1)} - \Theta^{(i)}| \leq \epsilon$ we get

$$(10) \quad |F(x, \Theta^{(i)}) - F(x, \Theta^{(i+1)})| \leq \epsilon\Psi(x).$$

By developing $F(x, \Theta^{(i)})$ in Taylor series up to the first term, expressing the remainder in the Lagrange form, and noting that $\nabla F(x^{(i)}) = 0$ as $x^{(i)} = \operatorname{argmin}_{x \in X} F(x, \Theta^{(i)})$ is an internal point (see Lemma 2), we get

$$F(x, \Theta^{(i)}) = F(x^{(i)}, \Theta^{(i)}) + \frac{1}{2}(x - x^{(i)})^T H^{(i)}(c^{(i)})(x - x^{(i)}),$$

where $\min\{x_j, x_j^{(i)}\} \leq c_j^{(i)} \leq \max\{x_j, x_j^{(i)}\}$, $j = 1, \dots, n$, $H^{(i)}$ is the Hessian matrix of $F(x, \Theta^{(i)})$ which is positive definite as $x^{(i)}$ is a minimum point. But then,

$$\begin{aligned} G(x) &= F(x, \Theta^{(i)}) - F(x, \Theta^{(i+1)}) = F(x^{(i)}, \Theta^{(i)}) - F(x^{(i+1)}, \Theta^{(i+1)}) \\ &+ \frac{1}{2}(x - x^{(i)})^T H^{(i)}(c^{(i)})(x - x^{(i)}) - \frac{1}{2}(x - x^{(i+1)})^T H^{(i+1)}(c^{(i+1)})(x - x^{(i+1)}). \end{aligned}$$

Evaluating this expression in $x^{(i)}$ and $x^{(i+1)}$ we get

$$(11) \quad \begin{aligned} &G(x^{(i)}) - G(x^{(i+1)}) \\ &= -(x^{(i)} - x^{(i+1)})^T \left(\frac{H^{(i+1)}(c^{(i+1)}) + H^{(i)}(c^{(i)})}{2} \right) (x^{(i)} - x^{(i+1)}). \end{aligned}$$

Moreover,

$$(12) \quad A_{i+1,i} = \frac{H^{(i+1)}(c^{(i+1)}) + H^{(i)}(c^{(i)})}{2}$$

is positive definite. Then by (10), (11), and (12),

$$\begin{aligned} 0 &\leq (x^{(i)} - x^{(i+1)})^T A_{i+1,i} (x^{(i)} - x^{(i+1)}) = |G(x^{(i)}) - G(x^{(i+1)})| \\ &\leq |G(x^{(i)})| + |G(x^{(i+1)})| \leq \epsilon|\Psi(x^{(i)})| + \epsilon|\Psi(x^{(i+1)})| \leq 2M\epsilon. \end{aligned}$$

Therefore, we have

$$(x^{(i)} - x^{(i+1)})^T A_{i+1,i} (x^{(i)} - x^{(i+1)}) \leq 2M\epsilon.$$

Let $A = UDU^T$ be the spectral decomposition of $A_{i+1,i}$, where we dropped the index i on the matrices A , U , and D for simplifying the notations. As A is positive definite, U is orthogonal and $D = \text{diag}\{d_1, \dots, d_n\}$, $0 < d_1 \leq d_2 \leq \dots \leq d_n$. Therefore, letting $w = U^T(x^{(i)} - x^{(i+1)})$ we have

$$\begin{aligned} 2M\epsilon &\geq (x^{(i)} - x^{(i+1)})^T UDU^T (x^{(i)} - x^{(i+1)}) = w^T Dw = \sum_{h=1}^n w_h^2 d_h \\ &\geq d_1 \sum_{h=1}^n w_h^2 = d_1 w^T w = d_1 \|x^{(i)} - x^{(i+1)}\|^2. \end{aligned}$$

Therefore

$$\|x^{(i)} - x^{(i+1)}\| \leq \sqrt{\frac{2M\epsilon}{d}}.$$

But then (see, e.g., [27, pp. 163, 1.4])

$$\begin{aligned} \left| \sqrt{q(x^{(i+1)})} - \sqrt{q(x^{(i)})} \right| &= \left| \|x^{(i)} - y\| - \|x^{(i+1)} - y\| \right| \leq \|x^{(i)} - y - x^{(i+1)} + y\| \\ &= \|x^{(i)} - x^{(i+1)}\| \leq \sqrt{\frac{2M\epsilon}{d}}. \quad \square \end{aligned}$$

Lemma 6 allows us to control the variation of $\sqrt{q(x^{(i)})}$ between two consecutive iterations as a function of the variation of $\Theta^{(i)}$. By using Lemmas 5 and 6 we can now prove that the value $T(x^{(i)})$ can be controlled in order to satisfy the constraint.

LEMMA 7. *There exists an iteration $i \geq 1$ such that, if $\xi(x^{(1)}) > \sigma$, then $T(x^{(i)}) \leq \tau$.*

Proof. By Lemma 3 we know that, when $i \geq 1$,

$$|\Theta^{(i+1)} - \Theta^{(i)}| \leq \frac{N \cdot p^{(0)}}{\delta^{i-1}} < N \cdot p^{(0)}.$$

Therefore, let $N \cdot p^{(0)} = \epsilon$. Now, by Lemma 6, $\frac{1}{\sigma} \left| \sqrt{\frac{q(x^{(i+1)})}{n}} - \sqrt{\frac{q(x^{(i)})}{n}} \right| < \frac{1}{\sigma} \sqrt{\frac{2M\epsilon}{nd}}$. It follows that if $p^{(0)} = \frac{\epsilon}{N} < \frac{\sigma^2 \tau^2 nd}{2MN}$, then

$$\frac{1}{\sigma} \sqrt{\frac{2M\epsilon}{nd}} = \frac{1}{\sigma} \sqrt{\frac{2MNp^{(0)}}{nd}} < \tau.$$

Hence, $\frac{1}{\sigma} \left| \sqrt{\frac{q(x^{(i+1)})}{n}} - \sqrt{\frac{q(x^{(i)})}{n}} \right| < \tau$.

Moreover, since by hypothesis at iteration 1 we start with a $x^{(1)}$ such that $\frac{1}{\sigma} \sqrt{\frac{q(x^{(1)})}{n}} > 1$, and by Lemma 5, $\lim_{i \rightarrow \infty} \frac{1}{\delta} \sqrt{\frac{q(x^{(i)})}{n}} = 0$, there must exist an iteration i such that

$$\frac{1}{\sigma} \left| \sqrt{\frac{q(x^{(i)})}{n}} - \sigma \right| \leq \tau. \quad \square$$

We can now prove the convergence theorem.

THEOREM 3. *If ϕ is convex, the algorithm FADN converges to $(\hat{x}, \hat{\Theta})$ such that $\hat{x} = \operatorname{argmin}_{x \in X} F(x, \hat{\Theta})$ and $\tau - T(\hat{x}) \geq 0$. Moreover, $\hat{x}$ is a global minimum of $F(x, \hat{\Theta})$.*

Proof. By Lemma 4 and Theorems 1 and 2, we are able to correctly initialize the algorithm in such a way that $\frac{1}{\sigma} \sqrt{\frac{q(x^{(1)})}{n}} > 1$. Hence, by Lemma 7 there exists an iteration i such that $T(x^{(i)}) \leq \tau$, where $x^{(i)} = \operatorname{argmin}_{x \in X} F(x, \Theta^{(i)})$. Taking $\hat{\Theta} = \Theta^{(i)}$, $\hat{x} = x^{(i)}$ is the unique global minimum of $F(x, \hat{\Theta})$ since F is convex on the compact convex set X . $\square$

3.2. The nonconvex case. In this section we assume that ϕ is nonconvex. In this case it is not possible in general to say anything about the convergence of FADN. The critical point is that the curve $\hat{x}(\Theta)$ is not a function since it might bifurcate. However, recalling that from Lemma 3 there exists $\Theta_{max} < +\infty$ such that $\Theta^{(i)} \leq \Theta_{max} \forall i \geq 1$, we have the following theorem.

THEOREM 4. *If ϕ is nonconvex, there exists $\Theta_{lim} > 0$ such that if one of the following conditions holds:*

- (a) $\Theta_{lim} \geq \Theta_{max}$,
- (b) $\Theta_{lim} < \Theta_{max}$ and $F(x, \Theta)$ has a unique minimizer when $\Theta \in [0, \Theta_{max}]$,

then FADN converges.

Proof.

(a) Since $q(x)$ is convex and F is continuous in Θ , there must exist an interval $[0, \Theta_{lim}]$ in which F is convex. In this case if $\Theta_{lim} \geq \Theta_{max}$ Lemmas 5, 6, and, hence, 7 hold. Moreover FADN can be initialized because Lemma 4 and Theorems 1 and 2 hold. Hence one can conclude that the algorithm converges even if ϕ is nonconvex. To find Θ_{lim} we observe that by Proposition 1 the Hessian matrix of $F(x, \Theta)$ is $H(x, \Theta) = 2I + \Theta D^T G(x) D$. Let us compute the largest value of Θ that makes $H(x, \Theta)$ positive definite. By the assumptions on ϕ , there exists a constant $c > 0$ such that $\phi''(t/\eta) \geq -c$. From [4, p. 132] the “worst case” for $H(x, \Theta)$ occurs when x is such that $G(x) = -2cI$. But, in this case, $H(x, \Theta) = 2I - 2c\Theta D^T D$. Hence, the minimum eigenvalue of $H(x, \Theta)$ in this case is $\rho = 2 - 2c\Theta \tilde{\lambda}_n$, where $\tilde{\lambda}_n$ is the maximum eigenvalue of $D^T D$ given in Proposition 1. In order to have $\rho > 0$ we need to take $\Theta_{lim} = \frac{1}{c\tilde{\lambda}_n}$.

(b) If the function F has a unique minimizer, Lemmas 5, 6, and, hence, 7 hold. Moreover, FADN can be initialized because Lemma 4 and Theorems 1 and 2 hold. Hence FADN converges. $\square$

Remark 3. We show in the appendix that the bound $\Theta_{lim} = \frac{1}{c\tilde{\lambda}_n}$ is tight. However, for a given y , F could have a unique minimum also if $\Theta \geq \Theta_{lim}$; but if one looks for a condition independent on the data, the convexity condition is the most general one.

4. The heuristic algorithm. In this section we describe a heuristic variant of the proposed algorithm in order to reduce the possible slow down produced by the computation of the minimum of F at each iteration and to get rid of the threshold τ . We notice in fact that the computational burden of FADN is concentrated on the computation of the minimum $x^{(i)}$. The idea is to compute sufficiently good values $x^{(i)}$ without necessarily getting the minimum of F . This can be accomplished by a procedure, called *DESCENT*, which, given $x^{(i)}$ and $\Theta^{(i)}$, computes an approximation $x^{(i+1)}$ of $\operatorname{argmin}_x F(x, \Theta^{(i)})$. In the following, we use, componentwise, a descent method based on auxiliary variables, which, in the convex case, converges to the

unique minimum of F [5]. Moreover, the function $T(x)$ is used as a test for the solutions offered by *DESCENT*: as far as $T(x^{(i+1)}) < T(x^{(i)})$ the algorithm proceeds as described above, while if $T(x^{(i+1)}) \geq T(x^{(i)})$, a second procedure called *NEWTON*, is used, starting from $x^{(i)}$ and keeping Θ fixed to the value $\Theta^{(i)}$. The procedure *NEWTON* is supposed to produce a sequence $x^{(i)}$ converging more quickly to an approximation of the minimum of F . We assume in fact that $T(x)$ has reached a (local) minimum at iteration i ; therefore we keep the function fixed and try to minimize it as quickly as possible. To this aim we use as *NEWTON* the Newton method (see, e.g., [1, 20]) componentwise. The heuristic algorithm, denoted by *HFADN*, is summarized in Table 2.

TABLE 2
The heuristic algorithm *HFADN*.

Step 1. $\Theta^{(1)} = p^{(0)} \frac{q(x^{avr})}{\Psi(x^{avr})}$; $p^{(0)} = 1$; $\delta > 1$; $x^{(0)} = y$; $i = 1$; $T(x^{(0)}) = +\infty$; Fix a maximum number of iterations; $x^{(1)} = \text{DESCENT}(x^{(0)}, \Theta^{(1)})$; Step 2. While $(T(x^{(i)}) < T(x^{(i-1)}))$ do Step 2.1. $\Theta^* = \Theta^{(i)}$ and $x^* = x^{(i)}$; Step 2.2. If the maximum number of iterations is met then stop; Step 2.3. $p^{(i)} = \frac{p^{(i-1)}}{\delta}$; Step 2.4. $i = i + 1$; Step 2.5. $\Theta^{(i)} = p^{(i-1)} \frac{q(x^{(i-1)})}{\Psi(x^{(i-1)})}$; Step 2.6. $x^{(i)} = \text{DESCENT}(x^{(i-1)}, \Theta^{(i)})$; Step 3. $x^{(i)} = x^*$, $i = i + 1$; Step 4. $x^{(i)} = \text{NEWTON}(x^{(i-1)}, \Theta^*)$; Step 5. While $(T(x^{(i)}) < T(x^{(i-1)}))$ do; Step 5.1. $i = i + 1$. Step 5.2. $x^{(i)} = \text{NEWTON}(x^{(i-1)}, \Theta^*)$;

In Step 1 the initial value of Θ is computed as follows. Each component of x is set to the average value over the neighbors of the corresponding component of y . If $x^{(avr)}$ is the resulting vector, we compute $\Theta^{(0)} = p^{(0)} \frac{q(x^{avr})}{\Psi(x^{avr})}$ in order to balance the contribution of $q(x)$ and $\Psi(x)$.

Remark 4. The complexity per iteration of the algorithm is linear in the data size. In fact the computational burden is concentrated in the evaluation of *DESCENT* and *NEWTON*. For both procedures this consists of visiting the pixels in a given order and for each pixel i performing a few steps (of *DESCENT* or *NEWTON*, respectively) toward the minimum of the local function

$$(13) \quad F_i(\zeta) = (y_i - \zeta)^2 + \Theta \sum_{\langle i, j \rangle} \phi \left(\frac{\zeta - x_j}{\eta} \right).$$

Assuming that the neighborhood of each pixel consists of a small fixed number of pixels (e.g., $O(1)$), performing a *DESCENT* or *NEWTON* step requires a number of flops proportional to the cardinality of the neighborhood of each pixel. Therefore, the cost of a full sweep is linear in the number of pixels.

Remark 5. We notice that when $\phi(t) = \log(1 + t^2)$ a *DESCENT* step is similar to a step of Perona–Malik [23] nonlinear diffusion algorithm. In fact the updating of pixel j by *DESCENT* can be written as

$$x_j^{(i+1)} = \frac{1}{C\sigma^2} y_j + \left[\frac{2\Theta^{(i)}}{C} \sum_{\langle j, h \rangle} \frac{1}{\eta^2 + (x_j^{(i)} - x_h^{(i)})^2} \right] x_j^{(i)} + \frac{2\Theta^{(i)}}{C} \sum_{\langle j, h \rangle} \frac{(x_h^{(i)} - x_j^{(i)})}{\eta^2 + (x_j^{(i)} - x_h^{(i)})^2},$$

where

$$C = \frac{1}{\sigma^2} + 2\Theta^{(i)} \sum_{\langle j, h \rangle} \frac{1}{\eta^2 + (x_j^{(i)} - x_h^{(i)})^2}$$

while the corresponding step of the Perona–Malik algorithm is

$$x_j^{(i+1)} = x_j^{(i)} + \frac{\lambda \eta^2}{|\langle j, h \rangle|} \sum_{\langle j, h \rangle} \frac{(x_h^{(i)} - x_j^{(i)})}{\eta^2 + (x_j^{(i)} - x_h^{(i)})^2}, \quad \lambda \geq 0,$$

where $|\langle j, h \rangle|$ is the number of neighbors of pixel j . Therefore, *DESCENT* can be seen as a generalization of nonlinear diffusion where a data term is added and the weight of each term is adjusted in order to keep the constraint that their sum is one. Of course *DESCENT* is just a step of *HFADN* which therefore can be interpreted as a generalization of nonlinear diffusion where the parameter Θ is adaptively estimated.

5. Experimental results. In this section we report some experimental findings on the algorithm *HFADN*. We first show the behavior of *HFADN* on synthetic data and real data. Then we provide a comparison of *HFADN* with several standard denoising methods. In all cases the heuristic nonconvex algorithm was used. All tests were performed on a Pentium III PC with 700 MHz and 128 MB RAM.

Both synthetic and real data are time series of 2-dimensional images. We then define a 3-dimensional neighborhood system by considering spatial and temporal neighbors of each pixel (i, j, t) as follows:

$$N_s^{(i,j,t)} = \{(i-1, j-1), (i-1, j), (i-1, j+1), (i, j-1), \\ (i, j+1), (i+1, j-1), (i+1, j), (i+1, j+1)\}$$

and

$$N_t^{(i,j,t)} = \{t-1, t+1\}.$$

The family of functionals used in the experiments is

$$F(x, \beta_s, \beta_t) = \sum_{i,j,t} (y_{ijt} - x_{ijt})^2 + \beta_s \cdot \sum_{i,j,t} \sum_{(h,k) \in N_s^{(i,j,t)}} \phi\left(\frac{x_{ijt} - x_{hkt}}{\eta_s}\right) \\ + \beta_t \cdot \sum_{i,j,t} \sum_{\tau \in N_t^{(i,j,t)}} \phi\left(\frac{x_{ijt} - x_{ij\tau}}{\eta_t}\right),$$

where $\phi(t) = \log(1+t^2)$ and η_s and η_t are two hyperparameters that can be estimated by the method proposed in [26] or simply by computing the sample standard deviation of the first differences in space and time, respectively, of the averaged data $x^{(avr)}$, as proposed in [12]. The hyperparameter Θ is the vector (β_s, β_t) ; hence $p^{(0)} = 0.5$ in order to balance the contribution of $q(x)$ and $\Psi(x)$.

Synthetic data. Synthetic data x is a series of 10 images made of 50×50 pixels taking values in the set $\{0, 1\}$. Three data sets y were generated according to (1) by using $\sigma = 0.5, 1, 3$. Figure 1 reports the true synthetic data x and the three noisy data y .

Real data. To test the algorithm on real data we considered a time series of 2-dimensional magnetic resonance images representing a region of interest of a human

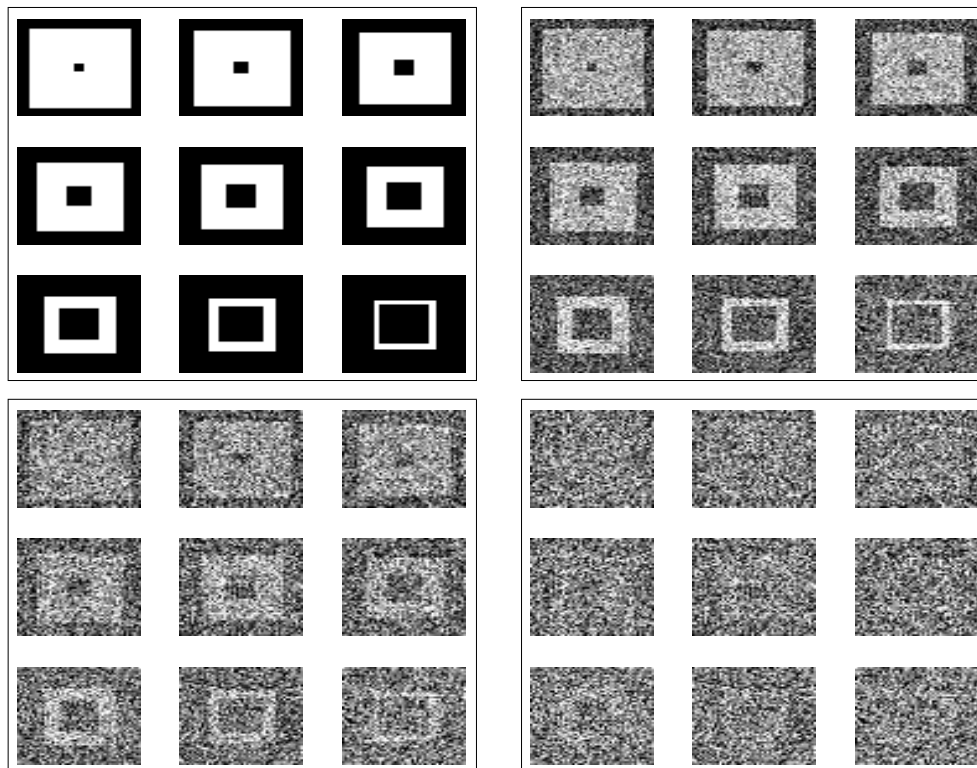


FIG. 1. The true synthetic $\{0,1\}$ -valued data (top left) and the true data corrupted with a Gaussian white noise with $\sigma = 0.5$ (top right), $\sigma = 1$ (bottom left), and $\sigma = 3$ (bottom right).

breast including a tumor. The image sequence represents the diffusion at different time instants of a contrast agent injected in the vessels.

Figure 2 reports the images associated with the real data set. It is worth mentioning that the set is a time series of 20 images each formed by 29×29 pixels with intensity in the range $\{0, 1300\}$ and the noise standard deviation is $\sigma = 74.33$.

5.1. Experimental analysis of the heuristic HFADN. In Figure 3 we report the output obtained by HFADN for the synthetic data set. In Table 3 we report the L_1 norm of the difference between the true and the restored time series of images, the values of δ used, the best values achieved for $T(x)$, the running times of the algorithm, and the number of iterations.

In Figure 4 we report the output obtained by HFADN on the real data set. Furthermore, as for synthetic data, in Table 3 we show the experimental parameters.

Remark 6. Note that the parameter δ is robust with respect to image variability; therefore, we are likely to get reasonable results even if δ is not accurately tuned provided that we guess the correct magnitude order. In fact, we notice that the same value of δ is used for all the synthetic images (see Table 3).

5.2. Comparison with different methods. In this subsection we experimentally compare HFADN with three known methods which share the feature to be fast: wavelets [6], median filter [16], and nonlinear diffusion [23, 31]. All methods were tuned to get the best possible result. For the sake of comparison with the other methods we experimented on a single image rather than on a time series of images

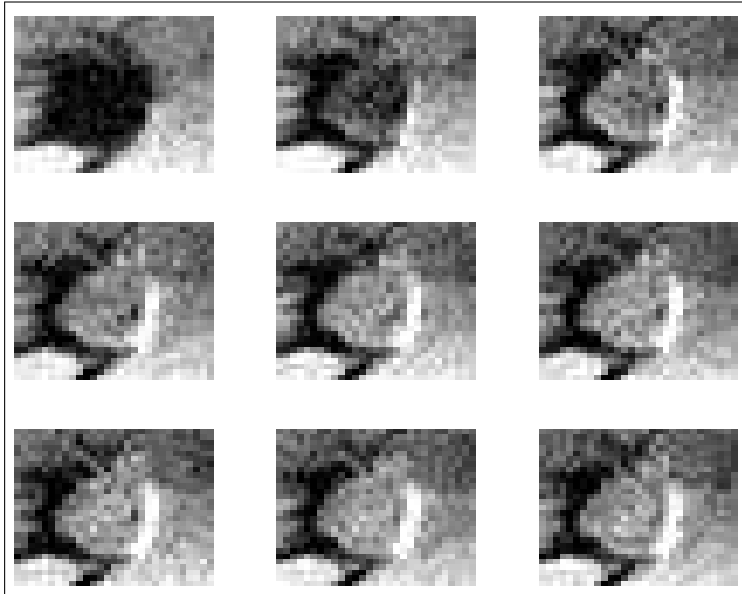


FIG. 2. The series of the first nine images associated with the real data set.

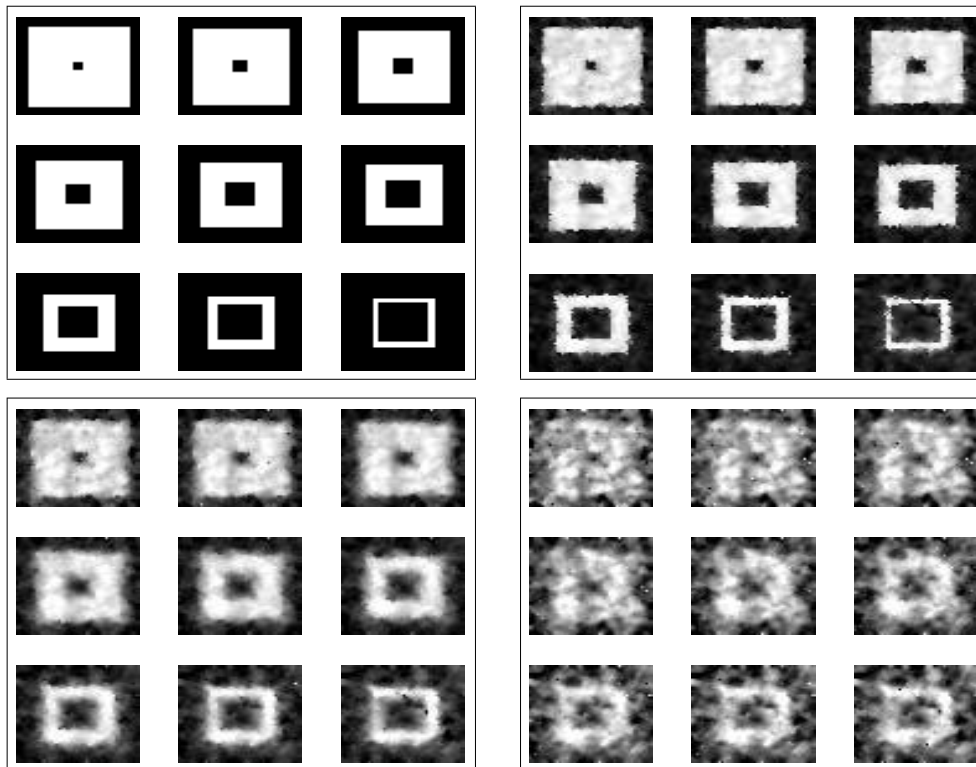


FIG. 3. The true synthetic data (top left) and HFADN restorations when $\sigma = 0.5$ (top right), when $\sigma = 1$ (bottom left), and when $\sigma = 3$ (bottom right).

TABLE 3

Experimental parameters of HFADN. The first three rows are related to as the nonconvex synthetic case and the last one as the real nonconvex case.

	σ	L_1	δ	$T(x)$	CPU time	Tot. iter.	DESCENT iter.	NEWTON iter.
	3	0.0203	10	0.01	9.12	12	6	6
Synthetic	1	0.0019	10	0.25	8.25	12	4	8
	0.5	0.0016	10	0.23	4.89	12	4	8
Real	74.33	-	20	0.18	6.89	19	7	12

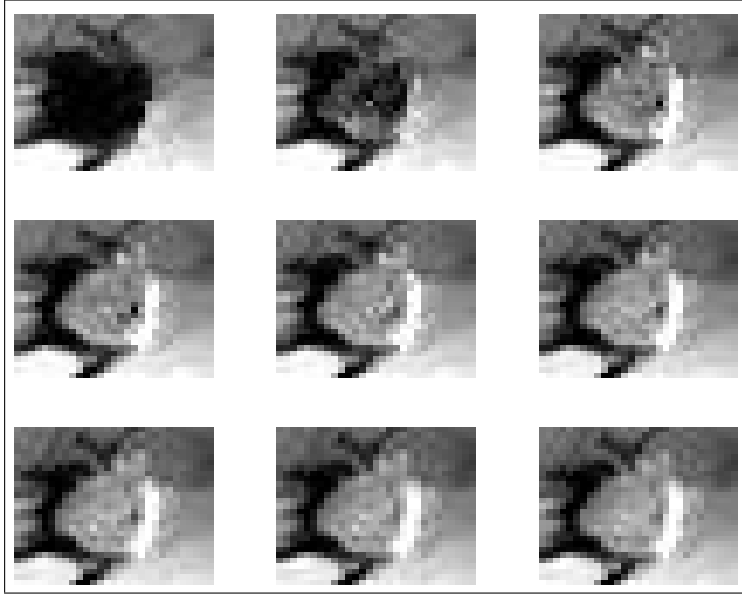


FIG. 4. Real data set: HFADN restoration of the first nine images.

as for the previous tests, thus neglecting the terms $\beta_t \cdot \sum_{i,j,t} \sum_{\tau \in N_t^{(i,j,t)}} \phi\left(\frac{x_{ijt} - x_{ij\tau}}{\eta_t}\right)$ in $F(x, \beta_s, \beta_t)$ and setting $p^{(0)} = 1$. We implemented nonlinear diffusion with the stopping criterion based on the minimization of the correlation of signal and noise in the filtered image proposed in [19].

Figure 5 reports the behavior of the algorithms on the well-known image of 512×512 pixels representing Lenna scaled in $\{0, 1\}$, which has been corrupted by a Gaussian white noise with $\sigma = \sqrt{0.05}$. The value $\delta = 3$ was used.

Table 4 reports $SNRs$ and running times of the four algorithms (notice that the running time of HFADN also includes the hyperparameter tuning). The SNR is a standard index for comparing pictures [16] and it is defined, e.g., by $SNR = 10 \log \frac{n(\max_j x_j)^2}{\|x - \hat{x}\|_2^2}$, where x_j is the true intensity of pixel j and $\hat{x}_j$ is the corresponding restored intensity. HFADN performed 6 iterations, getting a relative error $T(x) = 0.06$, while nonlinear diffusion had been stopped as soon as it reached the first local minimum, after 8 iterations.

Visual inspection shows that HFADN is able to restore some details of the true image better than the other three methods as confirmed by the $SNRs$ values in Table 4.

In [7] the magnetic resonance data set considered previously was restored by using



FIG. 5. The original (top left), the noisy image (top middle), HFADN reconstruction (top right), wavelets reconstruction (bottom left), median filter (bottom middle), and nonlinear diffusion (bottom right).

TABLE 4
SNRs and running times of the four algorithms.

Algorithm	SNR	CPU time (sec.)
HFADN	39.8	25
Wavelets	39.3	6
Nonlinear diffusion	38.9	18
Median filter	36.5	1

a Bayesian MAP estimator with a grid search to find the best hyperparameter value. Slightly better results are obtained by using *HFADN* which is, moreover, about seven times faster.

6. Conclusions and further research. We presented a fast and automatic algorithm for denoising arrays representing functions in any dimension. The algorithm solves a constrained optimization problem whose unknowns are the noise-free array and a set of hyperparameters which the objective function depends on linearly. However, usually, other nonlinear hyperparameters appear in the objective functions that in the present work were estimated from the data before starting the iterative process. It is therefore reasonable to conceive a method that includes these hyperparameters too in the optimization process estimating them adaptively. It is also possible to generalize the proposed method to the case of pixel-dependent regularization parameters, by imposing more stringent constraints. These topics require further research.

Appendix.

Proof of Proposition 1. Let us consider the model in dimension one, as the generalization to higher dimensions is straightforward by using tensor products. We then have

$$F(x, \Theta) = (x - y)^T(x - y) + 2\Theta u^T \underline{\phi}(D_0 x),$$

where $u \in \mathbb{R}^{n-1}$ is a vector of ones, $D_0 = [e_1, -e_1 + e_2, \dots, -e_{n-1} + e_n, -e_n]$, where e_j is the j th column of the identity matrix I_{n-1} and $\underline{\phi} \in \mathbb{R}^{n-1}$ has components $\phi(\frac{x_j - x_{j+1}}{\eta})$, $j = 1, \dots, n-1$. But then $\nabla F = 2(x - y) + 2\Theta D_0^T \underline{\phi}'$, where $\underline{\phi}'$ has components $\phi'(\frac{x_j - x_{j+1}}{\eta})$. Let us multiply and divide componentwise $\underline{\phi}'$ by $D_0 x$. We get $\underline{\phi}' = J(x) D_0 x$, where

$$J(x) = \text{diag} \left(\frac{\phi' \left(\frac{x_j - x_{j+1}}{\eta} \right)}{x_j - x_{j+1}} \right).$$

Therefore, $\frac{1}{2} \nabla F = (I + \Theta D_0^T J(x) D_0) x - y$, and $W(x) = D_0^T J(x) D_0$.

If $\frac{df}{dx}$ is the matrix whose (i, j) element is $\frac{\partial f_i}{\partial x_j}$ from $\nabla F = 2(x - y) + 2\Theta D_0^T \underline{\phi}'$ we get $\frac{d\nabla F}{dx} = H(x, \Theta) = 2I + \Theta D_0^T G(x) D_0$, where $G(x) = 2 \text{diag}(\phi''(\frac{x_j - x_{j+1}}{\eta}))$ and $V(x) = D_0^T G(x) D_0$.

It is easy to show that, in dimension two, $W(x)$ and $V(x)$ admit the factorization $W(x) = D^T J(x) D$ and $V(x) = D^T G(x) D$, with $J(x) = \text{diag}\{J_1(x), J_2(x)\}$, $D = [D_1^T D_2^T]^T$, where $J_1(x) = \text{diag}(\frac{\phi'(D_1 x / \eta)}{D_1 x})$ and $J_2(x) = \text{diag}(\frac{\phi'(D_2 x / \eta)}{D_2 x})$, and $D_1 = D_0 \otimes I_{\sqrt{n}}$, $D_2 = I_{\sqrt{n}} \otimes D_0$. Similarly, $G(x) = \text{diag}\{G_1(x), G_2(x)\}$, where $G_1(x) = 2(\frac{\phi''(D_1 x / \eta)}{D_1 x})$ and $G_2(x) = 2(\frac{\phi''(D_2 x / \eta)}{D_2 x})$. $\square$

Proof of Property 1. Property 1 follows by noticing that each row of D has two elements different from zero equal to 1 and -1 ; therefore, $\ker(J^{1/2} D)$ has dimension one because it contains only the constant vectors. Hence, $\text{rank}(J^{1/2} D) = n - 1$. But then the number of nonzero singular values of $J^{1/2} D$ is $n - 1$. As the eigenvalues of $W(x)$ are the singular values of $J^{1/2} D$ squared, the proof is concluded. $\square$

Proof of Property 2. $D^T D = (D_0^T D_0 \otimes I) + (I \otimes D_0^T D_0)$ and if μ_r are the eigenvalues of $D_0^T D_0$, then $\mu_h + \mu_k$, with $h, k = 1, \dots, \sqrt{n}$, are the eigenvalues λ of $D^T D$ (see [17, Chap. 12]). The eigenvalues μ_h are obtained by noticing that if $c = e_2 + e_{2\sqrt{n}}$, and $C \in \mathbb{R}^{2\sqrt{n} \times 2\sqrt{n}}$ is the circulant matrix whose first row is c , then $PCQ = I_2 \otimes A$, where P and Q are permutation matrices and $A = I_{\sqrt{n}} - \frac{1}{2} D_0^T D_0$. Hence, $\mu_h = 2(1 - v_h)$, where v_h are the eigenvalues of A which are the eigenvalues of C with multiplicity 1 (see [17]), i.e., $v_h = \cos \frac{2\pi h}{2\sqrt{n}} = \cos \frac{\pi h}{\sqrt{n}}$. $\square$

Proof of Lemma 4. Let us consider the estimator of σ^2 given by $\hat{\sigma}_n^2 = \frac{1}{n}(\|y\|_2^2 - \|x\|_2^2)$. As $E[\hat{\sigma}_n^2] = \sigma^2$ and it is easy to show that $\lim_{n \rightarrow \infty} \text{Var}[\hat{\sigma}_n^2] = 0$, it follows that $\hat{\sigma}_n^2$ is a consistent estimator of σ^2 . Therefore, given $a > 0$, $\forall \epsilon > 0$ there exists n_ϵ such that $\forall n > n_\epsilon$, $P(|\|y\|_2^2 - \|x\|_2^2 - \sigma^2 n| \geq a) < \epsilon$. But then $P(\|y\|_2^2 > \|x\|_2^2 + \sigma^2 n - a) > 1 - \epsilon$. Let us choose $0 < a < c = \lim_{n \rightarrow \infty} \|x\|_2^2 < \infty$. But then there exists m such that $\forall n > m$, $\|x\|_2^2 > a$. Let $p_\epsilon = \max\{m, n_\epsilon\}$. Then, $\forall n > p_\epsilon$, $1 - \epsilon < P(\|y\|_2^2 > \|x\|_2^2 + \sigma^2 n - a) \leq P(\|y\|_2^2 > \sigma^2 n)$. $\square$

Proof of the tightness of the bound on Θ_{lim} . We want to show that there exists at least one function $F(x, \Theta) = q(x) + \Theta \Psi(x)$ with Ψ nonconvex for which the maximum value of Θ such that F has only one minimum is Θ_{lim} .

Let us consider the one-dimensional case with two pixels, with $\phi(t) = \log(1 + t^2)$. We have $F(x, \Theta) = (x_1 - y_1)^2 + (x_2 - y_2)^2 + 2\Theta \log(1 + (x_1 - x_2)^2)$. In this case the eigenvalues of $D_0^T D_0$ are $\lambda_j = 2(1 - \cos(\frac{\pi j}{n}))$, and therefore, since $\phi''(t) \geq -c = -1/4$, we have that $\Theta_{lim} = \frac{4}{\lambda_n} = 2$. The NC (6) are

$$\begin{cases} 2(x_1 - y_1) + 4\Theta \frac{(x_1 - x_2)}{1 + (x_1 - x_2)^2} = 0, \\ 2(x_2 - y_2) - 4\Theta \frac{(x_1 - x_2)}{1 + (x_1 - x_2)^2} = 0. \end{cases}$$

Let $x_d = (x_1 - x_2)$ and $y_d = (y_1 - y_2)$; then we get the cubic equation

$$(14) \quad x_d^3 - y_d x_d^2 + x_d(4\Theta + 1) - y_d = 0$$

which admits one real solution if and only if

$$Q = \frac{q^2}{4} + \frac{p^3}{27} > 0,$$

where $p = -\frac{a^3}{3} + b$, $q = \frac{2}{27}a^3 - \frac{ab}{3} + c$ and $a = c = -y_d$, $b = 4\Theta + 1$. Hence we get

$$Q(y_d, \Theta) = \left(-\frac{y_d^2}{9} + \frac{1}{3} + \frac{4}{3}\Theta\right)^3 + \left[-\left(\frac{y_d}{3}\right)^3 + y_d\left(\frac{1+4\Theta}{6}\right) - \frac{y_d}{2}\right]^2.$$

To get a bound independent on the data, let us compute

$$\operatorname{argmin}_{y_d} Q(y_d, \Theta) = \frac{4}{27}\Theta(8 - 12\Theta + 6\Theta^2 - \Theta^3),$$

which is positive for $0 \leq \Theta < 2 = \Theta_{lim}$. $\square$

Acknowledgments. We wish to thank the anonymous referees for their useful comments which led to this improved version of the paper. Moreover, we wish to thank R. March and G. Sebastiani for many useful discussions, G. L. Torrisi for some technical help in proving Lemma 4, and M. Crecco and P. Gentile of “Polo Oncologico Istituto Regina Elena,” Rome, Italy, for providing the MRI data.

REFERENCES

- [1] D. P. BERTSEKAS, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1995.
- [2] J. BESAG, *Toward Bayesian image analysis*, J. Appl. Stat., 16 (1989), pp. 395–407.
- [3] M. BLACK, G. SAPIRO, D. MARIMONT, AND D. HEEGER, *Robust anisotropic diffusion*, IEEE Trans. Image Process., 7 (1998), pp. 421–432.
- [4] A. BLAKE AND A. ZISSERMAN, *Visual Reconstruction*, MIT Press, Cambridge, MA, 1987.
- [5] P. CHARBONNIER, L. BLANC-FERAUD, G. AUBERT, AND M. BARLAUD, *Deterministic edge-preserving regularization in computed imaging*, IEEE Trans. Image Process., 6 (1997), pp. 298–311.
- [6] R. R. COIFMAN AND D. L. DONOHO, *Translation-invariant de-noising in Wavelets and Statistics*, A. Antoniadis and G. Oppenheim, eds., Springer-Verlag, New York, 1995, pp. 125–150.
- [7] F. DE PASQUALE, P. BARONE, G. SEBASTIANI, AND J. STANDER, *Bayesian analysis of dynamic MR breast images*, Applied Statistics, 53 (2004), pp. 1–19.
- [8] D. DOBSON AND F. SANTOSA, *Recovery of blocky images from noisy and blurred data*, SIAM J. Appl. Math., 56 (1996), pp. 1181–1198.
- [9] N. GALATSANOS AND A. KATSAGGELOS, *Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation*, IEEE Trans. Image Process., IP1 (1992), pp. 322–336.
- [10] S. GEMAN AND D. E. MCCLURE, *Statistical methods for tomographic image reconstruction*, Bull. Inst. Internat. Statist., 52 (1987), pp. 5–21.
- [11] D. GEMAN AND G. REYNOLDS, *Constrained restoration and the recovery of discontinuities*, IEEE Trans. Pattern Anal. Machine Intell., 14 (1992), pp. 367–383.
- [12] I. K. GLAD AND G. SEBASTIANI, *A Bayesian approach to synthetic magnetic resonance imaging*, Biometrika, 82 (1995), pp. 237–250.
- [13] P. GREEN, *Bayesian reconstructions from emission tomography data using a modified EM algorithm*, IEEE Trans. Med. Imaging, 9 (1990), pp. 84–93.
- [14] C. HANSEN, *Analysis of discrete ill-posed problems by means of the L-curve*, SIAM Rev., 34 (1992), pp. 561–580.
- [15] P. HANSEN, B. JAUMARD, AND G. SAVARD, *New branch-and-bound rules for linear bilevel programming*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 1194–1217.

- [16] A. JAIN, *Fundamentals of Digital Image Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
- [17] P. LANCASTER AND M. TISMENETSKY, *The Theory of Matrices*, Academic Press, Orlando, FL, 1985.
- [18] G. D. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley & Sons, New York, 1969.
- [19] P. MRAZEK, *Selection of optimal stopping time for nonlinear diffusion filtering*, in Scale-Space and Morphology in Computer Vision, Proceedings of the 3rd International Conference Scale-Space 2001, M. Kerckhove, ed., Lecture Notes in Comput. Sci. 2106, Springer-Verlag, Berlin, pp. 290–298.
- [20] S. G. NASH AND A. SOFER, *Linear and Nonlinear Programming*, McGraw-Hill, New York, 1996.
- [21] M. NIKOLOVA, *Local strong homogeneity of a regularized estimator*, SIAM J. Appl. Math., 61 (2000), pp. 633–658.
- [22] M. ORTEGA AND W. C. RHEINOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
- [23] P. PERONA AND J. MALIK, *Scale-space and edge detection using anisotropic diffusion*, IEEE Trans. Pattern Anal. Machine Intell., 12 (1990), pp. 629–639.
- [24] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [25] S. SAQUIB, C. BOUMAN, AND K. SAUER, *ML parameter estimation for Markov random fields, with applications to Bayesian tomography*, IEEE Trans. Image Process., 7 (1998), pp. 1029–1044.
- [26] G. SEBASTIANI AND F. GODTLIEBSEN, *On the use of Gibbs priors for Bayesian image restoration*, Signal Process., 56 (1997), pp. 111–118.
- [27] G. W. STEWART, *Introduction to Matrix Computation*, Academic Press, New York, 1973.
- [28] S. TEBOUL, L. BLANCFERAUD, G. AUBERT, AND M. BARLAUD, *Variational approach for edge preserving regularization using coupled pde's*, IEEE Trans. Image Process., 7 (1998), pp. 387–397.
- [29] A. M. THOMSON, J. C. BROWN, J. W. KAY, AND D. M. TITTERINGTON, *A study of methods of choosing the smoothing parameter in image restoration by regularization*, IEEE Trans. Pattern Anal. Machine Intell., 13 (1991), pp. 326–339.
- [30] L. N. VICENTE AND P. H. CALAMAI, *Bilevel and multilevel programming: A bibliography review*, J. Global Optim., 5 (1994), pp. 291–306.
- [31] J. WEICKERT, *Anisotropic Diffusion in Image Processing*, B. G. Teubner, Stuttgart, Germany, 1998.

A COMBINED SMOOTHING AND REGULARIZATION METHOD FOR MONOTONE SECOND-ORDER CONE COMPLEMENTARITY PROBLEMS*

SHUNSUKE HAYASHI[†], NOBUO YAMASHITA[†], AND MASAO FUKUSHIMA[†]

Abstract. The second-order cone complementarity problem (SOCCP) is a wide class of problems containing the nonlinear complementarity problem (NCP) and the second-order cone programming problem (SOCP). Recently, Fukushima, Luo, and Tseng [*SIAM J. Optim.*, 12 (2001), pp. 436–460] extended some merit functions and their smoothing functions for NCP to SOCCP. Moreover, they derived computable formulas for the Jacobians of the smoothing functions and gave conditions for the Jacobians to be invertible. In this paper, we propose a globally and quadratically convergent algorithm, which is based on smoothing and regularization methods, for solving monotone SOCCP. In particular, we study strong semismoothness and Jacobian consistency, which play an important role in establishing quadratic convergence of the algorithm. Furthermore, we examine the effectiveness of the algorithm by means of numerical experiments.

Key words. second-order cone, complementarity problem, smoothing method, regularization method

AMS subject classifications. 90C33, 65K05

DOI. 10.1137/S1052623403421516

1. Introduction. The *second-order cone complementarity problem* (SOCCP) [11] is to find $(x, y, \zeta) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^l$ such that

$$(1.1) \quad x \in \mathcal{K}, \quad y \in \mathcal{K}, \quad x^T y = 0, \quad F(x, y, \zeta) = 0,$$

where $F : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^l \rightarrow \mathbb{R}^n \times \mathbb{R}^n$ is a continuously differentiable mapping and $\mathcal{K} \subset \mathbb{R}^n$ is the Cartesian product of second-order cones, that is, $\mathcal{K} = \mathcal{K}^{n_1} \times \mathcal{K}^{n_2} \times \cdots \times \mathcal{K}^{n_m}$ with $n = n_1 + \cdots + n_m$ and the n_i -dimensional second-order cone $\mathcal{K}^{n_i} \subset \mathbb{R}^{n_i}$ defined by

$$\mathcal{K}^{n_i} := \{(z_1, z_2^T)^T \in \mathbb{R} \times \mathbb{R}^{n_i-1} \mid \|z_2\|_2 \leq z_1\}.$$

The SOCCP contains a wide class of problems such as the nonlinear complementarity problem (NCP) and the second-order cone programming problem (SOCP) [13]. For example, SOCCP with $n_1 = n_2 = \cdots = n_m = 1$ and $F(x, y, \zeta) = f(x) - y$ is NCP, and the KKT conditions for SOCP (possibly including nonlinear functions) reduce to SOCCP.

In this paper, we focus on the special SOCCP: Find $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$ such that

$$(1.2) \quad x \in \mathcal{K}, \quad y \in \mathcal{K}, \quad x^T y = 0, \quad y = f(x),$$

where f is a continuously differentiable mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$. This SOCCP may seem rather restrictive. However, the KKT conditions for any SOCP with continuously differentiable functions can be written in the form of SOCCP (1.2). For more detail, see section 6.

*Received by the editors January 20, 2003; accepted for publication (in revised form) April 28, 2004; published electronically February 19, 2005. This research was supported in part by a Grant-in-Aid for Scientific Research (B) from the Japan Society for the Promotion of Science.

<http://www.siam.org/journals/siopt/15-2/42151.html>

[†]Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan (hayashi@amp.i.kyoto-u.ac.jp, nobuo@amp.i.kyoto-u.ac.jp, fuku@amp.i.kyoto-u.ac.jp).

Recently, a number of methods for solving SOCCP have been proposed. For SOCP including affine functions, primal-dual interior-point methods [13, 15, 25] are shown to be effective. For SOCCP, Fukushima, Luo, and Tseng [11]; Chen, Sun, and Sun [5]; and Chen, Chen, and Tseng [2] studied smoothing and nonsmooth approaches. In particular, Fukushima, Luo, and Tseng [11] showed that the min function and the Fischer–Burmeister function for the NCP can be extended to the SOCCP by means of Jordan algebra. Furthermore, they constructed smoothing functions for those functions and analyzed the properties of their Jacobians.

In this paper, we introduce not only smoothing methods but also regularization methods for SOCCPs. Smoothing methods have been adopted effectively to deal with nondifferentiable reformulations of complementarity problems [1, 3, 4, 12, 17, 19, 21, 26]. On the other hand, regularization methods have provided a fundamental tool for dealing with ill-posed problems [7, 8, 17]. By combining these methods, we develop a hybrid algorithm for solving monotone SOCCPs. Moreover, we show conditions for the algorithm to be globally and quadratically convergent. Particularly, we highlight the two properties called *strong semismoothness* and *Jacobian consistency*, which play a crucial role in establishing quadratic convergence of the algorithm.

This paper is organized as follows. In section 2, we review the spectral factorization associated with a second-order cone, which plays a key role in analyzing the properties of merit functions for the SOCCP. Moreover, we construct a merit function by means of the natural residual for the SOCCP. In section 3, we introduce smoothing and regularization methods, which smooth the natural residual and weaken the condition for the merit function to be level-bounded. In section 4, we propose a prototype algorithm for solving the SOCCP and show that it has global convergence when applied to monotone SOCCPs. Moreover, by incorporating Newton’s method, we construct a concrete algorithm and establish quadratic convergence. In section 5, we present some numerical results for the latter algorithm. In section 6, we conclude the paper with some remarks.

Throughout this paper, we use the following notation: $\|\cdot\|$ denotes the 2-norm defined by $\|z\| := \sqrt{z^T z}$ for a vector z and $\|M\| := \max_{z \in \mathbb{R}^n \setminus \{0\}} \|Mz\|/\|z\|$ for a matrix $M \in \mathbb{R}^{m \times n}$; $B(z, \varepsilon)$ denotes the open sphere with center z and radius ε ; for a given set S , $\text{cl } S$ denotes the closure of S , and $\text{co } S$ denotes the convex hull of S . We often write $z = (z_1, z_2)$ for $(z_1, z_2^T)^T$.

2. Preliminaries.

2.1. Spectral factorization. We briefly review some properties of the spectral factorization with respect to a second-order cone, which will be used in the subsequent analysis. Spectral factorization is one of the basic concepts in Jordan algebra. For more detail, see [10, 11].

For any vector $z = (z_1, z_2) \in \mathbb{R} \times \mathbb{R}^{n-1}$ ($n \geq 2$), its spectral factorization with respect to the second-order cone $\mathcal{K}^n$ is defined as

$$z = \lambda_1 u^1 + \lambda_2 u^2,$$

where λ_1 and λ_2 are the *spectral values* given by

$$(2.1) \quad \lambda_i = z_1 + (-1)^i \|z_2\|, \quad i = 1, 2,$$

and u^1 and u^2 are the *spectral vectors* given by

$$(2.2) \quad u^i = \begin{cases} \frac{1}{2} \begin{pmatrix} 1, (-1)^i \frac{z_2}{\|z_2\|} \end{pmatrix} & (z_2 \neq 0), \\ \frac{1}{2} (1, (-1)^i w) & (z_2 = 0), \end{cases} \quad i = 1, 2,$$

with $w \in \Re^{n-1}$ such that $\|w\| = 1$. The spectral values and vectors have the following properties. For any $z \in \Re^n$, the inequality $\lambda_1 \leq \lambda_2$ holds, and $\lambda_1 \geq 0 \Leftrightarrow z \in \mathcal{K}^n$. Moreover, for any $z \in \Re^n$, we have $\|u^i\| = 1/\sqrt{2}$ for $i = 1, 2$, and $(u^1)^T u^2 = 0$.

For any $z \in \Re^n$, let $P_{\mathcal{K}^n}(z)$ denote the projection of z onto the second-order cone $\mathcal{K}^n$, that is,

$$(2.3) \quad P_{\mathcal{K}^n}(z) := \operatorname{argmin}_{z' \in \mathcal{K}^n} \|z' - z\|.$$

In particular, when $n = 1$, $P_{\mathcal{K}^1}(\alpha) = \max\{0, \alpha\}$ for any $\alpha \in \Re$. For $n \geq 2$, the projection function $P_{\mathcal{K}^n}$ can be calculated by using the spectral values and vectors of z (see [11]); that is,

$$(2.4) \quad P_{\mathcal{K}^n}(z) = \max\{0, \lambda_1\} u^1 + \max\{0, \lambda_2\} u^2,$$

where λ_1 and λ_2 are the spectral values of z defined by (2.1), and u^1 and u^2 are the spectral vectors of z defined by (2.2).

2.2. Merit function. In this paper, we consider the unconstrained optimization reformulation of SOCCP (1.2):

$$\text{Minimize } \Psi(x, y),$$

where Ψ is a real-valued function on $\Re^n \times \Re^n$. The objective function Ψ is called a merit function for SOCCP (1.2). In order to construct a merit function for SOCCP (1.2), it is convenient to introduce a function $\Phi : \Re^n \times \Re^n \rightarrow \Re^n$ satisfying

$$(2.5) \quad \Phi(x, y) = 0 \iff x \in \mathcal{K}, \quad y \in \mathcal{K}, \quad x^T y = 0.$$

By using such a function, we define $H : \Re^n \times \Re^n \rightarrow \Re^{2n}$ by

$$H(x, y) := \begin{pmatrix} \Phi(x, y) \\ f(x) - y \end{pmatrix}.$$

It is obvious that SOCCP (1.2) is equivalent to the equation $H(x, y) = 0$. Moreover, we define function $\Psi : \Re^n \times \Re^n \rightarrow \Re$ by

$$(2.6) \quad \Psi(x, y) := \frac{1}{2} \|H(x, y)\|^2 = \frac{1}{2} \|\Phi(x, y)\|^2 + \frac{1}{2} \|f(x) - y\|^2.$$

Then, it is easy to see that $\Psi(x, y) \geq 0$ for any $(x, y) \in \Re^n \times \Re^n$, and that $\Psi(x, y) = 0$ if and only if (x, y) is a solution of SOCCP (1.2). Therefore, the function Ψ defined by (2.6) can serve as a merit function for SOCCP (1.2).

Note that the complementarity condition on $\mathcal{K} = \mathcal{K}^{n_1} \times \cdots \times \mathcal{K}^{n_m}$ can be decomposed into complementarity conditions on each $\mathcal{K}^{n_i}$; that is,

$$(2.7) \quad \begin{aligned} & x \in \mathcal{K}, \quad y \in \mathcal{K}, \quad x^T y = 0 \\ \iff & x^i \in \mathcal{K}^{n_i}, \quad y^i \in \mathcal{K}^{n_i}, \quad (x^i)^T y^i = 0 \quad (i = 1, \dots, m), \end{aligned}$$

where $x = (x^1, \dots, x^m) \in \mathbb{R}^{n_1} \times \dots \times \mathbb{R}^{n_m}$ and $y = (y^1, \dots, y^m) \in \mathbb{R}^{n_1} \times \dots \times \mathbb{R}^{n_m}$. This fact naturally leads us to construct a function Φ satisfying (2.5) as

$$\Phi(x, y) := \begin{pmatrix} \varphi^1(x^1, y^1) \\ \vdots \\ \varphi^m(x^m, y^m) \end{pmatrix},$$

where $\varphi^i : \mathbb{R}^{n_i} \times \mathbb{R}^{n_i} \rightarrow \mathbb{R}^{n_i}$ is a function satisfying

$$(2.8) \quad \varphi^i(x^i, y^i) = 0 \iff x^i \in \mathcal{K}^{n_i}, \quad y^i \in \mathcal{K}^{n_i}, \quad (x^i)^T y^i = 0$$

for each $i = 1, \dots, m$. Fukushima, Luo, and Tseng [11] showed that (2.8) holds for the *natural residual* function $\varphi_{\text{NR}}^i : \mathbb{R}^{n_i} \times \mathbb{R}^{n_i} \rightarrow \mathbb{R}^{n_i}$ defined by

$$(2.9) \quad \varphi_{\text{NR}}^i(x^i, y^i) := x^i - P_{\mathcal{K}^{n_i}}(x^i - y^i).$$

Using this function, we define function $\Phi_{\text{NR}} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ by

$$\Phi_{\text{NR}}(x, y) := \begin{pmatrix} \varphi_{\text{NR}}^1(x^1, y^1) \\ \vdots \\ \varphi_{\text{NR}}^m(x^m, y^m) \end{pmatrix},$$

and $H_{\text{NR}} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^{2n}$ by

$$H_{\text{NR}}(x, y) := \begin{pmatrix} \Phi_{\text{NR}}(x, y) \\ f(x) - y \end{pmatrix}.$$

Then, we can construct a merit function $\Psi_{\text{NR}} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ for SOCCP (1.2) as

$$\Psi_{\text{NR}}(x, y) := \frac{1}{2} \|H_{\text{NR}}(x, y)\|^2 = \frac{1}{2} \sum_{i=1}^m \|\varphi_{\text{NR}}^i(x^i, y^i)\|^2 + \frac{1}{2} \|f(x) - y\|^2.$$

In what follows, we write φ_{NR} for φ_{NR}^i for simplicity of notation.

3. Smoothing and regularization. In the previous section, we have constructed the merit function Ψ_{NR} from the natural residual φ_{NR} . We can solve SOCCP (1.2) by minimizing Ψ_{NR} by an appropriate descent algorithm. However, the function Ψ_{NR} is not differentiable, and hence, methods that use the gradient of the function, such as the steepest descent method and Newton's method, are not applicable. In order to get rid of this difficulty, we first introduce a smoothing method that solves a sequence of differentiable approximations to the original nondifferentiable problem. To ensure global convergence of a descent method, the level-boundedness of the objective function plays an important role. If the function involved in the SOCCP is strongly monotone, then the merit function Ψ_{NR} is level-bounded. However, the assumption of strong monotonicity is quite restrictive from a practical standpoint. To be amenable to a merely monotone problem, we propose to combine a regularization method with a smoothing method.

In the remainder of the paper, we assume $\mathcal{K} = \mathcal{K}^n$. Then we can rewrite SOCCP (1.2) as follows: Find $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$ such that

$$(3.1) \quad x \in \mathcal{K}^n, \quad y \in \mathcal{K}^n, \quad x^T y = 0, \quad y = f(x).$$

The assumption $\mathcal{K} = \mathcal{K}^n$ is only for simplicity of presentation. In view of (2.7), the results obtained in the following can be extended to the general $\mathcal{K}$ in a straightforward manner (see section 6). Under this assumption, $\Phi_{\text{NR}}(x, y)$ is reduced to $\varphi_{\text{NR}}(x, y)$, and hence, the vector-valued function H_{NR} and the merit function Ψ_{NR} are respectively rewritten as

$$(3.2) \quad \begin{aligned} H_{\text{NR}}(x, y) &= \begin{pmatrix} \varphi_{\text{NR}}(x, y) \\ f(x) - y \end{pmatrix}, \\ \Psi_{\text{NR}}(x, y) &= \frac{1}{2} \|H_{\text{NR}}(x, y)\|^2 = \frac{1}{2} \|\varphi_{\text{NR}}(x, y)\|^2 + \frac{1}{2} \|f(x) - y\|^2. \end{aligned}$$

3.1. Smoothing functions. In this subsection, we introduce a class of smoothing functions of the merit function Ψ_{NR} . For a nondifferentiable function $h : \mathbb{R}^n \rightarrow \mathbb{R}^m$, we consider a function $h_\mu : \mathbb{R}^n \rightarrow \mathbb{R}^m$ with a parameter $\mu > 0$ that has the following properties:

- (a) h_μ is differentiable for any $\mu > 0$,
- (b) $\lim_{\mu \downarrow 0} h_\mu(x) = h(x)$ for any $x \in \mathbb{R}^n$.

Such a function h_μ is called a *smoothing function* of h . Instead of handling the nonsmooth equation $h(x) = 0$ directly, the smoothing method solves a family of smoothed subproblems $h_\mu(x) = 0$ for $\mu > 0$ and obtains a solution of the original problem by letting $\mu \downarrow 0$. Fukushima, Luo, and Tseng [11] extended Chen and Mangasarian's class [1] of smoothing functions for NCP to SOCCP, which may be regarded as a smoothing function of the natural residual φ_{NR} .

First we define a smoothing function of the projection function $P_{\mathcal{K}^n}$ defined by (2.3). To this end, we consider a continuously differentiable convex function $\hat{g} : \mathbb{R} \rightarrow \mathbb{R}$ such that

$$(3.3) \quad \lim_{\alpha \rightarrow -\infty} \hat{g}(\alpha) = 0, \quad \lim_{\alpha \rightarrow \infty} (\hat{g}(\alpha) - \alpha) = 0, \quad 0 < \hat{g}'(\alpha) < 1.$$

For example, $\hat{g}_1(\alpha) = (\sqrt{\alpha^2 + 4} + \alpha)/2$ and $\hat{g}_2(\alpha) = \ln(e^\alpha + 1)$ satisfy (3.3). By using $\hat{g}$, we define function $P_\mu : \mathbb{R}^n \rightarrow \mathbb{R}^n$ by

$$(3.4) \quad P_\mu(z) := \mu \hat{g}\left(\frac{\lambda_1}{\mu}\right) u^1 + \mu \hat{g}\left(\frac{\lambda_2}{\mu}\right) u^2,$$

where λ_1 and λ_2 are the spectral values of z given by (2.1), and u^1 and u^2 are the spectral vectors of z given by (2.2). Fukushima, Luo, and Tseng [11] showed that the function P_μ defined by (3.4) is a smoothing function of $P_{\mathcal{K}^n}$, by using the facts that $\lim_{\mu \downarrow 0} \mu \hat{g}(\lambda/\mu) = \max\{0, \lambda\}$ and that $\gamma_\mu(\lambda) := \mu \hat{g}(\lambda/\mu)$ is differentiable for any $\mu > 0$. Hence, from the definition (2.9) of φ_{NR} , the function $\varphi_\mu : \mathbb{R}^n \rightarrow \mathbb{R}^n$ defined by

$$\varphi_\mu(x, y) := x - P_\mu(x - y)$$

becomes a smoothing function of φ_{NR} . In particular, by [11, Proposition 5.1], there exists a positive constant ν such that

$$(3.5) \quad \|\varphi_\mu(x, y) - \varphi_{\text{NR}}(x, y)\| \leq \nu\mu$$

for any $\mu > 0$ and $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$. Likewise, function $\Psi_\mu : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ given by $\Psi_\mu(x, y) := (1/2) \|\varphi_\mu(x, y)\|^2 + (1/2) \|f(x) - y\|^2$ serves as a smoothing function of the merit function Ψ_{NR} . In the following, we define $P_0(z) := P_{\mathcal{K}^n}(z)$, $\varphi_0(x, y) := \varphi_{\text{NR}}(x, y)$, and $\Psi_0(x, y) := \Psi_{\text{NR}}(x, y)$.

3.2. Regularization method. Since Ψ_μ is differentiable, we may apply an appropriate gradient-based descent method to obtain a minimum (x_μ, y_μ) of the function Ψ_μ for $\mu > 0$. In order that a sequence generated by such a descent method have an accumulation point, it is desirable that Ψ_μ be level-bounded, that is, that the level sets $\mathcal{L}_\alpha := \{(x, y) \mid \Psi_\mu(x, y) \leq \alpha\}$ be bounded for all $\alpha \in \mathbb{R}$. Actually, by using (3.5) and the results for a natural residual of variational inequality problems [16, 27], we can show that Ψ_μ is level-bounded for any $\mu \geq 0$ if f is strongly monotone; that is, for some $\varepsilon > 0$,

$$(3.6) \quad (x - y)^T (f(x) - f(y)) \geq \varepsilon \|x - y\|^2 \quad \forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n.$$

However, strong monotonicity is quite a severe condition. As a remedy for this inconvenience, we employ a regularization method.

Let the function $f_\varepsilon : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be defined by $f_\varepsilon(x) := f(x) + \varepsilon x$ with a positive parameter ε . The regularization method solves a sequence of SOCCPs involving f_ε with $\varepsilon > 0$, so that a solution of the original SOCCP is obtained by taking the limit $\varepsilon \downarrow 0$. Define functions $H_{\mu,\varepsilon} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^{2n}$ and $\Psi_{\mu,\varepsilon} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ by

$$(3.7) \quad H_{\mu,\varepsilon}(x, y) := \begin{pmatrix} \varphi_\mu(x, y) \\ f_\varepsilon(x) - y \end{pmatrix},$$

$$(3.8) \quad \Psi_{\mu,\varepsilon}(x, y) := \frac{1}{2} \|H_{\mu,\varepsilon}(x, y)\|^2 = \frac{1}{2} \|\varphi_\mu(x, y)\|^2 + \frac{1}{2} \|f_\varepsilon(x) - y\|^2.$$

If f is monotone, that is, (3.6) holds for $\varepsilon = 0$, then f_ε is strongly monotone for any $\varepsilon > 0$, and hence, the function $\Psi_{\mu,\varepsilon}$ is level-bounded for any $\mu \geq 0$ and $\varepsilon > 0$.

Finally, we give explicit expressions of functions $\nabla H_{\mu,\varepsilon}$ and ∇P_μ , which will be useful for the subsequent analysis.

Remark 3.1. From the definition of $H_{\mu,\varepsilon}$, φ_μ , and f_ε , we can write $\nabla H_{\mu,\varepsilon}(x, y)$ as

$$(3.9) \quad \nabla H_{\mu,\varepsilon}(x, y) = \begin{pmatrix} I - \nabla P_\mu(x - y) & \nabla f(x) + \varepsilon I \\ \nabla P_\mu(x - y) & -I \end{pmatrix}.$$

Let $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be defined by $g(z) := \hat{g}(\lambda_1)u^1 + \hat{g}(\lambda_2)u^2$, where λ_1 and λ_2 are the spectral values of z and u^1 and u^2 are the spectral vectors of z . Then we have $\nabla P_\mu(z) = \nabla g(z/\mu)$ since $P_\mu(z) = \mu g(z/\mu)$ from (3.4). Therefore, by [11, Proposition 5.2], $\nabla P_\mu(z)$ is written as

$$(3.10) \quad \nabla P_\mu(z) = \begin{cases} \hat{g}'(z_1/\mu)I & \text{if } z_2 = 0, \\ \begin{pmatrix} b_\mu & \frac{c_\mu z_2^T}{\|z_2\|} \\ \frac{c_\mu z_2}{\|z_2\|} & a_\mu I + (b_\mu - a_\mu) \frac{z_2 z_2^T}{\|z_2\|^2} \end{pmatrix} & \text{if } z_2 \neq 0, \end{cases}$$

where

$$(3.11) \quad \begin{aligned} a_\mu &= \frac{\hat{g}(\lambda_2/\mu) - \hat{g}(\lambda_1/\mu)}{\lambda_2/\mu - \lambda_1/\mu}, & b_\mu &= \frac{1}{2} \left(\hat{g}'\left(\frac{\lambda_2}{\mu}\right) + \hat{g}'\left(\frac{\lambda_1}{\mu}\right) \right), \\ c_\mu &= \frac{1}{2} \left(\hat{g}'\left(\frac{\lambda_2}{\mu}\right) - \hat{g}'\left(\frac{\lambda_1}{\mu}\right) \right). \end{aligned}$$

4. Algorithm. In this section, we present a globally and quadratically convergent algorithm for solving SOCCP (3.1). We first give a globally convergent prototype algorithm that utilizes smoothing and regularization techniques. Next, we study two properties called strong semismoothness and Jacobian consistency, which play an important role in establishing quadratic convergence of the algorithm. Finally, we construct a quadratically convergent algorithm by incorporating Newton's method into the prototype algorithm.

4.1. Prototype algorithm. As we have stated in the previous section, for any $\mu > 0$ and $\varepsilon > 0$, the function $\Psi_{\mu,\varepsilon}$ defined by (3.8) is differentiable and level-bounded, provided that f is monotone. Therefore, by applying an appropriate descent method, we may obtain a minimum $(x_{\mu,\varepsilon}, y_{\mu,\varepsilon})$ of the function $\Psi_{\mu,\varepsilon}$. Moreover, letting (μ, ε) converge to $(0, 0)$, we can expect that $(x_{\mu,\varepsilon}, y_{\mu,\varepsilon})$ converges to a solution of the original SOCCP. However, in practice, it is usually impossible to compute an exact minimum of $\Psi_{\mu,\varepsilon}$. Thus, we consider the following algorithm in which the function $\Psi_{\mu,\varepsilon}$ is minimized only approximately at each iteration.

ALGORITHM 1.

Step 0. Choose $(x^{(0)}, y^{(0)}) \in \mathbb{R}^n \times \mathbb{R}^n$, $\mu_0 \in (0, \infty)$, $\varepsilon_0 \in (0, \infty)$, and $\alpha_0 \in (0, \infty)$. Set $k := 0$.

Step 1. Terminate the iteration if $\Psi_{\text{NR}}(x^{(k)}, y^{(k)}) = 0$.

Step 2. Find a pair $(x^{(k+1)}, y^{(k+1)}) \in \mathbb{R}^n \times \mathbb{R}^n$ such that

$$(4.1) \quad \Psi_{\mu_k, \varepsilon_k}(x^{(k+1)}, y^{(k+1)}) \leq \alpha_k.$$

Step 3. Update parameters $\mu_{k+1} \in (0, \mu_k)$, $\varepsilon_{k+1} \in (0, \varepsilon_k)$, and $\alpha_{k+1} \in (0, \alpha_k)$ so that they converge to 0 eventually. Set $k := k + 1$. Go back to Step 1.

To obtain $(x^{(k+1)}, y^{(k+1)})$ in Step 2, we may use any suitable unconstrained minimization technique. These issues will be discussed in detail in subsection 4.4.

In order for Algorithm 1 to be well defined, there must exist a pair $(x^{(k+1)}, y^{(k+1)})$ satisfying (4.1) for any $\alpha_k > 0$. For this purpose, it suffices to show that any stationary point of $\Psi_{\mu,\varepsilon}$ is a global minimum of $\Psi_{\mu,\varepsilon}$.

PROPOSITION 4.1. *If $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is monotone, then, for any $\mu > 0$ and $\varepsilon \geq 0$, every stationary point $(\bar{x}, \bar{y})$ of the function $\Psi_{\mu,\varepsilon}$ satisfies $\Psi_{\mu,\varepsilon}(\bar{x}, \bar{y}) = 0$.*

Proof. Note that $\nabla \Psi_{\mu,\varepsilon}(\bar{x}, \bar{y}) = \nabla H_{\mu,\varepsilon}(\bar{x}, \bar{y}) H_{\mu,\varepsilon}(\bar{x}, \bar{y}) = 0$. By Proposition 6.1 of [11], $\nabla H_{\mu,\varepsilon}(\bar{x}, \bar{y})$ is nonsingular. Hence, we have $H_{\mu,\varepsilon}(\bar{x}, \bar{y}) = 0$; that is, $\Psi_{\mu,\varepsilon}(\bar{x}, \bar{y}) = (1/2) \|H_{\mu,\varepsilon}(\bar{x}, \bar{y})\|^2 = 0$. $\square$

We proceed to showing the global convergence property of Algorithm 1. To this end, we introduce the weak univalence property. We say that function $H : \mathcal{D} \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^n$ is weakly univalent if it is continuous and there exists a sequence of continuous and injective functions $\{H_k\}$ converging to H uniformly on a bounded subset of $\mathcal{D}$. In particular, if f is monotone, then H_{NR} is weakly univalent for the following reasons. For any $\mu > 0$ and $\varepsilon \geq 0$, $\nabla H_{\mu,\varepsilon}(x, y)$ is nonsingular for any $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$ [11, Proposition 6.1], and hence $H_{\mu,\varepsilon}$ is injective. Moreover, $H_{\mu,\varepsilon}$ converges to H_{NR} uniformly on a bounded set as $(\mu, \varepsilon) \downarrow (0, 0)$ since we have from (3.5) that

$$(4.2) \quad \begin{aligned} \|H_{\mu,\varepsilon}(x, y) - H_{\text{NR}}(x, y)\| &\leq \|\varphi_\mu(x, y) - \varphi_{\text{NR}}(x, y)\| + \|f_\varepsilon(x) - f(x)\| \\ &\leq \nu\mu + \varepsilon\|x\|. \end{aligned}$$

The next lemma indicates a property which weakly univalent functions possess.

LEMMA 4.2 (see [9, Corollary 3.6.5]). *Let $H : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a weakly univalent function such that the inverse image $H^{-1}(0)$ is nonempty and compact. Then, for any $\varepsilon > 0$, there exists $\delta > 0$ such that the following statement holds: For every weakly univalent function $\tilde{H} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that $\|H(z) - \tilde{H}(z)\| \leq \delta$ for any $z \in \text{cl}(H^{-1}(0) + B(0, \varepsilon))$, we have $\emptyset \neq \tilde{H}^{-1}(0) \subseteq H^{-1}(0) + B(0, \varepsilon)$, and $\tilde{H}^{-1}(0)$ is connected.*

By using this lemma, we establish the global convergence of Algorithm 1.

THEOREM 4.3. *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a monotone function. Assume that the solution set $\mathcal{S}$ of SOCCP (3.1) is nonempty and bounded. Let $\{(x^{(k)}, y^{(k)})\}$ be a sequence generated by Algorithm 1. Then, $\{(x^{(k)}, y^{(k)})\}$ is bounded, and every accumulation point is a solution of SOCCP (3.1).*

Proof. From a simple continuity argument, we can easily show that every accumulation point of $\{(x^{(k)}, y^{(k)})\}$ is a solution of SOCCP (3.1). Thus we now show only the boundedness of $\{(x^{(k)}, y^{(k)})\}$. For any $\varepsilon > 0$, let $\Omega := \text{cl}(H_{\text{NR}}^{-1}(0) + B(0, \varepsilon))$, which is nonempty and compact by the given assumption. Then there exists $\delta > 0$ such that Lemma 4.2 holds for $H = H_{\text{NR}}$. Let continuous and injective functions $G_k : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^{2n}$ be defined by $G_k(x, y) := H_{\mu_{k-1}, \varepsilon_{k-1}}(x, y) - H_{\mu_{k-1}, \varepsilon_{k-1}}(x^{(k)}, y^{(k)})$. From (4.1) and (4.2), there exists $\bar{k}$ such that $\|G_k(x, y) - H_{\text{NR}}(x, y)\| \leq \|H_{\mu_{k-1}, \varepsilon_{k-1}}(x, y) - H_{\text{NR}}(x, y)\| + \|H_{\mu_{k-1}, \varepsilon_{k-1}}(x^{(k)}, y^{(k)})\| \leq \nu\mu_{k-1} + \varepsilon_{k-1}\|x\| + \sqrt{2\alpha_{k-1}} \leq \delta$ for any $k \geq \bar{k}$ and $(x, y) \in \Omega$. Moreover, G_k is a weakly univalent function. Hence, by Lemma 4.2 with $\tilde{H} = G_k$ and $H = H_{\text{NR}}$, we have $G_k^{-1}(0) \subseteq \Omega$ for all $k \geq \bar{k}$. This together with $(x^{(k)}, y^{(k)}) \in G_k^{-1}(0)$ implies the boundedness of $\{(x^{(k)}, y^{(k)})\}$. $\square$

4.2. Semismoothness and strong semismoothness. Semismoothness is a generalized concept of smoothness, which was originally introduced by Mifflin [14] for functionals and extended to vector-valued functions by Qi and Sun [20]. Strong semismoothness is also a generalized concept of smoothness, which is stronger than semismoothness. These concepts play an important role in establishing fast local convergence of the algorithm presented later. In this subsection, we first show that the function $P_{\mathcal{K}^n}$ is strongly semismooth at an arbitrary point. Then, by using this result, we show the strong semismoothness of the function H_{NR} defined by (3.2).

Let $H : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be a locally Lipschitzian function. Then H is differentiable almost everywhere by Rademacher's theorem (see [6]). Let $D_H \subseteq \mathbb{R}^n$ be the set of points where H is differentiable. Then, Clarke subdifferential $\partial H(x)$ of H at x is defined by $\partial H(x) := \text{co}\{\lim_{\hat{x} \rightarrow x, \hat{x} \in D_H} \nabla H(\hat{x})\}$ [6, 18]. We note that $\partial H(x) = \{\nabla H(x)\}$ if H is continuously differentiable at x . By using the concept of subdifferential, we give the definitions of semismoothness and strong semismoothness.

DEFINITION 4.4. *A directionally differentiable and locally Lipschitzian function $H : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is said to be semismooth at z if $V^T d - H'(z; d) = o(\|d\|)$ for any $d \in \mathbb{R}^n \setminus \{0\}$ sufficiently small and $V \in \partial H(z + d)$, where $H'(z; d) := \lim_{h \downarrow 0} (H(z + hd) - H(z))/h$ is the directional derivative of H at z along the direction d . In particular, if $o(\|d\|)$ can be replaced by $O(\|d\|^2)$, then function H is said to be strongly semismooth.*

It is known that any piecewise smooth function¹ is semismooth. Moreover, any piecewise smooth function is strongly semismooth if all the derivatives of smooth pieces are locally Lipschitzian. Next we show that the projection function $P_{\mathcal{K}^n}$ is strongly semismooth everywhere.

¹Function $H : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is said to be piecewise smooth at z if there exist $\varepsilon > 0$ and a finite number of continuously differentiable functions $H_i : B(z, \varepsilon) \rightarrow \mathbb{R}^m$ ($i = 1, \dots, N$) such that for any $z' \in B(z, \varepsilon)$ there exists an index i such that $H(z') = H_i(z')$.

PROPOSITION 4.5. *The projection function $P_{\mathcal{K}^n}$ is strongly semismooth at any $z \in \mathbb{R}^n$.*

Proof. Let λ_1 and λ_2 be the spectral values of z . Since we have from (2.4)

$$P_{\mathcal{K}^n}(z) = \begin{cases} z & (\lambda_1 \geq 0, \lambda_2 \geq 0), \\ \frac{1}{2}(z_1 + \|z_2\|) \begin{pmatrix} 1, \frac{z_2}{\|z_2\|} \end{pmatrix} & (\lambda_1 < 0, \lambda_2 > 0), \\ 0 & (\lambda_1 \leq 0, \lambda_2 \leq 0), \end{cases}$$

$P_{\mathcal{K}^n}$ is continuously differentiable at any z such that $\lambda_1 \neq 0$ and $\lambda_2 \neq 0$, and its derivative is locally Lipschitzian. Moreover, $P_{\mathcal{K}^n}$ is piecewise smooth at any z such that $\lambda_2 > \lambda_1 = 0$ or $\lambda_1 < \lambda_2 = 0$. Hence, we have only to show the strong semismoothness at the origin. Note that $P_{\mathcal{K}^n}(tz) = tP_{\mathcal{K}^n}(z)$ for any $t \geq 0$. Since any locally Lipschitzian positively homogeneous function is strongly semismooth at the origin [23], $P_{\mathcal{K}^n}$ is strongly semismooth at the origin. $\square$

Recently, Chen, Sun, and Sun [5] also showed the strong semismoothness of $P_{\mathcal{K}^n}$, and Chen, Chen, and Tseng [2] gave more general results. We note, however, that their approaches are quite different from ours. Now we are ready to prove the strong semismoothness of H_{NR} .

THEOREM 4.6. *The function H_{NR} defined by (3.2) is semismooth at every point $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$. Moreover, if $\nabla f : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times n}$ is locally Lipschitzian, then H_{NR} is strongly semismooth at every point $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$.*

Proof. Let $(\xi, \eta) \in \mathbb{R}^n \times \mathbb{R}^n$ be a pair of vectors sufficiently small, and U be an arbitrary element of $\partial H_{\text{NR}}(x + \xi, y + \eta)$. Note that $P_{\mathcal{K}^n}$ is strongly semismooth at $x - y$. Then, since

$$\partial H_{\text{NR}}(x + \xi, y + \eta) = \left\{ \begin{pmatrix} I - V & \nabla f(x + \xi) \\ V & -I \end{pmatrix} \middle| V \in \partial P_{\mathcal{K}^n}((x - y) + (\xi - \eta)) \right\}$$

and

$$H'_{\text{NR}}((x, y); (\xi, \eta)) = \begin{pmatrix} \xi - P'_{\mathcal{K}^n}(x - y; \xi - \eta) \\ \nabla f(x)^T \xi - \eta \end{pmatrix},$$

we have

$$\begin{aligned} U^T \begin{pmatrix} \xi \\ \eta \end{pmatrix} - H'_{\text{NR}}((x, y); (\xi, \eta)) &= \begin{pmatrix} I - V & \nabla f(x) \\ V & -I \end{pmatrix}^T \begin{pmatrix} \xi \\ \eta \end{pmatrix} - \begin{pmatrix} \xi - P'_{\mathcal{K}^n}(x - y; \xi - \eta) \\ \nabla f(x)^T \xi - \eta \end{pmatrix} \\ &= \begin{pmatrix} P'_{\mathcal{K}^n}(x - y; \xi - \eta) - V^T(\xi - \eta) \\ (\nabla f(x + \xi) - \nabla f(x))^T \xi \end{pmatrix} \\ (4.3) \quad &= \begin{pmatrix} O(\|\xi - \eta\|^2) \\ o(\|\xi\|) \end{pmatrix} \\ &= o(\|(\xi, \eta)\|), \end{aligned}$$

where the last equality follows from $\|\xi - \eta\|^2 \leq 2\|\xi\|^2 + 2\|\eta\|^2 = 2\|(\xi, \eta)\|^2$ and $\|\xi\| \leq \|(\xi, \eta)\|$. Hence, H_{NR} is semismooth at (x, y) . Furthermore, if ∇f is locally Lipschitzian, then the strong semismoothness of H_{NR} at (x, y) readily follows since $o(\|\xi\|)$ in (4.3) can be replaced by $O(\|\xi\|^2)$. $\square$

4.3. Jacobian consistency of the smoothing function. Jacobian consistency, which was first introduced by Chen, Qi, and Sun [4], is a concept relating the generalized Jacobian of a nonsmooth function with the Jacobian of a smoothing function. Like strong semismoothness, Jacobian consistency plays an important role in establishing rapid convergence of smoothing methods. In this subsection, we show that the function $H_{\mu,\varepsilon}$ defined by (3.7) enjoys the Jacobian consistency and give some results that will be useful in constructing a rapidly convergent algorithm.

DEFINITION 4.7. Let $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a continuous function. Let $F_{\mu,\varepsilon} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a function such that $F_{\mu,\varepsilon}$ is continuously differentiable for any $\mu > 0$ and $\varepsilon \geq 0$, and that $\lim_{(\mu,\varepsilon) \downarrow (0,0)} F_{\mu,\varepsilon}(z) = F(z)$ for any $z \in \mathbb{R}^n$. Then we say that $F_{\mu,\varepsilon}$ satisfies the Jacobian consistency if

$$\lim_{(\mu,\varepsilon) \downarrow (0,0)} \text{dist}(\nabla F_{\mu,\varepsilon}(z), \partial F(z)) = 0$$

holds for any $z \in \mathbb{R}^n$. Here $\text{dist}(X, S)$ denotes $\min\{\|X - Y\| \mid Y \in S\}$.

We note that Chen, Qi, and Sun [4] define the Jacobian consistency in terms of $\partial_C F(z) := \partial F_1(z) \times \cdots \times \partial F_n(z) \supseteq \partial F(z)$ instead of $\partial F(z)$. Moreover, their definition contains only a smoothing parameter μ .

In order to show the Jacobian consistency of $H_{\mu,\varepsilon}$, we need the following proposition. Since it can be proved from (2.4) and (3.10) in a straightforward manner, we omit the proof.

PROPOSITION 4.8. The Clarke subdifferential of the projection function P_{κ^n} and the matrix $J_P^o(z) := \lim_{\mu \downarrow 0} \nabla P_\mu(z)$ are given as follows:

$$(4.4) \quad \partial P_{\kappa^n}(z) = \begin{cases} \{I\} & (\lambda_1 > 0, \lambda_2 > 0), \\ \{(\lambda_2/(\lambda_1 + \lambda_2))I + Z\} & (\lambda_1 < 0, \lambda_2 > 0), \\ \{O\} & (\lambda_1 < 0, \lambda_2 < 0), \\ \text{co}\{I, I + Z\} & (\lambda_1 = 0, \lambda_2 > 0), \\ \text{co}\{O, Z\} & (\lambda_1 < 0, \lambda_2 = 0), \\ \text{co}(\{O\} \cup \{I\} \cup \mathcal{S}) & (\lambda_1 = 0, \lambda_2 = 0), \end{cases}$$

$$(4.5) \quad J_P^o(z) = \begin{cases} I & (\lambda_1 > 0, \lambda_2 > 0), \\ (\lambda_2/(\lambda_1 + \lambda_2))I + Z & (\lambda_1 < 0, \lambda_2 > 0), \\ O & (\lambda_1 < 0, \lambda_2 < 0), \\ I + (1 - \hat{g}'(0))Z & (\lambda_1 = 0, \lambda_2 > 0), \\ \hat{g}'(0)Z & (\lambda_1 < 0, \lambda_2 = 0), \\ \hat{g}'(0)I & (\lambda_1 = 0, \lambda_2 = 0), \end{cases}$$

where $(r_1, r_2) := (z_1, z_2)/\|z_2\|$,

$$Z = \frac{1}{2} \begin{pmatrix} -r_1 & r_2^T \\ r_2 & -r_1 r_2^T \end{pmatrix},$$

$$\mathcal{S} := \left\{ \frac{1}{2}(1 + \beta)I + \frac{1}{2} \begin{pmatrix} -\beta & w^T \\ w & -\beta w w^T \end{pmatrix} \mid -1 \leq \beta \leq 1, \|w\| = 1 \right\}.$$

By using this proposition, we show the Jacobian consistency of $H_{\mu,\varepsilon}$.

THEOREM 4.9. $H_{\mu,\varepsilon}$ satisfies the Jacobian consistency.

Proof. The formulas (4.4), (4.5), and $0 < \hat{g}'(0) < 1$ from (3.3) yield $J_P^o(z) \in$

$\partial P_{\kappa^n}(z)$ for any $z \in \mathbb{R}^n$. Moreover, we have

$$(4.6) \quad J_H^o(x, y) := \lim_{(\mu, \varepsilon) \downarrow (0, 0)} \nabla H_{\mu, \varepsilon}(x, y) = \begin{pmatrix} I - J_P^o(x - y) & \nabla f(x) \\ J_P^o(x - y) & -I \end{pmatrix}$$

from (3.9), and

$$\partial H_{\text{NR}}(x, y) = \left\{ \begin{pmatrix} I - D & \nabla f(x) \\ D & -I \end{pmatrix} \mid D \in \partial P_{\kappa^n}(x - y) \right\}.$$

We thus have $J_H^o(x, y) \in \partial H_{\text{NR}}(x, y)$ for any $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$, which implies that $H_{\mu, \varepsilon}$ satisfies the Jacobian consistency. $\square$

The Jacobian consistency of $H_{\mu, \varepsilon}$ guarantees that, for any $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$ and $\delta' > 0$, there exist $\mu > 0$ and $\varepsilon > 0$ such that $\text{dist}(\nabla H_{\mu, \varepsilon}(x, y), \partial H_{\text{NR}}(x, y)) < \delta'$. To find such μ and ε for given δ' and (x, y) is an essential issue for constructing a rapidly convergent algorithm. In the remainder of this subsection, we study how we can find such μ and ε .

In what follows, we use the following notation for convenience. Let $\hat{g} : \mathbb{R} \rightarrow \mathbb{R}$ be a continuously differentiable convex function satisfying (3.3). Then we define for any $\mu > 0$

$$(4.7) \quad \gamma_\mu(\alpha) := \mu \hat{g}\left(\frac{\alpha}{\mu}\right).$$

Note that, as $\mu \downarrow 0$, there exist limits of $\gamma_\mu(\alpha)$ and $\gamma'_\mu(\alpha)$ for any fixed α , which we denote by

$$(4.8) \quad \gamma_0(\alpha) := \lim_{\mu \downarrow 0} \gamma_\mu(\alpha) = \max\{0, \alpha\},$$

$$(4.9) \quad \gamma_0^+(\alpha) := \lim_{\mu \downarrow 0} \gamma'_\mu(\alpha) = \begin{cases} 0 & (\alpha < 0), \\ \hat{g}'(0) & (\alpha = 0), \\ 1 & (\alpha > 0). \end{cases}$$

Moreover, we suppose that the function $\hat{g}$ satisfies not only (3.3) but also

$$(4.10) \quad \hat{g}(\alpha) - \alpha = \hat{g}(-\alpha)$$

for any $\alpha \in \mathbb{R}$. For example, $\hat{g}_1(\alpha) = (\sqrt{\alpha^2 + 4} + \alpha)/2$ and $\hat{g}_2(\alpha) = \ln(e^\alpha + 1)$ satisfy (3.3) and (4.10). For such functions, we have the following lemma.

LEMMA 4.10. *Let $\hat{g}$ be a continuously differentiable convex function satisfying (3.3) and (4.10). Let γ_μ , γ_0 , and γ_0^+ be defined by (4.7)–(4.9). Then it holds that (a) $\gamma_\mu(\alpha) - \gamma_0(\alpha) = \gamma_\mu(-\alpha) - \gamma_0(-\alpha)$ for any $\alpha \in \mathbb{R}$, and (b) $|\gamma'_\mu(0) - \gamma_0^+(0)| = 0 < |\gamma'_\mu(\alpha_2) - \gamma_0^+(\alpha_2)| \leq |\gamma'_\mu(\alpha_1) - \gamma_0^+(\alpha_1)|$ for any $(\alpha_1, \alpha_2) \in \mathbb{R} \times \mathbb{R}$ such that $0 < |\alpha_1| \leq |\alpha_2|$.*

Proof. Since (a) can be easily seen from (4.7), (4.8), and (4.10), we show only (b). Let $(\alpha_1, \alpha_2) \in \mathbb{R} \times \mathbb{R}$ be arbitrary scalars such that $0 < |\alpha_1| \leq |\alpha_2|$. From (4.10), we have $\hat{g}'(\alpha) - 1 = -\hat{g}'(-\alpha)$ for any $\alpha \in \mathbb{R}$. This, together with $\gamma'_\mu(\alpha) = \hat{g}'(\alpha/\mu)$ and (4.9), yields $\gamma'_\mu(\alpha) - \gamma_0^+(\alpha) = -\{\gamma'_\mu(-\alpha) - \gamma_0^+(-\alpha)\}$. We therefore have $|\gamma'_\mu(0) - \gamma_0^+(0)| = 0$ and

$$(4.11) \quad |\gamma'_\mu(\alpha_i) - \gamma_0^+(\alpha_i)| = |\gamma'_\mu(|\alpha_i|) - \gamma_0^+(|\alpha_i|)|$$

for each $i = 1, 2$. Now we note that, for any $\beta > 0$, we have $|\gamma'_\mu(\beta) - \gamma_0^+(\beta)| = 1 - \gamma'_\mu(\beta) > 0$, where the inequality follows from $\gamma'_\mu(\beta) = \hat{g}'(\beta/\mu)$ and (3.3). Moreover, $1 - \gamma'_\mu(\beta)$ is monotonically nonincreasing since γ_μ is convex. Hence, we have $0 < |\gamma'_\mu(|\alpha_2|) - \gamma_0^+(|\alpha_2|)| \leq |\gamma'_\mu(|\alpha_1|) - \gamma_0^+(|\alpha_1|)|$. Combining this inequality with (4.11), we have (b). $\square$

We further define function $\tilde{\lambda} : \mathbb{R}^n \rightarrow [0, +\infty)$ by

$$(4.12) \quad \tilde{\lambda}(z) := \begin{cases} \min_{i \in \mathcal{I}(z)} |\lambda_i(z)| & (\mathcal{I}(z) \neq \emptyset), \\ 0 & (\mathcal{I}(z) = \emptyset), \end{cases}$$

where $\lambda_i(z)$ ($i = 1, 2$) are the spectral values of z and $\mathcal{I}(z) \subseteq \{1, 2\}$ is the index set defined by $\mathcal{I}(z) := \{i \mid \lambda_i(z) \neq 0\}$. Then we can easily see from Lemma 4.10(b) that

$$(4.13) \quad |\gamma'_\mu(\lambda_i(z)) - \gamma_0^+(\lambda_i(z))| \leq |\gamma'_\mu(\tilde{\lambda}(z)) - \gamma_0^+(\tilde{\lambda}(z))| \quad (i = 1, 2).$$

From this fact, we can estimate the upper bound of $\text{dist}(\nabla H_{\mu, \varepsilon}(x, y), \partial H_{\text{NR}}(x, y))$.

PROPOSITION 4.11. *Let $\hat{g}$ be a function satisfying (3.3) and (4.10), and let $\tilde{\lambda}$ be the function defined by (4.12). Then, there exists $M > 0$ such that*

$$\text{dist}(\nabla H_{\mu, \varepsilon}(x, y), \partial H_{\text{NR}}(x, y)) \leq M(|\gamma'_\mu(\tilde{\lambda}(x - y)) - \gamma_0^+(\tilde{\lambda}(x - y))| + \varepsilon)$$

for any $\mu > 0$, $\varepsilon \geq 0$, and $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$.

Proof. By (3.9), (4.6), and the Jacobian consistency of $H_{\mu, \varepsilon}$, there exists $M' > 0$ such that $\text{dist}(\nabla H_{\mu, \varepsilon}(x, y), \partial H_{\text{NR}}(x, y)) \leq M'(\|\nabla P_\mu(x - y) - J_P^o(x - y)\| + \varepsilon)$ for any $\mu > 0$, $\varepsilon \geq 0$, and $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$. Thus it suffices to show the existence of $N > 0$ such that

$$\|\nabla P_\mu(z) - J_P^o(z)\| \leq N|\gamma'_\mu(\tilde{\lambda}(z)) - \gamma_0^+(\tilde{\lambda}(z))|$$

for any $\mu > 0$ and $z \in \mathbb{R}^n$. Let λ_1 and λ_2 be the spectral values of z , and a_μ , b_μ , and c_μ be defined by (3.11). Moreover, let

$$(4.14) \quad a_0 := \lim_{\mu \downarrow 0} a_\mu = \frac{\gamma_0(\lambda_2) - \gamma_0(\lambda_1)}{\lambda_2 - \lambda_1},$$

$$(4.15) \quad b_0 := \lim_{\mu \downarrow 0} b_\mu = \frac{1}{2}(\gamma_0^+(\lambda_2) + \gamma_0^+(\lambda_1)),$$

$$(4.16) \quad c_0 := \lim_{\mu \downarrow 0} c_\mu = \frac{1}{2}(\gamma_0^+(\lambda_2) - \gamma_0^+(\lambda_1)).$$

When $z_2 = 0$, that is, $z_1 = \lambda_1 = \lambda_2$, we have $\|\nabla P_\mu(z) - J_P^o(z)\| = \|g'(z_1/\mu)I - \gamma_0^+(z_1)I\| = |\gamma'_\mu(z_1) - \gamma_0^+(z_1)| = |\gamma'_\mu(\tilde{\lambda}(z)) - \gamma_0^+(\tilde{\lambda}(z))|$, where the first equality follows from (3.10), (4.5), and (4.9), and the last equality follows from $\tilde{\lambda}(z) = |z_1|$ and (4.11).

When $z_2 \neq 0$, that is, $\lambda_1 < \lambda_2$ by (3.10) and (4.5), we have for some $N' > 0$

$$(4.17) \quad \begin{aligned} & \|\nabla P_\mu(z) - J_P^o(z)\| \\ &= \left\| (a_\mu - a_0)I + \begin{pmatrix} (b_\mu - b_0) - (a_\mu - a_0) & (c_\mu - c_0)r_2^T \\ (c_\mu - c_0)r_2 & ((b_\mu - b_0) - (a_\mu - a_0))r_2 r_2^T \end{pmatrix} \right\| \\ &\leq N'(|a_\mu - a_0| + |b_\mu - b_0| + |c_\mu - c_0|), \end{aligned}$$

where $r_2 := z_2/\|z_2\|$. First we consider the case where $0 \leq \lambda_1 < \lambda_2$. Note that there exists $\bar{\lambda} \in [\lambda_1, \lambda_2]$ such that $\gamma'_\mu(\bar{\lambda}) = (\gamma_\mu(\lambda_2) - \gamma_\mu(\lambda_1))/(\lambda_2 - \lambda_1) = a_\mu$. Moreover, we have from (4.9), (4.14), and $\bar{\lambda} \in [\lambda_1, \lambda_2]$ that $a_0 = 1 = \gamma_0^+(\bar{\lambda})$. Hence,

we obtain $|a_\mu - a_0| = |\gamma'_\mu(\bar{\lambda}) - \gamma_0^+(\bar{\lambda})|$. Furthermore, by (3.11) and (4.15), we have $|b_\mu - b_0| \leq (1/2)|\gamma'_\mu(\lambda_2) - \gamma_0^+(\lambda_2)| + (1/2)|\gamma'_\mu(\lambda_1) - \gamma_0^+(\lambda_1)|$, and, by (3.11) and (4.16), we have $|c_\mu - c_0| \leq (1/2)|\gamma'_\mu(\lambda_2) - \gamma_0^+(\lambda_2)| + (1/2)|\gamma'_\mu(\lambda_1) - \gamma_0^+(\lambda_1)|$. Combining these inequalities, we have

$$\begin{aligned} & |a_\mu - a_0| + |b_\mu - b_0| + |c_\mu - c_0| \\ & \leq |\gamma'_\mu(\lambda_2) - \gamma_0^+(\lambda_2)| + |\gamma'_\mu(\lambda_1) - \gamma_0^+(\lambda_1)| + |\gamma'_\mu(\bar{\lambda}) - \gamma_0^+(\bar{\lambda})| \\ & \leq 3|\gamma'_\mu(\tilde{\lambda}(z)) - \gamma_0^+(\tilde{\lambda}(z))|, \end{aligned}$$

where the last inequality follows from $0 \leq \lambda_1 < \lambda_2$, $\bar{\lambda} \in [\lambda_1, \lambda_2]$, and Lemma 4.10(b). Hence, by (4.17), we have for some $N > 0$ that $\|\nabla P_\mu(z) - J_P^o(z)\| \leq N|\gamma'_\mu(\tilde{\lambda}(z)) - \gamma_0^+(\tilde{\lambda}(z))|$. When $\lambda_1 < \lambda_2 \leq 0$, we can obtain the desired result in a similar way.

Finally, we consider the case where $\lambda_1 < 0 < \lambda_2$. First we assume $0 < |\lambda_1| \leq \lambda_2$. When $|\lambda_1| < \lambda_2$, there exists $\bar{\lambda} \in [|\lambda_1|, \lambda_2] = [-\lambda_1, \lambda_2]$ such that

$$\begin{aligned} (4.18) \quad |a_\mu - a_0| &= \left| \frac{\gamma_\mu(\lambda_2) - \gamma_\mu(\lambda_1)}{\lambda_2 - \lambda_1} - \frac{\gamma_0(\lambda_2) - \gamma_0(\lambda_1)}{\lambda_2 - \lambda_1} \right| \\ &= \left| \frac{\gamma_\mu(\lambda_2) - \gamma_\mu(-\lambda_1)}{\lambda_2 - \lambda_1} - \frac{\gamma_0(\lambda_2) - \gamma_0(-\lambda_1)}{\lambda_2 - \lambda_1} \right| \\ &= \frac{\lambda_2 + \lambda_1}{\lambda_2 - \lambda_1} \left| \frac{\gamma_\mu(\lambda_2) - \gamma_\mu(-\lambda_1)}{\lambda_2 - (-\lambda_1)} - 1 \right| \\ &= \frac{\lambda_2 + \lambda_1}{\lambda_2 - \lambda_1} |\gamma'_\mu(\bar{\lambda}) - \gamma_0^+(\bar{\lambda})| \\ &\leq |\gamma'_\mu(\bar{\lambda}) - \gamma_0^+(\bar{\lambda})|, \end{aligned}$$

where the second equality follows from Lemma 4.10(a), the third equality follows from (4.8), the fourth equality follows from (4.9), and the inequality follows from $(\lambda_2 + \lambda_1)/(\lambda_2 - \lambda_1) = (|\lambda_2| - |\lambda_1|)/(|\lambda_2| + |\lambda_1|) \leq 1$. When $|\lambda_1| = \lambda_2$, we have $|a_\mu - a_0| = 0$ from (4.18) and Lemma 4.10(a). In a way similar to the previous case, we have $|b_\mu - b_0| \leq (1/2)|\gamma'_\mu(\lambda_2) - \gamma_0^+(\lambda_2)| + (1/2)|\gamma'_\mu(\lambda_1) - \gamma_0^+(\lambda_1)|$ and $|c_\mu - c_0| \leq (1/2)|\gamma'_\mu(\lambda_2) - \gamma_0^+(\lambda_2)| + (1/2)|\gamma'_\mu(\lambda_1) - \gamma_0^+(\lambda_1)|$. Hence, by $0 < |\lambda_1| \leq \lambda_2$, $\bar{\lambda} \in [|\lambda_1|, \lambda_2]$, Lemma 4.10(b), and (4.13), we have $\|\nabla P_\mu(z) - J_P^o(z)\| \leq N|\gamma'_\mu(\tilde{\lambda}(z)) - \gamma_0^+(\tilde{\lambda}(z))|$ for some $N > 0$. In the case where $\lambda_1 < 0 < \lambda_2$ and $0 < \lambda_2 \leq |\lambda_1|$, we can also show the desired result in a similar manner. $\square$

Proposition 4.11 indicates that, for given $\alpha \in \mathfrak{R}$, if we can reduce the value of $|\gamma'_\mu(\alpha) - \gamma_0^+(\alpha)|$ arbitrarily close to zero by choosing a sufficiently small $\mu > 0$, then we can make $\nabla H_{\mu,\varepsilon}(x, y)$ arbitrarily close to $\partial H_{\text{NR}}(x, y)$. Since $\gamma_0^+(\alpha) = \lim_{\mu \downarrow 0} \gamma'_\mu(\alpha)$, for any $\delta > 0$ and $\alpha \in \mathfrak{R}$, there exists $\bar{\mu}(\alpha, \delta) > 0$ such that $|\gamma'_\mu(\alpha) - \gamma_0^+(\alpha)| < \delta$ for any $\mu \in (0, \bar{\mu}(\alpha, \delta))$. The following proposition gives an explicit expression of such $\bar{\mu}(\alpha, \delta)$ when $\hat{g}(\alpha)$ is given by $\hat{g}(\alpha) = (\alpha + \sqrt{\alpha^2 + 4})/2$.

PROPOSITION 4.12. *Let the function $\hat{g}$ be defined by $\hat{g}(\alpha) = (\alpha + \sqrt{\alpha^2 + 4})/2$, and γ_μ and γ_0^+ be defined by (4.7) and (4.9), respectively. Moreover, let $\bar{\mu}(\alpha, \delta)$ be defined by*

$$\bar{\mu}(\alpha, \delta) := \begin{cases} +\infty & (\delta \geq 1/2 \text{ or } \alpha = 0), \\ \frac{1}{2}|\alpha|\sqrt{\delta} & (\delta < 1/2 \text{ and } \alpha \neq 0). \end{cases}$$

Then, for any $\alpha \in \mathfrak{R}$, $\delta > 0$, and $\mu \in (0, \bar{\mu}(\alpha, \delta))$, we have

$$(4.19) \quad |\gamma'_\mu(\alpha) - \gamma_0^+(\alpha)| < \delta.$$

Proof. Let μ be an arbitrary scalar in the interval $(0, \bar{\mu}(\alpha, \delta))$. By easy calculation, we have $\gamma_\mu(\alpha) = (\alpha + \sqrt{\alpha^2 + 4\mu^2})/2$, $\gamma'_\mu(\alpha) = (1 + \alpha/\sqrt{\alpha^2 + 4\mu^2})/2$, and

$$\gamma_0^+(\alpha) = \begin{cases} 1 & (\alpha > 0), \\ 1/2 & (\alpha = 0), \\ 0 & (\alpha < 0). \end{cases}$$

Moreover, these equalities yield

$$|\gamma'_\mu(\alpha) - \gamma_0^+(\alpha)| = \begin{cases} 0 & (\alpha = 0), \\ \frac{1}{2} \left(1 - \frac{|\alpha|}{\sqrt{\alpha^2 + 4\mu^2}} \right) & (\alpha \neq 0). \end{cases}$$

When $\alpha = 0$, (4.19) holds trivially since $|\gamma'_\mu(\alpha) - \gamma_0^+(\alpha)| = 0 < \delta$. When $\delta \geq 1/2$, (4.19) always holds since $|\gamma'_\mu(\alpha) - \gamma_0^+(\alpha)| < 1/2$ for any $\delta > 0$ and $\mu > 0$. When $\alpha \neq 0$ and $0 < \delta < 1/2$, we have

$$\begin{aligned} |\gamma'_\mu(\alpha) - \gamma_0^+(\alpha)| - \delta &= \frac{1}{2} \left(1 - \frac{|\alpha|}{\sqrt{\alpha^2 + 4\mu^2}} \right) - \delta \\ &< \frac{1}{2} \left(1 - \frac{|\alpha|}{\sqrt{\alpha^2 + \alpha^2 \delta}} \right) - \delta \\ &= \frac{1}{2} \left(1 - \frac{1}{\sqrt{1 + \delta}} \right) - \delta \\ &< 0, \end{aligned}$$

where the first inequality follows from $\mu < \bar{\mu}(\alpha, \delta) = (1/2)|\alpha|\sqrt{\delta}$ and the last inequality follows since the function $f(\delta) := (1/2)(1 - (1 + \delta)^{-1/2}) - \delta$ satisfies $f(0) = 0$ and $f'(\delta) = (1/4)(1 + \delta)^{-3/2} - 1 < 0$ for any $\delta > 0$. This completes the proof. $\square$

4.4. Quadratically convergent algorithm with Newton's method. In this subsection we propose a more specific algorithm, based on Algorithm 1, in which Newton's method is applied for solving subproblem (4.1). Moreover, we show that the algorithm is quadratically convergent under appropriate assumptions including strong semismoothness and Jacobian consistency. We first state the algorithm. For convenience, we define

$$w := \begin{pmatrix} x \\ y \end{pmatrix}, \quad w^{(k)} := \begin{pmatrix} x^{(k)} \\ y^{(k)} \end{pmatrix}.$$

ALGORITHM 2. Choose $\eta, \rho \in (0, 1)$, $\bar{\eta} \in (0, \eta]$, $\sigma \in (0, 1/2)$, $\kappa > 0$, and $\hat{\kappa} > 0$.

Step 0. Choose $w^{(0)} \in \mathfrak{R}^{2n}$ and $\beta_0 \in (0, \infty)$. Let $\mu_0 := \|H_{\text{NR}}(w^{(0)})\|$ and

$\varepsilon_0 := \|H_{\text{NR}}(w^{(0)})\|$. Set $k := 0$.

Step 1. Terminate if $\|H_{\text{NR}}(w^{(k)})\| = 0$.

Step 2.

Step 2.0. Set $v^{(0)} := w^{(k)}$ and $j := 0$.

Step 2.1. Find a vector $\hat{d}^{(j)}$ such that

$$H_{\mu_k, \varepsilon_k}(v^{(j)}) + \nabla H_{\mu_k, \varepsilon_k}(v^{(j)})^T \hat{d}^{(j)} = 0.$$

Step 2.2. If $\|H_{\mu_k, \varepsilon_k}(v^{(j)} + \hat{d}^{(j)})\| \leq \beta_k$, then let $w^{(k+1)} := v^{(j)} + \hat{d}^{(j)}$ and go to Step 3. Otherwise, go to Step 2.3.

Step 2.3. Find the smallest nonnegative integer m such that

$$\Psi_{\mu_k, \varepsilon_k}(v^{(j)} + \rho^m \hat{d}^{(j)}) \leq (1 - 2\sigma\rho^m) \Psi_{\mu_k, \varepsilon_k}(v^{(j)}).$$

Let $m_j := m$, $\tau_j := \rho^{m_j}$, and $v^{(j+1)} := v^{(j)} + \tau_j \hat{d}^{(j)}$.

Step 2.4. If

$$(4.20) \quad \|H_{\mu_k, \varepsilon_k}(v^{(j+1)})\| \leq \beta_k,$$

then let $w^{(k+1)} := v^{(j+1)}$ and go to Step 3. Otherwise, set $j := j + 1$ and go back to Step 2.1.

Step 3. Update the parameters as follows:

$$\begin{aligned} \mu_{k+1} &:= \min\{\kappa \|H_{\text{NR}}(w^{(k+1)})\|^2, \mu_0 \bar{\eta}^{k+1}, \\ &\quad \bar{\mu}(\tilde{\lambda}(x^{(k+1)} - y^{(k+1)}), \hat{\kappa} \|H_{\text{NR}}(w^{(k+1)})\|)\}, \\ \varepsilon_{k+1} &:= \min\{\kappa \|H_{\text{NR}}(w^{(k+1)})\|^2, \varepsilon_0 \bar{\eta}^{k+1}\}, \\ \beta_{k+1} &:= \beta_0 \eta^{k+1}. \end{aligned}$$

Set $k := k + 1$. Go back to Step 1.

In Step 3, $\tilde{\lambda}$ is the function given by (4.12), and $\bar{\mu}(\alpha, \delta)$ is determined so that $|\gamma'_\mu(\alpha) - \gamma_0^+(\alpha)| < \delta$ for any $\mu \in (0, \bar{\mu}(\alpha, \delta))$. An explicit formula for $\bar{\mu}(\alpha, \delta)$ is given as in Proposition 4.12 when $\hat{g}(\alpha) = (\alpha + \sqrt{\alpha^2 + 4})/2$.

In the inner iterations, Steps 2.0–2.4, a damped Newton method seeks a point $w^{(k+1)}$ such that $\|H_{\mu_k, \varepsilon_k}(w^{(k+1)})\| \leq \beta_k$. Note that, by letting $\alpha_k := \beta_k^2/2$, the termination criterion (4.20) for the inner iterations becomes equivalent to (4.1) in Algorithm 1. Step 3 specifies the updating rule of the parameters, where $\{\beta_k\}$, $\{\mu_k\}$, and $\{\varepsilon_k\}$ converge to 0 since $0 < \bar{\eta} \leq \eta < 1$. Algorithm 2 is well defined in the sense that Steps 2.0–2.4 find $v^{(j+1)}$ satisfying (4.20) in a finite number of iterations for each k .

Now we show that Algorithm 2 is quadratically convergent under appropriate assumptions. Note that, from (4.2) and $\|x\| \leq \|(x, y)\|$, there exists $\nu > 0$ such that

$$(4.21) \quad \|H_{\mu, \varepsilon}(w) - H_{\text{NR}}(w)\| \leq \nu\mu + \varepsilon\|w\|$$

for any $w \in \mathbb{R}^{2n}$. By using this inequality and the two properties shown in subsections 4.2 and 4.3, we establish the main theorem of this subsection.

THEOREM 4.13. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a monotone function such that ∇f is locally Lipschitzian. Let $\{w^{(k)}\}$ be a sequence generated by Algorithm 2. Moreover, suppose that the following assumptions hold true:

- (i) The solution set of SOCCP (3.1) is nonempty and bounded.
- (ii) Every accumulation point of $\{\nabla H_{\mu_k, \varepsilon_k}(w^{(k)})\}$ is nonsingular.

Then we have the following two statements:

- (a) For all k sufficiently large, the inequality in Step 2.2 of Algorithm 2 holds for $j = 0$.
- (b) The sequence $\{w^{(k)}\}$ converges to a solution w^* of SOCCP (3.1) quadratically.

Proof. By assumption (i) and Theorem 4.3, $\{w^{(k)}\}$ is bounded, and an arbitrary accumulation point w^* is a solution of SOCCP (3.1), that is, $H_{\text{NR}}(w^*) = 0$. Let B be a positive number such that $\|w^{(k)}\| \leq B$ for all k . Let $L > 0$ be a

Lipschitzian constant of H_{NR} on a bounded set $\Omega \supset \{w^{(k)}\}$. Let $d^{(k)} := \hat{d}^{(0)} = -\nabla H_{\mu_k, \varepsilon_k}(w^{(k)})^{-T} H_{\mu_k, \varepsilon_k}(w^{(k)})$, where J^{-T} denotes $J^{-T} := (JT)^{-1} = (J^{-1})^T$. From the nonsingularity of $\nabla H_{\mu_k, \varepsilon_k}(w^{(k)})$ [11, Proposition 6.1] and assumption (ii), there exists $C > 0$ such that

$$(4.22) \quad \|\nabla H_{\mu_k, \varepsilon_k}(w^{(k)})^{-T}\| = \|\nabla H_{\mu_k, \varepsilon_k}(w^{(k)})^{-1}\| \leq C$$

for all k .

First we show $\|w^{(k)} + d^{(k)} - w^*\| = O(\|w^{(k)} - w^*\|^2)$. Let V_k be an element of $\partial H_{\text{NR}}(w^{(k)})$ such that $\|\nabla H_{\mu_k, \varepsilon_k}(w^{(k)}) - V_k\| = \text{dist}(\nabla H_{\mu_k, \varepsilon_k}(w^{(k)}), \partial H_{\text{NR}}(w^{(k)}))$. It then follows from (4.22) and $H_{\text{NR}}(w^*) = 0$ that

$$(4.23) \quad \begin{aligned} & \|w^{(k)} + d^{(k)} - w^*\| \\ &= \|w^{(k)} - \nabla H_{\mu_k, \varepsilon_k}(w^{(k)})^{-T} H_{\mu_k, \varepsilon_k}(w^{(k)}) - w^*\| \\ &\leq \|\nabla H_{\mu_k, \varepsilon_k}(w^{(k)})^{-T}\| \|\nabla H_{\mu_k, \varepsilon_k}(w^{(k)})^T (w^{(k)} - w^*) - H_{\mu_k, \varepsilon_k}(w^{(k)})\| \\ &\leq C\{\|(\nabla H_{\mu_k, \varepsilon_k}(w^{(k)}) - V_k)^T (w^{(k)} - w^*)\| + \|V_k^T (w^{(k)} - w^*) - H'_{\text{NR}}(w^*; w^{(k)} - w^*)\| \\ &\quad + \|H'_{\text{NR}}(w^*; w^{(k)} - w^*) - H_{\text{NR}}(w^{(k)}) + H_{\text{NR}}(w^*)\| + \|H_{\text{NR}}(w^{(k)}) - H_{\mu_k, \varepsilon_k}(w^{(k)})\|\}. \end{aligned}$$

Moreover, each term of (4.23) is $O(\|w^{(k)} - w^*\|^2)$ for the following reasons. Since $H_{\mu, \varepsilon}$ satisfies the Jacobian consistency, we have for some $M > 0$ that

$$(4.24) \quad \begin{aligned} \|\nabla H_{\mu_k, \varepsilon_k}(w^{(k)}) - V_k\| &\leq M(|\gamma'_{\mu_k}(\tilde{\lambda}(x^{(k)} - y^{(k)})) - \gamma_0^+(\tilde{\lambda}(x^{(k)} - y^{(k)}))| + \varepsilon_k) \\ &\leq M(\hat{\kappa}\|H_{\text{NR}}(w^{(k)})\| + \kappa\|H_{\text{NR}}(w^{(k)})\|^2) \\ &\leq M(\hat{\kappa} + \kappa\|H_{\text{NR}}(w^{(k)})\|)\|H_{\text{NR}}(w^{(k)}) - H_{\text{NR}}(w^*)\| \\ &\leq M(\hat{\kappa} + \kappa\|H_{\text{NR}}(w^{(k)})\|)L\|w^{(k)} - w^*\|, \end{aligned}$$

where the first inequality follows from Proposition 4.11, the second inequality follows from Step 3 of the algorithm, the third inequality follows from $H_{\text{NR}}(w^*) = 0$, and the last inequality is due to the local Lipschitz continuity of H_{NR} . By (4.24) and the boundedness of $\{\|H_{\text{NR}}(w^{(k)})\|\}$, we have $\|(\nabla H_{\mu_k, \varepsilon_k}(w^{(k)}) - V_k)^T (w^{(k)} - w^*)\| = O(\|w^{(k)} - w^*\|^2)$. The second and third terms of (4.23) are also $O(\|w^{(k)} - w^*\|^2)$ since H_{NR} is strongly semismooth and directionally differentiable. Finally, we consider the fourth term of (4.23). From (4.21) and Step 3 of Algorithm 2, we have $\|H_{\text{NR}}(w^{(k)}) - H_{\mu_k, \varepsilon_k}(w^{(k)})\| \leq \nu\mu_k + B\varepsilon_k = O(\|H_{\text{NR}}(w^{(k)})\|^2)$. Moreover, $O(\|H_{\text{NR}}(w^{(k)})\|^2) = O(\|w^{(k)} - w^*\|^2)$ from $H_{\text{NR}}(w^*) = 0$ and local Lipschitz continuity of H_{NR} . Therefore, the fourth term of (4.23) is $O(\|w^{(k)} - w^*\|)$. Consequently, we have

$$(4.25) \quad \|w^{(k)} + d^{(k)} - w^*\| = O(\|w^{(k)} - w^*\|^2).$$

Next we show (a), that is, $\|H_{\mu_k, \varepsilon_k}(w^{(k)} + d^{(k)})\| \leq \beta_k$ for all k sufficiently large. Note that $\{d^{(k)}\}$ is bounded from the inequality $\|d^{(k)}\| \leq \|w^{(k)} + d^{(k)} - w^*\| + \|w^{(k)} - w^*\|$ and (4.25). We therefore have for some $\Gamma > 0$ that

$$(4.26) \quad \begin{aligned} & \|H_{\mu_k, \varepsilon_k}(w^{(k)} + d^{(k)})\| \\ &\leq \|H_{\mu_k, \varepsilon_k}(w^{(k)} + d^{(k)}) - H_{\text{NR}}(w^{(k)} + d^{(k)})\| + \|H_{\text{NR}}(w^{(k)} + d^{(k)})\| \\ &\leq \Gamma\|H_{\text{NR}}(w^{(k)})\|^2 + \|H_{\text{NR}}(w^{(k)} + d^{(k)})\|, \end{aligned}$$

where the second inequality follows from (4.21), the boundedness of $\{w^{(k)} + d^{(k)}\}$, and Step 3 of Algorithm 2. Hence, it suffices to show $\|H_{\text{NR}}(w^{(k)} + d^{(k)})\| = O(\|H_{\text{NR}}(w^{(k)})\|^2)$ and $\|H_{\text{NR}}(w^{(k)})\| = O(\eta^k)$. Notice that assumption (ii) implies that H_{NR} satisfies the local error bound property at w^* , that is, $\|w^{(k)} - w^*\| = O(\|H_{\text{NR}}(w^{(k)})\|)$. Then, we have $\|H_{\text{NR}}(w^{(k)} + d^{(k)})\| \leq \bar{L}\|w^{(k)} + d^{(k)} - w^*\| = O(\|w^{(k)} - w^*\|^2) = O(\|H_{\text{NR}}(w^{(k)})\|^2)$, where $\bar{L}$ is a Lipschitzian constant for H_{NR} on a bounded set $\bar{\Omega} \supset \{w^{(k)} + d^{(k)}\}$. On the other hand, $\|H_{\text{NR}}(w^{(k)})\| = O(\eta^k)$ since

$$\begin{aligned} \|H_{\text{NR}}(w^{(k)})\| &\leq \|H_{\mu_{k-1}, \varepsilon_{k-1}}(w^{(k)})\| + \|H_{\mu_{k-1}, \varepsilon_{k-1}}(w^{(k)}) - H_{\text{NR}}(w^{(k)})\| \\ &\leq \beta_{k-1} + \nu\mu_{k-1} + B\varepsilon_{k-1} \\ &\leq \beta_0\eta^{k-1} + \nu\mu_0\eta^{k-1} + B\varepsilon_0\eta^{k-1} \\ &= \eta^{-1}(\beta_0 + \nu\mu_0 + B\varepsilon_0)\eta^k, \end{aligned}$$

where the second inequality follows from Step 2.4 of Algorithm 2 and (4.21), and the third inequality follows from Step 3 in Algorithm 2 and $0 < \bar{\eta} \leq \eta < 1$. Combining these results with (4.26), we have $\|H_{\mu_k, \varepsilon_k}(w^{(k)} + d^{(k)})\| \leq \|H_{\text{NR}}(w^{(k)})\|O(\eta^k)$, which implies (a). Finally, (b) directly follows from (4.25) and (a). $\square$

We note that assumption (ii) holds if f is strongly monotone around a solution. This can be observed as follows. From [11, Proposition 6.1], we have $O \prec \nabla P_\mu(z) \prec I$ for any $\mu > 0$ and $z \in \mathbb{R}^n$, which implies $O \preceq J_P^o(z) \preceq I$. Moreover, the matrix $J_H^o(x, y)$ defined by (4.6) is nonsingular since the matrix $I - J_P^o(x - y) + J_P^o(x - y)\nabla f(x)$ is nonsingular by the positive definiteness of $\nabla f(x)$ and [26, Proposition 2.1(b)].

Many Newton-type methods with smoothing and regularization techniques have been studied for solving NCPs and box constrained variational inequalities (BVI). Our algorithm is regarded as an extension of those methods to SOCCP. However, it is different from the existing Newton-type methods for NCP and BVI in some respects. The Jacobian smoothing method [4, 12] is one of the popular Newton-type methods. In this method, an approximate Newton direction $d = -\nabla H_\mu(z)^{-T}H(z)$ of H is used instead of Newton direction $-\nabla H_\mu(z)^{-T}H_\mu(z)$ of H_μ itself, which our algorithm adopts. Another approach [17, 22] is to consider the nonlinear equation $G(x, \mu) = 0$, where $G(x, \mu) := \begin{pmatrix} H_\mu(z) \end{pmatrix}$, and to solve it by Newton's method that treats parameter μ as a variable. Moreover, in [17], a regularization parameter is also included. On the other hand, our method distinguishes between parameters and variables strictly.

5. Numerical experiments. In order to evaluate the efficiency of Algorithm 2, we have conducted some numerical experiments. In our experiments, we chose $\hat{g}(\alpha) = (\sqrt{\alpha^2 + 4} + \alpha)/2$ as the function satisfying (3.3) and (4.10) and employed $\|H_{\text{NR}}(w^{(k)})\| < 10^{-8}$ as the termination criterion. Moreover, we adopted

$$\bar{\mu}(\alpha, \delta) := \begin{cases} 10^{10} & (\delta \geq 1/2 \text{ or } \alpha = 0), \\ \frac{1}{2}|\alpha|\sqrt{\delta} & (\delta < 1/2 \text{ and } \alpha \neq 0) \end{cases}$$

in Step 3 of the algorithm. Note that, when $\tilde{\lambda}(x^{(k)} - y^{(k)})$ defined by (4.12) is very small but positive while $\|H_{\text{NR}}(x^{(k)}, y^{(k)})\|$ is not small enough, μ_k becomes almost as small as $\tilde{\lambda}(x^{(k)} - y^{(k)})$, although $(x^{(k)}, y^{(k)})$ is not sufficiently close to a solution. Since this may diminish the effects of smoothing, we regard $\lambda_i(x^{(k)} - y^{(k)})$ as 0 when $\lambda_i(x^{(k)} - y^{(k)})/\|H_{\text{NR}}(x^{(k)}, y^{(k)})\| \leq \bar{\delta}$ holds for a sufficiently small $\bar{\delta} > 0$. In our implementation, we therefore modified the definition (4.12) of $\tilde{\lambda}(z)$ such that the

TABLE 5.1
Results for linear SOCCPs of various problem sizes.

n	#Ite	#Newton	cpu(s)
100	5.28	7.12	0.151
200	5.58	7.93	1.026
300	5.86	8.48	2.848
400	5.65	8.80	5.960
500	5.77	8.89	10.603
600	5.80	9.12	17.346
700	5.96	9.37	26.252
800	5.99	9.36	37.185
900	5.93	9.71	52.223
1000	5.98	9.52	67.316

index set $\mathcal{I}(z) := \{i \mid |\lambda_i(z)| > 10^{-4} \|H_{\text{NR}}(x, y)\|\}$ was used instead of $\mathcal{I}(z) = \{i \mid \lambda_i(z) \neq 0\}$. The program was coded in MATLAB 6.5 and run on a machine with AMD Athlon(TM) XP 2000+ CPU and 1 GB RAM.

5.1. Linear case. We solved the following linear SOCCP: Find $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$ such that

$$(5.1) \quad x \in \mathcal{K}^n, \quad y \in \mathcal{K}^n, \quad x^T y = 0, \quad y = Mx + q,$$

where $M \in \mathbb{R}^{n \times n}$ is a rank-deficient positive semidefinite matrix. We chose $\eta = 0.01$, $\bar{\eta} = 0.001$, $\rho = 0.5$, $\sigma = 0.4$, $\kappa = 0.01$, and $\hat{\kappa} = 1$ in Algorithm 2.

In order to obtain a positive semidefinite matrix M with $\text{rank } M = r < n$, we let $M := nBB^T / \|BB^T\|$, where $B \in \mathbb{R}^{n \times r}$ is a matrix of which components are randomly chosen from the interval $[-1, 1]$. Furthermore, we let $q := 10^\alpha n^{1/2} p - Me$, where $e = (1, 0, \dots, 0)^T \in \text{int } \mathcal{K}^n$, p is a vector such that $p \in \text{int } \mathcal{K}^n$ and $\|p\| = 1$, and α is randomly chosen from the interval $[-1, 1]$. Then, SOCCP (5.1) has a solution, since M is positive semidefinite and there exist $\bar{x} \in \text{int } \mathcal{K}^n$ and $\bar{y} \in \text{int } \mathcal{K}^n$ such that $\bar{y} = M\bar{x} + q$. In the experiments, we determined p as follows. Let θ be randomly chosen from $(0, \pi/2)$, let w be an $(n-1)$ -dimensional vector whose components are randomly chosen from $[-1, 1]$, and set $p := 2^{-1/2} \cos \theta(1, w/\|w\|) + 2^{-1/2} \sin \theta(1, -w/\|w\|)$. It is not difficult to see that a vector p thus determined satisfies $\|p\| = 1$ and $p \in \text{int } \mathcal{K}^n$.

In our first experiments, we generated 100 problem instances for each of $n = 100, 200, \dots, 1000$, and solved each problem instance by using 100 randomly chosen initial points. We let the rank r of M be an integer randomly chosen from $[0.9n, n-1]$, and selected an initial point $(x^{(0)}, y^{(0)})$ as $(x^{(0)}, y^{(0)}) := 10^\beta (a, b) / \|(a, b)\|$, where β is randomly chosen from $[-3, 3]$ and each component of $(a, b) \in \mathbb{R}^n \times \mathbb{R}^n$ is randomly chosen from $[-1, 1]$. Table 5.1 shows the results of our experiments, in which n denotes the size of problems, #Ite denotes the number of outer iterations, #Newton denotes the total number of inner Newton iterations, and cpu(s) denotes the CPU time in seconds. The values of #Ite, #Newton, and cpu(s) are the average of 100 runs for each n . Table 5.1 reveals that the problem size only slightly affects the number of iterations.

In our second experiment, we fixed the size of the problems at 100 and varied the rank of $M \in \mathbb{R}^{100 \times 100}$ as $r = 10, 20, \dots, 90$ and 99. We solved 1000 different problem instances for each r , where initial points are selected in a way similar to the previous

TABLE 5.2
Results for linear SOCCPs with various degrees of rank deficiency.

r	#Ite	#Newton
99	5.396	7.346
90	5.361	7.260
80	5.284	7.207
70	5.245	7.115
60	5.242	7.189
50	5.172	7.004
40	5.233	7.128
30	5.116	6.988
20	5.072	6.809
10	4.894	6.371

experiments. Table 5.2 shows the results of the experiments. The values of #Ite and #Newton are the average of 1000 runs for each r . Table 5.2 indicates that neither #Ite nor #Newton is affected by the rank of matrix M .

5.2. Comparison with the interior-point method. We solved the following SOCP by our method and the interior-point method:

$$(5.2) \quad \begin{aligned} & \text{Minimize} && c^T z \\ & \text{subject to} && z \in \mathcal{K}_1, \quad Az + b \in \mathcal{K}_2, \end{aligned}$$

where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$, $\mathcal{K}_1 := \mathcal{K}^{n_1} \times \cdots \times \mathcal{K}^{n_N} \subset \mathbb{R}^n$, and $\mathcal{K}_2 := \mathcal{K}^{m_1} \times \cdots \times \mathcal{K}^{m_M} \subset \mathbb{R}^m$. Since the KKT conditions for SOCP (5.2) are written as

$$(5.3) \quad \begin{aligned} & \begin{pmatrix} z \\ \lambda \end{pmatrix} \in \mathcal{K}_1 \times \mathcal{K}_2, \quad \begin{pmatrix} O & -A^T \\ A & O \end{pmatrix} \begin{pmatrix} z \\ \lambda \end{pmatrix} + \begin{pmatrix} c \\ b \end{pmatrix} \in \mathcal{K}_1 \times \mathcal{K}_2, \\ & \begin{pmatrix} z \\ \lambda \end{pmatrix}^T \left\{ \begin{pmatrix} O & -A^T \\ A & O \end{pmatrix} \begin{pmatrix} z \\ \lambda \end{pmatrix} + \begin{pmatrix} c \\ b \end{pmatrix} \right\} = 0, \end{aligned}$$

we can obtain a solution of SOCP (5.2) by solving SOCCP (5.3). In this experiment, we chose eight pairs of $(\mathcal{K}_1, \mathcal{K}_2)$ such that $n_1 = \cdots = n_N$ and $m_1 = \cdots = m_M$ and generated 100 problem instances for each $(\mathcal{K}_1, \mathcal{K}_2)$, where matrix A is randomly chosen so that its components are contained in the interval $[-1, 1]$, and vectors b and c are randomly chosen so that SOCP (5.2) is solvable.

In applying Algorithm 2 to SOCCP (5.3), we chose parameters and initial points in a way similar to the previous experiment. As the interior-point method for SOCP, we used the SDPT3 solver [24]. We give the results in Table 5.3, in which (n, N) denotes the dimension of $\mathcal{K}_1$ and the number of second-order cones comprising $\mathcal{K}_1$, (m, M) applies similarly to $\mathcal{K}_2$, and the numbers of iterations are the average of 100 runs for each $(\mathcal{K}_1, \mathcal{K}_2)$. As this table shows, our method tends to require fewer iterations than the interior-point method when the dimension of the second-order cone is large.

5.3. Nonlinear case. We solved the following nonlinear SOCCP: Find $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$ such that

$$(5.4) \quad x \in \mathcal{K}, \quad y \in \mathcal{K}, \quad x^T y = 0, \quad y = f(x),$$

TABLE 5.3

Comparison of Algorithm 2 and the interior point method on linear SOCPs.

(n, N)	(m, M)	Our method		SDPT3
		# Ite	# Newton	# Ite
(100,2)	(100,4)	5.07	11.07	12.58
(100,20)	(100,25)	5.21	18.55	15.81
(400,8)	(300,5)	5.30	9.23	14.05
(400,40)	(300,50)	5.53	11.52	15.24
(800,4)	(900,3)	5.74	9.67	15.19
(800,50)	(900,60)	5.71	12.24	16.12

TABLE 5.4

Results for nonlinear SOCCPs with various choices of $(\eta, \bar{\eta})$.

η	$\bar{\eta}$	# Ite	# Newton
0.5	0.5	16.79	26.34
0.5	0.1	12.20	13.70
0.5	0.01	10.50	10.75
0.5	0.001	9.58	9.85
0.1	0.1	8.35	19.11
0.1	0.01	8.13	13.51
0.1	0.001	7.66	11.25
0.01	0.01	5.99	14.32
0.01	0.001	5.73	12.35

where $\mathcal{K} = \mathcal{K}^3 \times \mathcal{K}^2$ and $f : \Re^5 \rightarrow \Re^5$ is given by

$$f(x) := \begin{pmatrix} 24(2x_1 - x_2)^3 + \exp(x_1 - x_3) - 4x_4 + x_5 \\ -12(2x_1 - x_2)^3 + 3(3x_2 + 5x_3)/\sqrt{1 + (3x_2 + 5x_3)^2} - 6x_4 - 7x_5 \\ -\exp(x_1 - x_3) + 5(3x_2 + 5x_3)/\sqrt{1 + (3x_2 + 5x_3)^2} - 3x_4 + 5x_5 \\ 4x_1 + 6x_2 + 3x_3 - 1 \\ -x_1 + 7x_2 - 5x_3 + 2 \end{pmatrix}.$$

We note that, in view of the KKT conditions (6.4) for the SOCP (6.1) (see section 6), SOCCP (5.4) is equivalent to the following SOCP:

$$\begin{aligned} & \text{Minimize} \quad \exp(z_1 - z_3) + 3(2z_1 - z_2)^4 + \sqrt{1 + (3z_2 + 5z_3)^2} \\ & \text{subject to} \quad \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} \in \mathcal{K}^3, \quad \begin{pmatrix} 4 & 6 & 3 \\ -1 & 7 & -5 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} + \begin{pmatrix} -1 \\ 2 \end{pmatrix} \in \mathcal{K}^2. \end{aligned}$$

Since the objective function of this SOCP is convex, we can easily see that the function f is monotone. In the experiments, we set the parameters $\rho = 0.5$, $\sigma = 0.4$, $\kappa = 0.01$, and $\hat{\kappa} = 1$.

We used nine different pairs $(\eta, \bar{\eta})$ and ran Algorithm 2 with 100 initial points $(x^{(0)}, y^{(0)})$ for each $(\eta, \bar{\eta})$ determined as $(x^{(0)}, y^{(0)}) := \Gamma(a, b)/\|(a, b)\|$, where Γ is randomly chosen from $[0, 10]$ and each component of $(a, b) \in \Re^n \times \Re^n$ is randomly chosen from $[-1, 1]$. Table 5.4 shows the results of the experiments, in which # Ite and # Newton are the average of 100 runs for each $(\eta, \bar{\eta})$. As Table 5.4 shows, the number

of outer iterations decreases as η becomes smaller. On the other hand, the number of inner iterations decreases as $\bar{\eta}$ becomes smaller and η becomes larger. Since the computation time is largely dependent on the number of inner iterations, it would be recommended to choose η as large as possible and $\bar{\eta}$ as small as possible. However, it should be kept in mind that, when $\bar{\eta}$ is too small, μ and ε tend to 0 very rapidly, which may deteriorate the advantage of smoothing and regularization methods.

6. Concluding remarks. In this section, we make some comments that complement the results obtained in the paper.

In this paper, the function F in SOCCP (1.1) is assumed to be of the form $F(x, y, \zeta) = f(x) - y$. This assumption may seem rather restrictive. However, the KKT conditions for the SOCP,

$$(6.1) \quad \begin{aligned} & \text{Minimize} \quad \theta(z) \\ & \text{subject to} \quad \gamma(z) \in \mathcal{K}, \end{aligned}$$

can be written as the SOCCP with $F(x, y, \zeta) = f(x) - y$ as follows. In SOCP (6.1), let $z = z' - z''$ with $z' \in \mathbb{R}_+^s$ and $z'' \in \mathbb{R}_+^s$, and denote $\hat{z} := \begin{pmatrix} z' \\ z'' \end{pmatrix} \in \mathbb{R}_+^{2s}$, where $\mathbb{R}_+^n$ is the n -dimensional nonnegative orthant. Moreover, define $\hat{\theta} : \mathbb{R}^{2s} \rightarrow \mathbb{R}$ by $\hat{\theta}(\hat{z}) = \theta(z' - z'')$ and $\hat{\gamma} : \mathbb{R}^{2s} \rightarrow \mathbb{R}^t$ by $\hat{\gamma}(\hat{z}) = \gamma(z' - z'')$. Then SOCP (6.1) can be reformulated as

$$(6.2) \quad \begin{aligned} & \text{Minimize} \quad \hat{\theta}(\hat{z}) \\ & \text{subject to} \quad \begin{pmatrix} \hat{\gamma}(\hat{z}) \\ \hat{z} \end{pmatrix} \in \mathcal{K} \times \mathbb{R}_+^{2s}, \end{aligned}$$

and the KKT conditions for (6.2) are written as

$$(6.3) \quad \begin{aligned} & \nabla \hat{\theta}(\hat{z}) - (\nabla \hat{\gamma}(\hat{z}) \ I) \begin{pmatrix} \hat{\lambda}_1 \\ \hat{\lambda}_2 \end{pmatrix} = 0, \\ & \begin{pmatrix} \hat{\lambda}_1 \\ \hat{\lambda}_2 \end{pmatrix} \in \mathcal{K} \times \mathbb{R}_+^{2s}, \quad \begin{pmatrix} \hat{\gamma}(\hat{z}) \\ \hat{z} \end{pmatrix} \in \mathcal{K} \times \mathbb{R}_+^{2s}, \quad \begin{pmatrix} \hat{\lambda}_1 \\ \hat{\lambda}_2 \end{pmatrix}^T \begin{pmatrix} \hat{\gamma}(\hat{z}) \\ \hat{z} \end{pmatrix} = 0. \end{aligned}$$

Now, let $\hat{\mu}_1 = \hat{\gamma}(\hat{z})$, and notice that (2.7) holds. Then, (6.3) can be rewritten as

$$(6.4) \quad \begin{aligned} & \begin{pmatrix} \hat{\gamma}(\hat{z}) \\ \nabla \hat{\theta}(\hat{z}) - \nabla \hat{\gamma}(\hat{z}) \hat{\lambda}_1 \end{pmatrix} = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\lambda}_2 \end{pmatrix}, \\ & \begin{pmatrix} \hat{\lambda}_1 \\ \hat{z} \end{pmatrix} \in \mathcal{K} \times \mathbb{R}_+^{2s}, \quad \begin{pmatrix} \hat{\mu}_1 \\ \hat{\lambda}_2 \end{pmatrix} \in \mathcal{K} \times \mathbb{R}_+^{2s}, \quad \begin{pmatrix} \hat{\lambda}_1 \\ \hat{z} \end{pmatrix}^T \begin{pmatrix} \hat{\mu}_1 \\ \hat{\lambda}_2 \end{pmatrix} = 0. \end{aligned}$$

Setting

$$x = \begin{pmatrix} \hat{\lambda}_1 \\ \hat{z} \end{pmatrix}, \quad y = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\lambda}_2 \end{pmatrix}, \quad f(x) = \begin{pmatrix} \hat{\gamma}(\hat{z}) \\ \nabla \hat{\theta}(\hat{z}) - \nabla \hat{\gamma}(\hat{z}) \hat{\lambda}_1 \end{pmatrix},$$

the KKT conditions (6.4) for SOCP (6.1) can be reduced to the SOCCP with $F(x, y, \zeta) = f(x) - y$. Note that the KKT conditions (6.4) for SOCP (6.1) contain more variables than the original KKT conditions. Furthermore, some desirable properties of the functions involved in SOCP (6.1) may be lost. For example, even if θ and γ in SOCP (6.1) are strictly convex, $\hat{\theta}$ and $\hat{\gamma}$ in (6.2) are merely convex. Hence, it would be useful to develop a method that can directly deal with the original KKT conditions, or more

generally, SOCCP involving the function $F(x, y, \zeta)$ which is not restricted to be of the form $F(x, y, \zeta) = f(x) - y$.

In sections 3 and 4, it is also assumed that $\mathcal{K} = \mathcal{K}^n$. For the general case where $\mathcal{K} = \mathcal{K}^{n_1} \times \cdots \times \mathcal{K}^{n_m}$, $H_{\mu, \varepsilon}(x, y)$ and $\nabla H_{\mu, \varepsilon}(x, y)$ are given by

$$H_{\mu, \varepsilon}(x, y) = \begin{pmatrix} \varphi_{\mu}(x^1, y^1) \\ \vdots \\ \varphi_{\mu}(x^m, y^m) \\ f_{\varepsilon}(x) - y \end{pmatrix},$$

$$\nabla H_{\mu, \varepsilon}(x, y) = \begin{pmatrix} I - \text{diag}\{\nabla P_{\mu}(x^j - y^j)\}_{j=1}^m & \nabla f(x) + \varepsilon I \\ \text{diag}\{\nabla P_{\mu}(x^j - y^j)\}_{j=1}^m & -I \end{pmatrix},$$

where $x = (x^1, \dots, x^m) \in \mathfrak{R}^{n_1} \times \cdots \times \mathfrak{R}^{n_m}$, $y = (y^1, \dots, y^m) \in \mathfrak{R}^{n_1} \times \cdots \times \mathfrak{R}^{n_m}$, and $\text{diag}\{\nabla P_{\mu}(x^j - y^j)\}_{j=1}^m$ denotes the block-diagonal matrix with block elements $\nabla P_{\mu}(x^j - y^j)$. Moreover, the definition (4.12) of $\tilde{\lambda}(z)$ is replaced by

$$\tilde{\lambda}(z) := \begin{cases} \min_{(i,j) \in \mathcal{I}(z)} |\lambda_i(z^j)| & (\mathcal{I}(z) \neq \emptyset), \\ 0 & (\mathcal{I}(z) = \emptyset), \end{cases}$$

where $z = (z^1, \dots, z^m) \in \mathfrak{R}^{n_1} \times \cdots \times \mathfrak{R}^{n_m}$, $\lambda_i(z^j)$ ($i = 1, 2$) are the spectral values of z^j , and $\mathcal{I}(z) \subseteq \{1, 2\} \times \{1, \dots, m\}$ is the index set defined by $\mathcal{I}(z) := \{(i, j) \mid \lambda_i(z^j) \neq 0\}$. Then, we can also show that Theorems 4.3 and 4.13 hold for monotone f in a similar way.

Acknowledgments. The authors are grateful to the anonymous referees for their very helpful comments.

REFERENCES

- [1] C. CHEN AND O. L. MANGASARIAN, *A class of smoothing functions for nonlinear and mixed complementarity problems*, Comput. Optim. Appl., 5 (1996), pp. 97–138.
- [2] J.-S. CHEN, X. CHEN, AND P. TSENG, *Analysis of nonsmooth vector-valued functions associated with second-order cones*, Math. Program., 101 (2004), pp. 95–117.
- [3] X. CHEN, *Smoothing methods for complementarity problems and their applications: A survey*, J. Oper. Res. Soc. Japan, 43 (2000), pp. 32–47.
- [4] X. CHEN, L. QI, AND D. SUN, *Global and superlinear convergence of the smoothing Newton method and its application to general box constrained variational inequalities*, Math. Comput., 67 (1998), pp. 519–540.
- [5] X. D. CHEN, D. SUN, AND J. SUN, *Complementarity functions and numerical experiments for second-order-cone complementarity problems*, Comput. Optim. Appl., 25 (2003), pp. 39–56.
- [6] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.
- [7] A. L. DONTCHEV AND T. ZOLEZZI, *Well-Posed Optimization Problems*, Lecture Notes in Math. 1543, Springer-Verlag, Berlin, 1993.
- [8] F. FACCHINEI AND C. KANZOW, *Beyond monotonicity in regularization methods for nonlinear complementarity problems*, SIAM J. Control Optim., 37 (1999), pp. 1150–1161.
- [9] F. FACCHINEI AND J.-S. PANG, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Springer-Verlag, New York, 2003.
- [10] J. FARAUT AND A. KORÁNYI, *Analysis on Symmetric Cones*, Oxford Mathematical Monographs, Oxford University Press, New York, 1994.
- [11] M. FUKUSHIMA, Z.-Q. LUO, AND P. TSENG, *Smoothing functions for second-order cone complementarity problems*, SIAM J. Optim., 12 (2001), pp. 436–460.

- [12] C. KANZOW AND H. PIEPER, *Jacobian smoothing methods for nonlinear complementarity problems*, SIAM J. Optim., 9 (1999), pp. 342–373.
- [13] M. S. LOBO, L. VANDENBERGHE, S. BOYD, AND H. LEBRET, *Applications of second-order cone programming*, Linear Algebra Appl., 284 (1998), pp. 193–228.
- [14] R. MIFFLIN, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15 (1977), pp. 959–972.
- [15] R. D. C. MONTEIRO AND T. TSUCHIYA, *Polynomial convergence of primal-dual algorithms for the second-order cone program based on the MZ-family of directions*, Math. Program., 88 (2000), pp. 61–83.
- [16] J. M. PENG AND M. FUKUSHIMA, *A hybrid Newton method for solving the variational inequality problem via the D-gap function*, Math. Program., 86 (1999), pp. 367–386.
- [17] H.-D. QI, *A regularized smoothing Newton method for box constrained variational inequality problems with P_0 -functions*, SIAM J. Optim., 10 (1999), pp. 315–330.
- [18] L. QI, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res., 18 (1993), pp. 227–244.
- [19] L. QI AND X. CHEN, *A globally convergent successive approximation method for severely nonsmooth equations*, SIAM J. Control Optim., 33 (1995), pp. 402–418.
- [20] L. QI AND D. SUN, *A nonsmooth version of Newton's method*, Math. Programming, 58 (1993), pp. 353–367.
- [21] L. QI AND D. SUN, *Smoothing functions and a smoothing Newton method for complementarity and variational inequality problems*, J. Optim. Theory Appl., 113 (2002), pp. 121–147.
- [22] L. QI, D. SUN, AND G. ZHOU, *A new look at smoothing Newton methods for nonlinear complementarity problems and box constrained variational inequalities*, Math. Program., 87 (2000), pp. 1–35.
- [23] A. SHAPIRO, *On a class of nonsmooth composite functions*, Math. Oper. Res., 28 (2003), pp. 677–692.
- [24] K. C. TOH, R. H. TÜTÜNCÜ, AND M. J. TODD, *SDPT3 Version 3.02—A MATLAB Software for Semidefinite-Quadratic-Linear Programming*, 2002, available online at <http://www.math.nus.edu.sg/~mattohkc/sdpt3.html>.
- [25] T. TSUCHIYA, *A convergence analysis of the scaling-invariant primal-dual path-following algorithms for second-order cone programming*, Optim. Methods Softw., 11 & 12 (1999), pp. 141–182.
- [26] N. YAMASHITA AND M. FUKUSHIMA, *Modified Newton methods for solving a semismooth reformulation of monotone complementarity problems*, Math. Programming, 76 (1997), pp. 469–491.
- [27] N. YAMASHITA AND M. FUKUSHIMA, *On the Level-Boundedness of the Natural Residual Function for Variational Inequality Problems*, Tech. report 2002-005, Department of Applied Mathematics and Physics, Kyoto University, Kyoto, Japan, 2002.

REGULAR LAGRANGE MULTIPLIERS FOR CONTROL PROBLEMS WITH MIXED POINTWISE CONTROL-STATE CONSTRAINTS*

FREDI TRÖLTZSCH†

Abstract. A class of quadratic optimization problems in Hilbert spaces is considered, where pointwise box-constraints and constraints of bottleneck type are given. The main focus is on proving the existence of regular Lagrange multipliers in L^2 -spaces. This question is solved by investigating the solvability of a Lagrange dual quadratic problem. The theory is applied to different optimal control problems for elliptic and parabolic partial differential equations with mixed pointwise control-state constraints.

Key words. quadratic programming, Hilbert space, regular Lagrange multipliers, optimal control, elliptic and parabolic equations, mixed pointwise control-state constraint

AMS subject classifications. 49K20, 49N10, 49N15, 90C45

DOI. 10.1137/S1052623403426519

1. A class of optimization problems. We consider the following general class of optimization problems in Hilbert spaces,

$$(PP) \quad \text{minimize} \quad \frac{1}{2} \|Su\|_H^2 + \int_D \left(a(x)u(x) + \frac{\kappa}{2} u(x)^2 \right) dx,$$

subject to the pointwise constraints

$$\begin{aligned} u(x) &\leq c(x) + (Gu)(x), \\ u(x) &\leq b(x), \\ u(x) &\geq 0, \end{aligned}$$

to be fulfilled for almost all $x \in D$, where the function u is taken from $L^2(D)$.

In this problem, henceforth called the *primal problem*, $D \subset \mathbb{R}^N$ is a Lebesgue-measurable bounded set, S and G are linear continuous operators from $L^2(D)$ to a real Hilbert space H and $L^2(D)$, respectively, while a, b, c are fixed functions from $L^2(D)$ and $\kappa > 0$ is a fixed constant. Let us denote the natural inner product of $L^2(D)$ by $(\cdot, \cdot)$ and the associated norm by $\|\cdot\|$. In what follows, we denote the primal objective function by f , i.e.,

$$f(u) = \frac{1}{2} \|Su\|_H^2 + \frac{\kappa}{2} \|u\|^2 + (a, u).$$

A function $u \in L^2(D)$ is said to be *feasible* if it satisfies the pointwise constraints given above.

This class of problems is large enough to cover several types of linear-quadratic optimal control problems for elliptic or parabolic equations with pointwise control-state constraints. Below, we briefly sketch some possible examples, where Su and Gu stand for certain observations of the state y of the control system. For instance,

*Received by the editors April 25, 2003; accepted for publication (in revised form) April 20, 2004; published electronically February 19, 2005. Supported by the DFG Research Center “Mathematics for key technologies” (FZT86) in Berlin.

<http://www.siam.org/journals/siopt/15-2/42651.html>

†Institut für Mathematik, Technische Universität Berlin, Str. des 17. Juni 136, D-10623 Berlin, Germany (troeltz@math.tu-berlin.de).

$Su = Gu = y$ is a particular choice for example (i) below, where y solves the elliptic equation $\Delta y = u$ with homogeneous boundary conditions. Then $u \leq c + Gu$ is equivalent to the mixed control-state constraint $u - y \leq c$.

In all of these problems, the regularity of Lagrange multipliers associated with the state constraints is an important issue. The main aim of our paper is to show that, under natural assumptions, the multipliers can be assumed to be functions from $L^2(D)$.

Examples. The following optimal control problems for partial differential equations are covered by (PP).

(i) *Elliptic control problem with distributed control and distributed observation:*

$$\min \frac{1}{2} \|y - y_\Omega\|_{L^2(\Omega)}^2 + \frac{\kappa}{2} \|u\|_{L^2(\Omega)}^2$$

subject to

$$\begin{aligned} -\Delta y(x) &= u(x) & \text{in } \Omega, & & 0 \leq u(x) \leq b(x) & \text{in } \Omega, \\ y(x) &= 0 & \text{on } \Gamma, & & u(x) - y(x) \leq c(x) & \text{in } \Omega, \end{aligned}$$

where $\Omega \subset \mathbb{R}^N$, $N \geq 2$, is a bounded domain with $C^{0,1}$ -boundary Γ , and y_Ω is a fixed function from $L^2(\Omega)$. The function u is the *control*, while y is the associated *state*. The constraints $0 \leq u \leq b$ are called *box-constraints*, while the restriction $u - y \leq c$ is said to be a *mixed control-state constraint*.

It is known that the elliptic equation admits for each $u \in L^2(\Omega)$ a unique weak solution $y \in H_0^1(\Omega)$. The solution mapping $\hat{S} : u \mapsto y$ is continuous.

This problem is related to (PP) as follows: We define S as the solution operator, considered with range in $H = L^2(\Omega)$; that is, $S = E \hat{S}$, $S : L^2(\Omega) \rightarrow L^2(\Omega)$, where E denotes the embedding operator from $H_0^1(\Omega)$ to $L^2(\Omega)$.

Expanding the first norm square, we get $\frac{1}{2} \|y - y_\Omega\|_{L^2(\Omega)}^2 = \frac{1}{2} \|Su\|_{L^2(\Omega)}^2 - (S^*y_\Omega, u) + \text{const.}$; hence $a = -S^*y_\Omega$. The operator G is defined by $G = S$; i.e., the pointwise control-state constraints are considered in $L^2(D)$, and hence $D = \Omega$.

(ii) *Elliptic boundary control problem with distributed observation.* We consider a similar case of boundary control and introduce some more coefficients to make the setting more flexible for applications. Thus

$$\min \frac{1}{2} \|\alpha y - y_\Omega\|_{L^2(\Omega)}^2 + \frac{\kappa}{2} \|u\|_{L^2(\Gamma)}^2$$

subject to

$$\begin{aligned} -\Delta y(x) &= 0 & \text{in } \Omega, & & 0 \leq u(x) \leq b(x) & \text{on } \Gamma, \\ \partial_n y(x) + \beta(x) y(x) &= u(x) & \text{on } \Gamma, & & u(x) - \gamma(x) y(x) \leq c(x) & \text{on } \Gamma, \end{aligned}$$

where we use the same quantities as in (i), and $\alpha \in L^\infty(\Omega)$, $\beta, \gamma \in L^\infty(\Gamma)$ are fixed nonnegative functions. We assume $\beta \neq 0$ to guarantee the unique solvability of the elliptic boundary value problem. The solution mapping $\hat{S} : u \mapsto y$ is continuous from $L^2(\Gamma)$ to $H^1(\Omega)$. Let E be the embedding operator of $H^1(\Omega)$ into $L^2(\Omega)$. Now, $D = \Gamma$, $H = L^2(\Omega)$, $S = \alpha E \hat{S}$, $S : L^2(\Gamma) \rightarrow L^2(\Omega)$, and $G = \gamma \tau \hat{S} : L^2(\Gamma) \rightarrow L^2(\Gamma)$, where $\tau : H^1(\Omega) \rightarrow L^2(\Gamma)$ denotes the trace operator. By ∂_n we denote the normal derivative with respect to the outward normal vector.

(iii) *Parabolic boundary control problem.* Let us mention a further application:

$$\min \frac{1}{2} \|\alpha y(T) - y_\Omega\|_{L^2(\Omega)}^2 + \frac{\kappa}{2} \|u\|_{L^2(\Sigma)}^2$$

subject to

$$\begin{aligned} y_t(x, t) - \Delta y(x, t) &= 0 && \text{in } Q, \\ \partial_n y(x, t) + \beta(x, t) y(x, t) &= u(x, t) && \text{on } \Sigma, \\ y(x, 0) &= 0 && \text{in } \Omega, \end{aligned}$$

$$\begin{aligned} 0 \leq u(x, t) \leq b(x, t) &&& \text{on } \Sigma, \\ u(x, t) - \gamma(x, t) y(x, t) \leq c(x, t) &&& \text{on } \Sigma. \end{aligned}$$

Here, the notations $Q = \Omega \times (0, T)$, $\Sigma = \Gamma \times (0, T)$ are used. Moreover, functions $\alpha \in L^\infty(\Omega)$, $\beta \in L^\infty(\Gamma)$, $\gamma \geq 0 \in L^\infty(\Sigma)$ are given. It is known that the solution mapping $\hat{S} : u \mapsto y$ is continuous from $L^2(\Sigma)$ to

$$W(0, T) = \{y \in L^2(0, T; H^1(\Omega)) \mid y_t \in L^2(0, T; H^1(\Omega)')\}.$$

Therefore, the mapping $S : u \mapsto \alpha y(T)$ is well defined and continuous from $L^2(\Sigma)$ to $H = L^2(\Omega)$, and $G : u \mapsto \gamma \tau y$ is continuous in $L^2(\Sigma)$. The choice $D = \Sigma$ fits into our setting.

In all of these examples, the operator G is nonnegative; i.e., it transforms non-negative functions u into nonnegative functions. This follows from the maximum principle for elliptic and parabolic PDEs.

Problems with mixed pointwise control-state constraints of the type defined above have already been the subject of various papers devoted to *bottleneck problems*. In the 1960s, they were considered extensively for linear programming problems in Banach spaces in the context of the optimal control of ODEs; see [7], [15]. In [14], the author extended these ideas to the case of parabolic equations to derive necessary optimality conditions in the form of a minimum principle.

Later, more general linear and semilinear parabolic equations were discussed [1], [2], [3], and second-order sufficient optimality conditions were derived for the nonlinear parabolic case in [13]. Moreover, the Pontryagin maximum principle was proven for parabolic problems with mixed control-state constraints in [4].

In early papers on this subject, techniques of linear programming in Banach spaces were used. The Lagrange multiplier for the state constraints was obtained as the solution of a dual linear problem. The main difficulty was the proof of existence of solutions to the dual problem. This linear programming technique can also be extended to nonlinear problems by showing that an optimal control solves a linearized problem. Then the Lagrange multiplier can be obtained as a solution of the dual to the linearized program. This method was applied by Bergounioux and the current author in [2], [3]. Later, Arada and Raymond [1] introduced a more direct approach starting from the existence of Lagrange multipliers in the space $L^\infty(D)^*$. This existence is directly obtained from the Karush–Kuhn–Tucker theory in Banach spaces. Then the crucial step is to show higher regularity by a compactification technique, a quite general but rather technical approach.

In [2], [3] the inverse-monotonicity of the operator $(I - G)^{-1}$ is employed, which is only true if the associated Neumann series $I + G + G^2 + \dots$ converges. In the elliptic case that was the starting point of this paper, this convergence holds true only for a

“small” operator G . In the example (ii), this means that γ must have a sufficiently small L^∞ -norm. The reason for that is that eigenvalues come into play for elliptic equations. Therefore, the methods used in [2], [3] do not directly extend to the elliptic case, which has thus not yet been discussed in the literature.

The novelty of this paper is twofold. First, the detour on linear programming is avoided by direct use of quadratic dual programs. In this way, the Lagrange multipliers are obtained by a constructive approach that is fairly elementary. Second, and this is the main issue, an estimation technique is applied that avoids the use of inverse monotonicity. An assumption of the smallness of G is not needed.

Let us briefly explain at this point the main difficulty in proving the existence of regular Lagrange multipliers. Multipliers for the box-constraints $0 \leq u \leq b$ can be constructed in a simple way. Therefore, we concentrate for a while on the more complicated constraint $u - Gu - c \leq 0$. The box-constraints imply that u is bounded and measurable, and hence we might consider the mapping $u \mapsto u - Gu - c$ as one in $L^\infty(D)$. The cone of nonpositive functions of $L^\infty(D)$ has a nonempty interior, and hence we are justified in assuming the existence of a so-called *Slater point* $\tilde{u}$ such that $\tilde{u} - G\tilde{u} - c$ belongs to the interior of this cone; i.e.,

$$(1.1) \quad (\tilde{u} - G\tilde{u} - c)(x) \leq -\delta$$

holds a.e. on D with some $\delta > 0$. Then Lagrange multipliers are obtained in the dual space $L^\infty(D)^*$ by a standard result; see [8]. However, $L^\infty(D)^*$ is a space of measures. Therefore, this direct and easy way is not useful for finding regular Lagrange multipliers, and we should consider the mapping $u \mapsto u - Gu - c$ as one in $L^2(D)$. Here, the interior of the cone K of nonpositive functions is empty, and a (strong) Slater condition cannot be fulfilled. Nevertheless, let us keep in mind (1.1) as a “weak Slater condition.”

Slater conditions are also a standard assumption in the duality theory of convex programming in Hilbert spaces; see, e.g., [6], [8], [10], [11], [12]. To explain this, we write (PP) in the form

$$\min f(u), \quad u \in C, \quad u - Gu - c \leq 0,$$

and eliminate the constraint $u - Gu - c \leq 0$ by a Lagrange multiplier $\mu \in L^2(D)$. Defining the Lagrange function $\mathcal{L}$ by

$$\mathcal{L}(u, \mu) = f(u) + (u - Gu - c, \mu),$$

(PP) is seen to be equivalent to the min-max problem

$$\min_{u \in C} \sup_{\mu \geq 0} \mathcal{L}(u, \mu).$$

The associated dual problem (DP) is obtained by interchanging the order of min and sup, to find $\sup_{\mu \geq 0} \min_{u \in C} \mathcal{L}(u, \mu)$. It is easy to confirm that

$$(\tilde{u} - G\tilde{u} - c, \mu) \leq -\delta \|\mu\|_{L^1(D)}$$

holds for the (weak) Slater point $\tilde{u}$; hence

$$\lim_{\|\mu\|_{L^1(D)} \rightarrow \infty} \mathcal{L}(u, \mu) = -\infty$$

is satisfied. Now it is more or less clear that the search for a solution of (DP) can be restricted to a bounded set of $L^1(D)$. Unfortunately, $L^1(D)$ is not reflexive, and boundedness does not imply weak compactness. The existence of an optimal solution of the dual problem cannot be proven in this direct way.

Nevertheless, we shall employ the weak Slater condition (1.1) to draw on boundedness in $L^1(D)$. Next, after some selection and bootstrapping arguments, boundedness in $L^2(D)$ is obtained.

2. The dual problem. Let us first introduce the notion of Lagrange multipliers. We include all inequality constraints in the *Lagrange functional* that is redefined by

$$L(u, \mu, \lambda, \nu) = f(u) + (u - Gu - c, \mu) + (u - b, \lambda) - (u, \nu),$$

$L : L^2(D)^4 \rightarrow \mathbb{R}$. For convenience we denote the vector of Lagrange multipliers by $\omega := (\mu, \lambda, \nu)$ and write $\omega \geq 0$ whenever $\mu(x) \geq 0$, $\lambda(x) \geq 0$, and $\nu(x) \geq 0$ hold a.e. on D . Moreover, $\mu \geq 0$ means $\mu(x) \geq 0$ a.e. on D .

DEFINITION. Let $\bar{u}$ be a solution of the problem (PP). Functions $\bar{\mu}, \bar{\lambda}, \bar{\nu} \in L^2(D)$ are said to be *Lagrange multipliers associated with $\bar{u}$* if the following holds:

$$\frac{\partial L}{\partial u}(\bar{u}, \bar{\mu}, \bar{\lambda}, \bar{\nu}) = 0$$

and

$$\begin{aligned} (\bar{u} - G\bar{u} - c, \bar{\mu}) &= 0, & \bar{\mu} &\geq 0, \\ (\bar{u} - b, \bar{\lambda}) &= 0, & \bar{\lambda} &\geq 0, \\ (\bar{u}, \bar{\nu}) &= 0, & \bar{\nu} &\geq 0. \end{aligned}$$

The second block of relations forms the well-known *complementary slackness conditions*.

As pointed out in the introduction, we invoke the concept of Lagrange duality. It is easy to verify that the primal problem can be equivalently written in the form

$$(PP') \quad \inf_{u \in L^2(D)} \left\{ \sup_{\omega \in L^2(D)^3, \omega \geq 0} L(u, \omega) \right\}.$$

The *Lagrange dual problem* is defined by reversing the order of inf and sup:

$$(DP') \quad \sup_{\omega \in L^2(D)^3, \omega \geq 0} \left\{ \inf_{u \in L^2(D)} L(u, \omega) \right\}.$$

We shall prove that the Lagrange function has a saddle point, i.e., a pair $(\bar{u}, \bar{\omega})$ such that

$$(2.1) \quad L(\bar{u}, \omega) \leq L(\bar{u}, \bar{\omega}) \leq L(u, \bar{\omega})$$

holds for all $u \in L^2(D)$ and all $\omega \in L^2(D)^3$, $\omega \geq 0$. The property of $(\bar{u}, \bar{\omega})$ to be a saddle point is equivalent to the statement that $\bar{u}$ is a solution of (PP') (and hence of (PP)) and $\bar{\omega}$ solves (DP'). Let us denote the dual objective functional by g ,

$$g(\mu, \lambda, \nu) = g(\omega) = \inf_{u \in L^2(D)} L(u, \omega).$$

In this way, the dual problem is equivalent to

$$(DP) \quad \sup g(\omega), \quad \omega \geq 0, \quad \omega \in L^2(D)^3.$$

Denote the set of all feasible points of (PP) by $\mathcal{F} \subset L^2(D)$. For all $\omega \geq 0$ and all $u \in \mathcal{F}$, we obtain the following relation of *weak duality* between $g(\omega)$ and $f(u)$:

$$(2.2) \quad \begin{aligned} g(\omega) &= \inf_{v \in L^2(D)} \{f(v) + (v - Gv - c, \mu) + (v - b, \lambda) - (v, \nu)\} \\ &\leq \inf_{v \in \mathcal{F}} \{f(v) + (v - Gv - c, \mu) + (v - b, \lambda) - (v, \nu)\} \\ &\leq \inf_{v \in \mathcal{F}} f(v) \leq f(u). \end{aligned}$$

Next, we derive a more explicit expression for g .

LEMMA 2.1. *The dual objective function g is equal to*

$$g(\mu, \lambda, \nu) = -(c, \mu) - (b, \lambda) - \frac{1}{2} \|a + \mu - G^* \mu + \lambda - \nu\|_{\Lambda}^2,$$

where the norm $\|\cdot\|_{\Lambda}$ is defined by $\Lambda = \kappa I + S^* S$ and

$$\|d\|_{\Lambda}^2 = (d, \Lambda^{-1} d).$$

Proof. The result is obtained by straightforward computations. We get

$$(2.3) \quad \begin{aligned} g(\mu, \lambda, \nu) &= \min_{u \in L^2(D)} L(u, \mu, \lambda, \nu) = -(c, \mu) - (b, \lambda) \\ &\quad + \min_u \{f(u) + (u - Gu, \mu) + (u, \lambda) - (u, \nu)\}. \end{aligned}$$

The function in braces is convex and differentiable. It attains its unique minimum at $\hat{u} \in L^2(D)$. Notice that f is strictly convex by $\kappa > 0$. We find $\hat{u}$ by setting the derivative equal to zero,

$$\kappa \hat{u} + S^* S \hat{u} + a + \mu - G^* \mu + \lambda - \nu = 0;$$

hence

$$(2.4) \quad \hat{u} = -\Lambda^{-1}(a + \mu - G^* \mu + \lambda - \nu) = -\Lambda^{-1} d,$$

where

$$(2.5) \quad d = a + \mu - G^* \mu + \lambda - \nu.$$

Inserting $\hat{u}$ defined by (2.4) into $f(u) + (u - Gu, \mu) + (u, \lambda) - (u, \nu)$, the formula for g is found instantly,

$$\begin{aligned} f(\hat{u}) + (\hat{u} - G\hat{u}, \mu) + (\hat{u}, \lambda - \nu) &= \frac{1}{2} \|S\hat{u}\|_H^2 + \frac{\kappa}{2} \|\hat{u}\|^2 + (\hat{u}, a + \mu - G^* \mu + \lambda - \nu) \\ &= \frac{1}{2} (S\Lambda^{-1}d, S\Lambda^{-1}d)_H + \frac{\kappa}{2} (\Lambda^{-1}d, \Lambda^{-1}d) \\ &\quad - (d, \Lambda^{-1}d) \\ &= \left(\frac{1}{2} [\kappa \Lambda^{-1}d + S^* S \Lambda^{-1}d] - d, \Lambda^{-1}d \right) \\ &= \left(\frac{1}{2} \Lambda \Lambda^{-1}d - d, \Lambda^{-1}d \right) = -\frac{1}{2} (d, \Lambda^{-1}d) \\ &= -\frac{1}{2} \|d\|_{\Lambda}^2. \end{aligned}$$

Now the formula for g is obtained from (2.3). $\square$

LEMMA 2.2. *Assume the functions b and c to be nonnegative. Then the primal problem (PP) admits a unique optimal solution $\bar{u}$.*

Proof. The arguments are standard. The nonnegativity of b and c ensures that the feasible set of (PP) is not empty, since $u = 0$ satisfies all constraints. The objective functional f is strictly convex, continuous, and tends to infinity as $\|u\| \rightarrow \infty$. Therefore, the search for the minimum can be restricted to a convex, closed, and bounded set that is weakly sequentially compact. It is well known that under these assumptions f attains its minimum at a feasible $\bar{u}$. $\square$

Let us denote for convenience the negative objective functional $-g$ of (DP) by φ ,

$$\varphi(\omega) = (c, \mu) + (b, \lambda) + \frac{1}{2} \|a + \mu - G^* \mu + \lambda - \nu\|_\Lambda^2.$$

It is obvious that the dual problem (DP) is equivalent to minimizing φ .

Under natural assumptions, separation arguments apply to show that the optimal values of the primal and dual problems coincide; i.e., there is no duality gap. To prove this, we need the following convex set, an epigraph associated with the primal problem:

$$K := \{(r, z, v, w) \in \mathbb{R} \times L^2(D)^3 \mid \exists u \in L^2(D) : \\ f(u) \leq r, u - Gu - c \leq z, u - b \leq v, -u \leq w\}.$$

LEMMA 2.3. *The set K is convex and closed.*

Proof. Let $\{(r_n, z_n, v_n, w_n)\}$ be a sequence of elements of K converging to (r, z, v, w) . We have to show that $(r, z, v, w) \in K$. By the definition of K , functions u_n exist such that

$$f(u_n) \leq r_n, \quad u_n - Gu_n - c \leq z_n, \quad u_n - b \leq v_n, \quad -u_n \leq w_n.$$

Since $r_n \geq f(u_n) \geq \kappa \|u_n\|^2$ and $r_n \rightarrow r$, $\{u_n\}$ must be bounded in $L^2(D)$ so that we can select a weakly converging subsequence, say $u_n \rightharpoonup u$. Thanks to the lower semicontinuity of f , it holds that $f(u) \leq r$. Moreover, the set

$$\{(u, z, v, w) \in L^2(D)^4 \mid u - Gu - c \leq z, u - b \leq v, -u \leq w\}$$

is convex and closed, hence weakly closed. Therefore, the limit (u, z, v, w) belongs to this set. Together with the result above, $(r, z, v, w) \in K$ is shown. $\square$

THEOREM 2.4 (duality). *Assume that $c(x) > 0$ and $b(x) > 0$ hold a.e. on D . Then the duality relation $\inf (\text{PP}) = \sup (\text{DP})$; i.e.,*

$$(2.6) \quad f(\bar{u}) = \sup_{\omega \geq 0} g(\omega)$$

is satisfied.

Proof. Let $\bar{u}$ be the optimal solution of (PP). We take an arbitrary but fixed natural number n . Then the point $(f(\bar{u}) - 1/n, 0, 0, 0)$ does not belong to K . Notice that for each $u \in L^2(D)$ and all nonnegative r, z, v, w the element $(\tilde{r} = f(u) + r, \tilde{z} = u - Gu - c + z, \tilde{v} = u - b + v, \tilde{w} = -u + w)$ belongs to K . Since K is convex and closed, we can separate the point $(f(\bar{u}) - 1/n, 0, 0, 0)$ from K by a separating hyperplane; i.e., there is an element $(\sigma, \mu, \lambda, \nu) \neq 0$ such that

$$(2.7) \quad \begin{aligned} \sigma \left(f(\bar{u}) - \frac{1}{n} \right) + 0 &\leq \sigma (f(u) + r) + (\mu, u - Gu - c + z) \\ &\quad + (\lambda, u - b + v) + (\nu, -u + w) \end{aligned}$$

for all u and all $r \geq 0, z \geq 0, v \geq 0, w \geq 0$. It is obvious that this relation can hold only if σ, μ, λ , and ν are nonnegative. Next we show $\sigma \neq 0$. Assume the contrary. Since b and c are a.e. positive, the function $\tilde{u}(x) = 0.5 \min\{b(x), c(x)\}$ is a.e. positive and less than $\min\{b(x), c(x)\}$. By $G\tilde{u} \geq 0$ we have

$$(2.8) \quad \tilde{u}(x) - G\tilde{u}(x) - c(x) < 0 \quad \text{and} \quad 0 < \tilde{u}(x) < b(x) \quad \text{a.e. on } E.$$

Inserting $\tilde{u}$ together with $z = v = w = 0$ into (2.7), we would get, in view of the temporary assumption $\sigma = 0$,

$$0 \leq (\mu, \tilde{u} - G\tilde{u} - c) + (\lambda, \tilde{u} - b) - (\nu, \tilde{u}).$$

The right-hand side of this expression is strictly negative unless $\mu = \lambda = \nu = 0$, contradicting the property $(\sigma, \mu, \lambda, \nu) \neq 0$. Therefore we can assume $\sigma > 0$, without limitation of generality $\sigma = 1$ (otherwise, we divide the inequality by σ). Now we have

$$f(\bar{u}) - \frac{1}{n} \leq (f(u) + r) + (\mu, u - Gu - c + z) + (\lambda, u - b + v) + (\nu, -u + w)$$

for all suitable u, r, v, z, w . We insert $r = 0, z = v = w = 0$. Then

$$\begin{aligned} f(\bar{u}) - \frac{1}{n} &\leq -(\mu, c) - (\lambda, b) + (\mu, u - Gu) + (\lambda, u) - (\nu, u) + (a, u) \\ &\quad + \frac{1}{2} \|Su\|_H^2 + \frac{\kappa}{2} \|u\|^2 \\ &= -(\mu, c) - (\lambda, b) + (\mu - G^*\mu + a + \lambda - \nu, u) + \frac{1}{2} \|Su\|_H^2 + \frac{\kappa}{2} \|u\|^2. \end{aligned}$$

Next we insert the function $\hat{u} = -\Lambda^{-1}d$ defined in (2.4). Then

$$\begin{aligned} f(\bar{u}) - \frac{1}{n} &\leq -(\mu, c) - (\lambda, b) - (d, \Lambda^{-1}d) + \frac{\kappa}{2} \|\Lambda^{-1}d\|^2 + \frac{1}{2} \|S\Lambda^{-1}d\|_H^2 \\ &= -(\mu, c) - (\lambda, b) - (d, \Lambda^{-1}d) + \frac{\kappa}{2} (\Lambda^{-1}d, \Lambda^{-1}d) + \frac{1}{2} (S\Lambda^{-1}d, S\Lambda^{-1}d)_H \\ &= -(\mu, c) - (\lambda, b) - \frac{1}{2} (d, \Lambda^{-1}d) = g(\mu, \lambda, \nu) = g(\omega). \end{aligned}$$

In view of the relation (2.2) of weak duality, it holds that $g(\omega) \leq f(\bar{u})$. Since n was arbitrary, we have found an $\omega = \omega_n = (\mu_n, \lambda_n, \nu_n)$ such that

$$f(\bar{u}) - \frac{1}{n} \leq g(\omega_n) \leq f(\bar{u}).$$

The claim of the lemma is proven for $n \rightarrow \infty$. $\square$

Remark. Notice that in (2.8) a Slater-type condition is used that is even weaker than (1.1). It is worth mentioning that, for problems with inequality constraints defined in L^2 -spaces, weak Slater conditions of this type are sufficient for obtaining a Lagrange multiplier if a so-called Fritz–John-type theorem (Lagrange multiplier rule in nonqualified form) is known to hold true. Unfortunately, in the proof of Fritz–John-type theorems, the assumption cannot be avoided that the interior of the convex cone K of nonpositive elements is nonempty. This observation shows that the actual bottleneck in proving Lagrange multiplier theorems is the Fritz–John-type result.

The next results are standard. Their proofs are included for convenience only.

LEMMA 2.5. *Let $\bar{u}$ be the solution of (PP) and $\bar{\omega} = (\bar{\mu}, \bar{\lambda}, \bar{\nu}) \geq 0$ be a solution of the dual problem (DP). Then the complementary slackness conditions*

$$(\bar{u} - G\bar{u} - c, \bar{\mu}) = (\bar{u} - b, \bar{\lambda}) = (\bar{u}, \bar{\nu}) = 0$$

are satisfied, and the pair $\{\bar{u}, \bar{\omega}\}$ is a saddle point of the Lagrange functional L .

Proof. We know $f(\bar{u}) = g(\bar{\omega})$ from Theorem 2.4, and hence

$$\begin{aligned} f(\bar{u}) &= g(\bar{\omega}) = \min_{u \in L^2(D)} \{f(u) + (u - Gu - c, \bar{\mu}) + (u - b, \bar{\lambda}) - (u, \bar{\nu})\} \\ (2.9) \quad &\leq f(\bar{u}) + (\bar{u} - G\bar{u} - c, \bar{\mu}) + (\bar{u} - b, \bar{\lambda}) - (\bar{u}, \bar{\nu}) = L(\bar{u}, \bar{\omega}) \leq f(\bar{u}) \end{aligned}$$

since $\bar{u}$ is feasible and $\bar{\mu}, \bar{\lambda}, \bar{\nu}$ are nonnegative. Therefore,

$$(\bar{u} - G\bar{u} - c, \bar{\mu}) + (\bar{u} - b, \bar{\lambda}) - (\bar{u}, \bar{\nu}) = 0$$

must hold. All scalar products above are nonpositive, and hence they vanish altogether. Therefore, (2.9) yields

$$f(\bar{u}) = L(\bar{u}, \bar{\omega}) = g(\bar{\omega}) = \min_v L(v, \bar{\omega}) \leq L(u, \bar{\omega})$$

for all $u \in L^2(D)$. This is one half of the saddle point property. On the other hand, for $\omega \geq 0$,

$$L(\bar{u}, \omega) = f(\bar{u}) + (\bar{u} - G\bar{u} - c, \mu) + (\bar{u} - b, \lambda) - (\bar{u}, \nu) \leq f(\bar{u}) = L(\bar{u}, \bar{\omega})$$

because all scalar products are nonpositive for feasible $\bar{u}$. This is the second part of the proof. $\square$

THEOREM 2.6. *Suppose that a feasible function $\bar{u}$ is optimal for (PP), and that the dual problem (DP) admits a solution $\bar{\omega} = (\bar{\mu}, \bar{\lambda}, \bar{\nu})$. Then $\bar{\omega}$ is a triplet of Lagrange multipliers associated with $\bar{u}$.*

This is a simple consequence of the preceding lemma. From the saddle point property we know that $\bar{u}$ solves the problem $\min_u L(u, \bar{\omega})$. Therefore, the derivative of L with respect to u must vanish at $\bar{u}$. Together with the complementary slackness conditions, this shows that $\bar{\omega}$ is a vector of Lagrange multipliers.

3. Existence of regular Lagrange multipliers. In this section, we prove the main result of this paper, the solvability of the dual problem. By Theorem 2.6, this implies the existence of Lagrange multipliers in the space $L^2(D)$. The proof is based upon the following assumption.

ASSUMPTION OF NONNEGATIVITY AND BOUNDEDNESS. *The operator G is non-negative; i.e.,*

$$u(x) \geq 0 \text{ a.e. on } D \Rightarrow (Gu)(x) \geq 0 \text{ a.e. on } D.$$

Moreover, there are some $r > 0$ and a constant $c_r > 0$ such that G^* is bounded from $L^1(D)$ to $L^{1+r}(D)$ and from $L^2(D)$ to $L^{2+r}(D)$; i.e., the smoothing property

$$(3.1) \quad \|G^* \mu\|_{L^{i+r}(D)} \leq c_r \|\mu\|_{L^i(D)} \quad \forall \mu \in L^i(D), \quad i = 1, 2,$$

is satisfied.

We define

$$(3.2) \quad s = \min\{1 + r, 2\}.$$

Then G^* is bounded from $L^s(D)$ to $L^s(D)$. Moreover, we obtain by interpolation that

$$\|G^* \mu\|_{L^{p+r}(D)} \leq c_r \|\mu\|_{L^p(D)} \quad \forall \mu \in L^p(D), \forall p \in [1, 2].$$

The dual problem is equivalent to minimizing φ on the set of nonnegative functions of $L^2(D)^3$,

$$(3.3) \quad \inf_{\omega \geq 0} (c, \mu) + (b, \lambda) + \frac{1}{2} \|a + \mu - G^* \mu + \lambda - \nu\|_{\Lambda}^2.$$

Suppose that c and b are nonnegative. Then $\varphi(\omega) \geq 0$ holds for all $\omega \geq 0$, and the corresponding infimum j of φ exists as a nonnegative real number. Let $\{\omega_n\} = \{(\mu_n, \lambda_n, \nu_n)\}$ be a minimizing sequence, i.e.,

$$\lim_{n \rightarrow \infty} \varphi(\omega_n) = j.$$

Define d_n by

$$(3.4) \quad d_n = a + \mu_n - G^* \mu_n + \lambda_n - \nu_n.$$

LEMMA 3.1. *Assume that $c(x) \geq \delta_0$ and $b(x) \geq \delta_0$ holds a.e. on D with some constant $\delta_0 > 0$. Then a constant $M > 0$ exists such that the minimizing sequence $\omega_n = (\mu_n, \lambda_n, \nu_n)$ satisfies*

$$\|d_n\|_{L^2(D)} + \|\mu_n\|_{L^1(D)} + \|\lambda_n\|_{L^1(D)} \leq M$$

for all natural n .

Proof. Notice that $b(x) \geq \delta_0$, $c(x) \geq \delta_0$ implies

$$(c, \mu) \geq \delta_0 \|\mu\|_{L^1(D)}, \quad (b, \lambda) \geq \delta_0 \|\lambda\|_{L^1(D)}$$

for all nonnegative λ and μ . Hence, for all sufficiently large n ,

$$(3.5) \quad \delta_0 \|\mu_n\|_{L^1(D)} + \delta_0 \|\lambda_n\|_{L^1(D)} + \frac{1}{2} \|d_n\|_{\Lambda}^2 \leq j + 1$$

holds because $\{\omega_n\}$ is a minimizing sequence. Otherwise $\varphi(\omega_n)$ would tend to infinity. Therefore, the L^1 -norms of μ_n and λ_n are bounded. The operator $\Lambda = \kappa I + S^* S$ is positive definite, as

$$(e, \Lambda e) \geq \kappa \|e\|^2$$

holds for all $e \in L^2(D)$. This implies that, for all $d \in L^2(D)$,

$$(3.6) \quad \|d\|_{\Lambda}^2 = (d, \Lambda^{-1} d) = (\Lambda (\Lambda^{-1} d), \Lambda^{-1} d) \geq \kappa \|\Lambda^{-1} d\|^2 \geq \kappa c_{\Lambda}^{-2} \|d\|^2.$$

Here, we have used the boundedness of Λ , $\|\Lambda e\| \leq c_{\Lambda} \|e\|$, and hence $\|d\| \leq c_{\Lambda} \|\Lambda^{-1} d\|$.

In view of this, unboundedness of $\|d_n\|$ would imply unboundedness of $\|d_n\|_{\Lambda}$, contradicting (3.5). Therefore, $\|d_n\|$ is bounded, too. $\square$

Remark. It is easy to confirm that the assumption $c(x) \geq \delta_0$, $b(x) \geq \delta_0$ is equivalent to a weak Slater condition, since $\tilde{u}(x) = \varepsilon$ satisfies all constraints strictly if $\varepsilon > 0$ is sufficiently small.

LEMMA 3.2. *Let the assumptions of Lemma 3.1 be satisfied, and let $\omega = (\mu, \lambda, \nu) \geq 0$ be given. Then there is another $\tilde{\omega} = (\tilde{\mu}, \tilde{\lambda}, \tilde{\nu}) \geq 0$ such that*

$$(3.7) \quad \varphi(\tilde{\omega}) \leq \varphi(\omega)$$

and

$$(3.8) \quad \tilde{\mu}(x) + \tilde{\lambda}(x) + \tilde{\nu}(x) \leq 2(|d(x)| + |a(x)| + (G^*(\mu + \lambda))(x))$$

holds a.e. on D .

Proof. We define the measurable sets

$$\begin{aligned} I_0 &= \{x \in D \mid \nu(x) = 0\}, \\ I_+ &= \{x \in D \mid \nu(x) > 0\}. \end{aligned}$$

(i) *Update of (μ, λ) .* We update μ and λ by $\tilde{\mu} + \tilde{\lambda} = \mu + \lambda - \delta$ with a certain function $\delta = \delta(x)$ to be defined below.

On I_0 we have, by (2.5),

$$\mu(x) + \lambda(x) = d(x) - a(x) + (G^* \mu)(x).$$

The sum $\mu + \lambda$ is already bounded on I_0 by an expression of the type (3.8). There is no reason to update μ and λ on I_0 , and we define the update $\delta(x)$ by

$$\delta(x) = 0 \quad \text{on } I_0.$$

On I_+ we consider the two sets

$$\begin{aligned} I_+(\nu) &= \{x \in I_+ \mid \mu(x) + \lambda(x) \geq \nu(x)\}, \\ I_+(\mu) &= \{x \in I_+ \mid \mu(x) + \lambda(x) < \nu(x)\}, \end{aligned}$$

and set

$$\delta(x) = \begin{cases} \nu(x) & \text{for } x \in I_+(\nu), \\ \mu(x) + \lambda(x) & \text{for } x \in I_+(\mu). \end{cases}$$

Now, δ is defined on the whole set D , and we update μ and λ by

$$\tilde{\mu}(x) + \tilde{\lambda}(x) = \mu(x) + \lambda(x) - \delta(x),$$

while ensuring that $0 \leq \tilde{\mu}(x) \leq \mu(x)$ and $0 \leq \tilde{\lambda}(x) \leq \lambda(x)$ for all $x \in D$. This can be accomplished in many ways. To fix the setting, let us make the choice $\delta = \delta_\mu + \delta_\lambda$, where

$$\delta_\mu(x) = \begin{cases} 0 & \text{for } x \in I_0, \\ \nu(x)/2 & \text{for } x \in I_+(\nu), \\ \mu(x) & \text{for } x \in I_+(\mu), \end{cases} \quad \delta_\lambda(x) = \begin{cases} 0 & \text{for } x \in I_0, \\ \nu(x)/2 & \text{for } x \in I_+(\nu), \\ \lambda(x) & \text{for } x \in I_+(\mu), \end{cases}$$

and $\tilde{\mu} = \mu - \delta_\mu$, $\tilde{\lambda} = \lambda - \delta_\lambda$.

(ii) *Bound for $(\tilde{\mu}, \tilde{\lambda})$.* In view of our definitions, it holds that

$$(3.9) \quad \tilde{\mu}(x) + \tilde{\lambda}(x) = \begin{cases} \mu(x) + \lambda(x) = d(x) - a(x) + (G^* \mu)(x) & \text{on } I_0, \\ \mu(x) + \lambda(x) - \nu(x) = d(x) - a(x) + (G^* \mu)(x) & \text{on } I_+(\nu), \\ 0 & \text{on } I_+(\mu). \end{cases}$$

Hence (3.9) implies in turn that

$$(3.10) \quad \tilde{\mu}(x) + \tilde{\lambda}(x) \leq |d(x)| + |a(x)| + (G^*(\mu + \lambda))(x)$$

holds a.e. on D . Here we have used $G^* \mu \leq G^*(\mu + \lambda)$.

(iii) *Update of ν .* Next we adapt ν in a way that leaves $\|d\|_\Lambda$ unchanged. By (2.5), the following holds on D :

$$\mu + \lambda = d - a + G^* \mu + \nu.$$

Inserting $\mu + \lambda = \tilde{\mu} + \tilde{\lambda} + \delta$ and $\mu = \tilde{\mu} + \delta_\mu$, we find

$$\tilde{\mu} + \tilde{\lambda} = d - a + G^* \tilde{\mu} + G^* \delta_\mu + \nu - \delta.$$

Therefore, it is natural to define

$$(3.11) \quad \tilde{\nu} = G^* \delta_\mu + \nu - \delta.$$

Then $\tilde{\nu}$ is nonnegative on D , since $\nu - \delta \geq 0$, $\delta_\mu \geq 0$, and hence $G^* \delta_\mu \geq 0$. Moreover,

$$(3.12) \quad \tilde{\mu} + \tilde{\lambda} + a - G^* \tilde{\mu} - \tilde{\nu} = d$$

is satisfied everywhere on D .

On $I_0 \cup I_+(\nu)$, the relation $\nu(x) = \delta(x)$ holds true, and hence (3.11) yields

$$(3.13) \quad \tilde{\nu}(x) = (G^* \delta_\mu)(x) \quad \forall x \in I_0 \cup I_+(\nu).$$

Moreover, we know that $\tilde{\mu}(x) + \tilde{\lambda}(x) = 0$ on $I_+(\mu)$ and obtain from (3.12) that

$$\tilde{\nu}(x) = a(x) - d(x) - (G^* \tilde{\mu})(x) \quad \text{on } I_+(\mu).$$

Summarizing, $\tilde{\nu}$ is defined by

$$(3.14) \quad \tilde{\nu}(x) = \begin{cases} (G^* \delta_\mu)(x) \leq (G^* \delta)(x) \leq (G^*(\mu + \lambda))(x) & \text{on } I_+(\nu) \cup I_0, \\ a(x) - d(x) - (G^* \tilde{\mu})(x) & \text{on } I_+(\mu), \end{cases}$$

implying in turn

$$\tilde{\nu}(x) \leq |d(x)| + |a(x)| + (G^*(\mu + \lambda))(x)$$

a.e. on D , as $0 \leq \tilde{\mu} \leq \mu + \lambda$. Together with (3.10), this shows that $\tilde{\omega}$ satisfies (3.8).

(iv) *Verification of (3.7).* By definition, we have

$$\begin{aligned} 0 &\leq \tilde{\mu}(x) = \mu(x) - \delta_\mu(x) \leq \mu(x), \\ 0 &\leq \tilde{\lambda}(x) = \lambda(x) - \delta_\lambda(x) \leq \lambda(x), \end{aligned}$$

and hence

$$(c, \tilde{\mu}) \leq (c, \mu) \quad \text{and} \quad (b, \tilde{\lambda}) \leq (b, \lambda)$$

are satisfied. Moreover, our construction guarantees by (3.12) that

$$\tilde{\mu} + \tilde{\lambda} + a - G^* \tilde{\mu} - \tilde{\nu} = d = \mu + \lambda + a - G^* \mu - \nu.$$

This yields

$$\begin{aligned} \varphi(\tilde{\omega}) &= (c, \tilde{\mu}) + (b, \tilde{\lambda}) + \frac{1}{2} \|\tilde{\mu} + \tilde{\lambda} + a - G^* \tilde{\mu} - \tilde{\nu}\|_\Lambda^2 \\ &= (c, \tilde{\mu}) + (b, \tilde{\lambda}) + \frac{1}{2} \|\mu + \lambda + a - G^* \mu - \nu\|_\Lambda^2 \\ &\leq (c, \mu) + (b, \lambda) + \frac{1}{2} \|\mu + \lambda + a - G^* \mu - \nu\|_\Lambda^2 = \varphi(\omega). \quad \square \end{aligned}$$

THEOREM 3.3. *Assume that G satisfies the assumption of nonnegativity and boundedness and that b and c belong to $L^{s'}(D)$, where s is defined in (3.2) and s' is its conjugate exponent obtained from $1/s + 1/s' = 1$. Assume further that $b(x) \geq \delta_0$ and $c(x) \geq \delta_0$ are fulfilled for almost all $x \in D$ with some constant $\delta_0 > 0$. Then the dual problem admits at least one solution.*

Proof. Let ω_n be the minimizing sequence used in Lemma 3.1. By this lemma, the sequence $d_n = a + \mu_n - G^* \mu_n + \lambda_n - \nu_n$ is bounded in $L^2(D)$, and we can assume w.l.o.g. that $d_n \rightharpoonup d$ in $L^2(D)$ to some $d \in L^2(D)$. In particular, this sequence is weakly converging in $L^s(D)$. Moreover, we obtain from Lemma 3.1 the boundedness of $\{\mu_n\}$ in $L^1(D)$. The assumptions imposed on G^* imply that the sequence $\{G^* \mu_n\}$ is bounded in $L^s(D)$.

Thanks to Lemma 3.2, with the right-hand side of (3.8), the associated sequence $\tilde{\omega}_n$ is bounded in $L^s(D)$ too. Therefore, we can assume w.l.o.g. that $\tilde{\omega}_n \rightharpoonup \omega = (\mu, \lambda, \nu)$ in $L^s(D)$. The operator G^* is bounded in the L^s -norm, and hence G is bounded in $L^{s'}(D)$, and we get the weak convergence of $\{G^* \tilde{\mu}_n\}$ to $\{G^* \mu\}$ in $L^s(D)$. Altogether, we have

$$d_n = a + \tilde{\mu}_n - G^* \tilde{\mu}_n + \tilde{\lambda}_n - \tilde{\nu}_n,$$

where all sequences converge weakly in $L^s(D)$. Passing to the limit, we find

$$d = a + \mu - G^* \mu + \lambda - \nu.$$

It is obvious that

$$(c, \tilde{\mu}_n) \rightarrow (c, \mu) \quad \text{and} \quad (b, \tilde{\lambda}_n) \rightarrow (b, \lambda)$$

as $n \rightarrow \infty$, because b and c are assumed to be in $L^{s'}(D)$. Moreover, the functional $\|\cdot\|_\Lambda^2$ is continuous and convex in $L^2(D)$, and hence weakly lower semicontinuous. This yields

$$\begin{aligned} j &= \lim_{n \rightarrow \infty} \varphi(\omega_n) \geq \lim_{n \rightarrow \infty} \varphi(\tilde{\omega}_n) = \lim_{n \rightarrow \infty} \{(c, \tilde{\mu}_n) + (b, \tilde{\lambda}_n) + \|d_n\|_\Lambda^2\} \\ &= (c, \mu) + (b, \lambda) + \lim_{n \rightarrow \infty} \|d_n\|_\Lambda^2 \\ &\geq (c, \mu) + (b, \lambda) + \|d\|_\Lambda^2 \\ &= (c, \mu) + (b, \lambda) + \|a + \mu - G^* \mu + \lambda - \nu\|_\Lambda^2 = \varphi(\omega), \end{aligned}$$

and thus $\varphi(\omega) = j$.

It remains to show that the solution ω belongs to $L^2(D)$. This is done by a bootstrapping argument. So far, we have $\omega \in L^s(D)$ and

$$0 \leq \mu(x) + \lambda(x) + \nu(x) \leq 2(|d(x)| + |a(x)| + (G^*(\mu + \lambda))(x)).$$

We know that d and a belong to $L^2(D)$, and we obtain $G^*(\mu + \lambda) \in L^{s+r}(D)$ by the smoothing property (3.1) of G^* . In this way, the inequality above delivers the regularity $\omega \in L^{\min\{s+r, 2\}}(D)$. Repeating this argument, in the next step $\omega \in L^{\min\{s+2r, 2\}}(D)$ is obtained. After finitely many steps we arrive at the regularity $\omega \in L^2(D)$. $\square$

4. Other types of constraints. Let us mention a few more interesting cases, where the theory applies after minor changes.

4.1. Single control-state constraint. Consider the optimization problem for $u \in L^2(D)$,

$$(P1) \quad \min f(u), \quad u(x) \leq c(x) + (Gu)(x) \quad \text{a.e. on } D,$$

without any box-constraint on u . Existence and uniqueness of an optimal solution $\bar{u}$ follows from the strict convexity of f together with the convexity and closedness of the feasible set $\{u \mid u \leq c + Gu\}$. The existence of an associated regular Lagrange multiplier $\bar{\mu}$ can be proven in a direct and rather trivial way. We assume that $I - G$ has a continuous inverse operator in $L^2(D)$; in the applications to PDEs, G is compact, and hence this amounts to the assumption that $\alpha = 1$ is not an eigenvalue of G .

We substitute $v = u - Gu$ and have $u = (I - G)^{-1}v = Rv$ if we set $R = (I - G)^{-1}$. Define $\tilde{f}(v) = f(Rv)$. Then problem (P1) is equivalent to

$$(\tilde{P1}) \quad \min \tilde{f}(v), \quad v(x) \leq c(x) \quad \text{a.e. in } D, \quad v \in L^2(D),$$

where $\tilde{f}(v) = f(Rv)$. This is a problem without state constraints. For the optimal solution $\bar{v} = (I - G)\bar{u}$ we obtain by standard methods the variational inequality

$$(4.1) \quad \tilde{f}'(\bar{v})(v - \bar{v}) \geq 0 \quad \forall v \leq c.$$

By the Riesz representation theorem, the gradient $f'(\bar{v})$ can be identified with a function of $L^2(D)$. Then (4.1) implies a.e. on D

$$\tilde{f}'(\bar{v})(x) \begin{cases} = 0 & \text{where } \bar{v}(x) < c(x), \\ \leq 0 & \text{where } \bar{v}(x) = c(x). \end{cases}$$

We just define the multiplier $\bar{\mu}$ by

$$(4.2) \quad \bar{\mu}(x) := -\tilde{f}'(\bar{v})(x) = (-R^* f'(\bar{u}))(x).$$

This definition assures $\bar{\mu} \geq 0$,

$$(4.3) \quad \tilde{f}'(\bar{v}) + \bar{\mu} = 0,$$

and

$$(4.4) \quad (\bar{v} - c, \bar{\mu}) = 0.$$

THEOREM 4.1. Assume that $(I - G)^{-1}$ exists and is continuous in $L^2(D)$. Then the function $\bar{\mu}$ defined in (4.2) is a Lagrange multiplier associated with the optimal solution $\bar{u}$ of (P1).

Proof. The function $\bar{u}$ is optimal for (P1) if and only if $\bar{v} = \bar{u} - G\bar{u}$ is optimal for $(\tilde{P1})$. By definition, $\bar{\mu}$ is nonnegative. Moreover, inserting the concrete expression for $\bar{v}$, (4.4) yields

$$(\bar{u} - G\bar{u} - c, \bar{\mu}) = 0,$$

and the complementary slackness condition is satisfied. We have

$$\tilde{f}'(\bar{v}) = R^* f'(R\bar{v}) = R^* f'(\bar{u})$$

and $L(u, \mu) = f(u) + (u - Gu - c, \mu)$. Therefore, $\bar{\mu}$ is a Lagrange multiplier, provided that $\partial L / \partial u = 0$ also holds, i.e.,

$$(4.5) \quad f'(\bar{u}) + (I - G^*)\mu = 0.$$

In view of $\tilde{f}'(\bar{v}) = R^* f'(\bar{u})$, equation (4.3) implies

$$R^* f'(\bar{u}) + \bar{\mu} = 0,$$

which is equivalent to (4.5). $\square$

4.2. The case $b = \infty$. Consider the constraints

$$0 \leq u(x) \leq c(x) + (Gu)(x).$$

Here, the upper bound b is not imposed on u . This case fits into our setting by the formal choice $b = \infty$. The reader will easily verify that Theorem 3.3 remains true with $\bar{\lambda} = 0$. The assumption $b(x) \geq \delta_0 > 0$ is not needed; it is formally satisfied.

4.3. Arbitrary box-constraints. Let the more general box-constraints

$$a(x) \leq u(x) \leq b(x), \quad u(x) \leq c(x) + (Gu)(x)$$

be given. In this case, we set $v = u - a$. Then the constraints transform to

$$0 \leq v \leq b - a, \quad v \leq c - a + Ga + Gv.$$

The theory applies after introducing $\tilde{c} = c - a + Ga$. Here, the strict positivity of c required in Theorem 3.3 means $a + \delta_0 \leq c + Ga$. In other words, the lower bound a must strictly satisfy the state constraints. Then a triplet of Lagrange multipliers exists by Theorem 3.3.

4.4. Different orientation of constraints. Another interesting type of constraints has the form

$$a(x) \leq u(x) \leq b(x), \quad c(x) + (Gu)(x) \leq u(x).$$

These constraints can be transformed to our initial setting as well. We set $v = b - u$. Then the new constraints are

$$0 \leq v \leq b - a, \quad v \leq b - c - Gb + Gv.$$

We substitute $\tilde{c} = b - c - Gb$, and the strict positivity of $\tilde{c}$ amounts to the requirement that the upper bound b strictly fulfill the state constraint. If this assumption is satisfied, then Theorem 3.3 yields the existence of regular Lagrange multipliers.

4.5. Negative sign with the control. For several reasons, the discussion of the constraints

$$u \geq 0, \quad -u \leq c + Gu$$

is useful too. Here, it is useless to require $c \geq 0$, because then the inequality $-u \leq 0 \leq c + Gu$ is automatically satisfied for all nonnegative u , and the mixed control-state constraint would be without effect.

We have in mind an application to pointwise state constraints $y(x) \geq c(x)$ for all $x \in D$, i.e., $-c \leq Gu$ in our notation. The trick is to add a regularization term,

$$-\varepsilon u \leq c + Gu,$$

and to consider the limit $\varepsilon \downarrow 0$. Dividing by ε , we get constraints $-u \leq 0 \leq \tilde{c} + \tilde{G}u$ of the type defined above.

The negative sign in front of the control cannot be bypassed by a transformation to the standard case. However, this case turns out to be even simpler. The proof of existence of $\bar{\mu}$ in $L^2(D)$ is rather easy. The only point at which the negative sign influences the theory appears in (3.3). We have to show that the minimization in (3.3)

admits a solution. To this aim, let us consider the associated problem (3.3). Now, it reads

$$\min_{\mu \geq 0, \nu \geq 0} \varphi(\mu, \nu), \quad \varphi(\mu, \nu) = (c, \mu) + \frac{1}{2} \|a - \mu - G^* \mu - \nu\|_\Lambda^2.$$

We show that this problem admits a solution. Let $\{(\mu_n, \nu_n)\}$ be an associated minimizing sequence. Then $\|a - \mu_n - G^* \mu_n - \nu_n\|$ must be bounded, since otherwise $\|a - \mu_n - G^* \mu_n - \nu_n\|_\Lambda$ would tend to infinity and, for n sufficiently large,

$$\varphi(\mu_n, \nu_n) \geq \varphi(0, 0) + 1 = \frac{1}{2} \|a\|_\Lambda^2 + 1$$

would hold. Therefore, $\|\mu_n + G^* \mu_n + \nu_n\|$ is bounded as well. All functions under this norm are nonnegative, and hence $\|\mu_n\| + \|\nu_n\|$ are bounded, too. This is the decisive point, where we can select weakly converging subsequences w.l.o.g., $\mu_n \rightarrow \bar{\mu}$ and $\nu_n \rightarrow \bar{\nu}$. The optimality of $\bar{\mu}, \bar{\nu}$ follows by lower semicontinuity of the objective functional f . The remaining part of the theory is along the lines of the preceding section. In this way, we have proven the following result.

THEOREM 4.2. *Assume that G satisfies the assumption of nonnegativity and boundedness. Then the problem*

$$(P2) \quad \min f(u), \quad u \geq 0, \quad -u \leq c + Gu,$$

has a unique solution $\bar{u}$. Moreover, associated regular Lagrange multipliers $\bar{\mu}, \bar{\nu}$ exist in $L^2(D)$.

The constraints $u \leq 0$, $Gu + c \leq -u$ can be transformed to the form of the constraints in (P2) by substituting $v = -u$.

5. Application to the examples (i)–(iii).

5.1. Elliptic distributed problem (i). In this problem, $S = E \hat{S}$ was the solution operator for the elliptic boundary value problem, $G = S$, $a = -S^* y_\Omega$, considered with range $L^2(\Omega)$. Moreover, it is known that the operator S^* is given by $S^* z = p$, where $p \in H_0^1(\Omega)$ solves the adjoint boundary value problem

$$\begin{aligned} -\Delta p &= z, \\ p|_\Gamma &= 0. \end{aligned}$$

From $\partial L / \partial u = 0$ it follows that

$$(5.1) \quad S^*(S\bar{u} - y_\Omega - \bar{\mu}) + \kappa \bar{u} + \bar{\mu} + \bar{\lambda} - \bar{\nu} = 0.$$

Consequently, we introduce an adjoint state p by $p = S^*(\bar{y} - y_\Omega - \bar{\mu})$,

$$\begin{aligned} -\Delta p &= \bar{y} - y_\Omega - \bar{\mu}, \\ p|_\Gamma &= 0. \end{aligned}$$

Then (5.1) reads $p + \kappa \bar{u} + \bar{\mu} + \bar{\lambda} - \bar{\nu} = 0$. The control $u = \bar{u}$ with state $y = \bar{y} = S\bar{u}$ is optimal if and only if the optimality system

$$\begin{aligned} -\Delta y &= u, & -\Delta p &= y - y_\Omega - \mu, \\ y|_\Gamma &= 0, & p|_\Gamma &= 0, \end{aligned}$$

$$(p + \kappa u + \mu + \lambda - \nu)(x) = 0,$$

$$\begin{aligned} u(x) &\geq 0, & \nu(x) &\geq 0, & u(x)\nu(x) &= 0, \\ u(x) &\leq b(x), & \lambda(x) &\geq 0, & (u(x) - b(x))\lambda(x) &= 0, \\ u(x) - c(x) - y(x) &\leq 0, & \mu(x) &\geq 0, & (u(x) - c(x) - y(x))\mu(x) &= 0 \end{aligned}$$

is satisfied for almost all $x \in \Omega$ with the corresponding adjoint state p and associated Lagrange multipliers μ, λ, ν from $L^2(\Omega)$. In the system above, $(\cdot, \cdot)$ stands for the inner product of $L^2(\Omega)$.

The existence of the Lagrange multipliers follows from Theorem 3.3, because the operator G satisfies all assumptions. Its nonnegativity follows from the maximum principle for elliptic equations. By the definition of weak solutions (cf. example (ii) below) it is easy to see that G is self-adjoint. Therefore, by elliptic regularity, $G^* = G$ is bounded from $L^1(\Omega)$ to $L^s(\Omega)$ for all $s < N/(N-2)$ [5] and from $L^2(\Omega)$ to $H_0^1(\Omega) \subset L^s(\Omega)$ for all $s < 2N/(N-2)$. This assures the boundedness property.

Remark. The same system can formally be obtained by the Lagrange function as follows. We append all inequality constraints for u by Lagrange multipliers and consider the following elliptic optimal control problem without inequality constraints.

Minimize with respect to (y, u) , for fixed μ, λ, ν ,

$$\begin{aligned} L(y, u, \mu, \lambda, \nu) = & \frac{1}{2} \|y - y_\Omega\|_{L^2(\Omega)}^2 + \frac{\kappa}{2} \|u\|_{L^2(\Omega)}^2 \\ & + (u - y - c, \mu) + (u - b, \lambda) - (u, \nu) \end{aligned}$$

subject to

$$\begin{aligned} -\Delta y &= u, \\ y|_\Gamma &= 0. \end{aligned}$$

In other words, we formally eliminate the pointwise constraints by Lagrange multipliers while keeping the PDE as an explicit constraint. Then we establish the standard first-order necessary optimality conditions for this elliptic control problem. In this way, we arrive at the optimality system established above. We illustrate this technique below.

5.2. Elliptic boundary control problem (ii). Let us establish the optimality system in the formal way explained above. For given Lagrange multipliers μ, λ, ν , we consider the following elliptic optimal control problem for the unknowns (y, u) .

Minimize

$$\begin{aligned} J(y, u) = L(y, u, \mu, \lambda, \nu) = & \frac{1}{2} \|\alpha y - y_\Omega\|_{L^2(\Omega)}^2 + \frac{\kappa}{2} \|u\|_{L^2(\Omega)}^2 \\ & + (u - \gamma y - c, \mu) + (u - b, \lambda) - (u, \nu) \end{aligned}$$

subject to

$$\begin{aligned} -\Delta y &= 0, \\ \partial_n y + \beta y &= u. \end{aligned}$$

In this problem, $(\cdot, \cdot)$ stands for the inner product of $L^2(\Gamma)$. We know that the adjoint state is given by the equations $-\Delta p = D_y L|_\Omega$ and $\partial_n p + \beta p = D_y L|_\Gamma$, and hence the adjoint equation is

$$\begin{aligned} -\Delta p &= \alpha(\alpha y - y_\Omega), \\ \partial_n p + \beta p &= -\gamma \mu. \end{aligned}$$

Moreover, $D_u L = 0$ must hold; i.e.,

$$p + \kappa u + \mu + \lambda - \nu = 0.$$

The existence of regular Lagrange multipliers in $L^2(\Gamma)$ follows from Theorem 3.3. The associated assumptions are satisfied: Again, $G \geq 0$ follows from the maximum principle. The adjoint operator $G^* : L^2(\Gamma) \rightarrow L^2(\Gamma)$ is given by $G^*v = \tau z$, where $z \in H^1(\Omega)$ is the weak solution of $-\Delta z = 0$, $\partial_n z + \beta z = \gamma v$. This follows easily by the definition of weak solutions: The state $y = Su$ solves with test function z

$$\int_{\Omega} \nabla y \cdot \nabla z \, dx + \int_{\Gamma} \beta y z \, d\Gamma = \int_{\Gamma} u z \, d\Gamma,$$

while z satisfies with test function y

$$\int_{\Omega} \nabla y \cdot \nabla z \, dx + \int_{\Gamma} \beta y z \, d\Gamma = \int_{\Gamma} \gamma v y \, d\Gamma.$$

The left-hand sides coincide; hence the right-hand sides must be equal, too. We obtain

$$(Gu, v)_{L^2(\Gamma)} = (\gamma y, v)_{L^2(\Gamma)} = (u, z)_{L^2(\Gamma)} = (u, G^*v)_{L^2(\Gamma)}.$$

By the regularity theory for elliptic equations, G^* transforms $L^1(\Gamma)$ into $L^s(\Gamma)$ for all $s < (N-1)/(N-2)$ [5] and $L^2(\Gamma)$ to $H^{1/2}(\Gamma) \subset L^s(\Gamma)$ for all $s < 2(N-1)/(N-2)$. Again, the boundedness property is easy to verify.

The optimality system reads

$$\begin{aligned} -\Delta y &= 0, & -\Delta p &= \alpha(\alpha y - y_{\Omega}), \\ \partial_n y + \beta y &= u, & \partial_n p + \beta p &= -\gamma \mu, \\ (p + \kappa u + \mu + \lambda - \nu)(x) &= 0, \\ u(x) &\geq 0, \quad \nu(x) \geq 0, \quad u(x)\nu(x) = 0, \\ u(x) &\leq b(x), \quad \lambda(x) \geq 0, \quad (u(x) - b(x))\lambda(x) = 0, \\ u(x) - c(x) - \gamma(x)y(x) &\leq 0, \quad \mu(x) \geq 0, \\ (u(x) - c(x) - \gamma(x)y(x))\mu(x) &= 0 \end{aligned}$$

for almost all $x \in \Gamma$.

5.3. Parabolic control problem (iii). The parabolic example is discussed analogously. Theorem 3.3 assures the existence of regular Lagrange multipliers. The associated assumptions are met. Also here, the nonnegativity of G follows from the maximum principle. The adjoint operator G^* is represented by $G^*v = \tau z$, where z solves

$$-z_t - \Delta z = 0, \quad \partial_n z + \beta z = \gamma v, \quad z(\cdot, T) = 0.$$

This follows by standard techniques for computing adjoints. Therefore, the parabolic regularity theory yields that G^* is bounded from $L^1(\Sigma)$ to $L^s(\Sigma)$ for all $s < (N+1)/N$ [9, Theorem 4.2] and from $L^2(\Sigma)$ to $L^s(\Sigma)$ for all $s < 2(N+1)/(N-1)$ [9, Theorem 4.1]. The boundedness property is thus satisfied.

Applying the formal Lagrange technique, the optimality system

$$\begin{aligned} y_t - \Delta y &= 0, & -p_t - \Delta p &= 0, \\ \partial_n y + \beta y &= u, & \partial_n p + \beta p &= -\gamma \mu, \\ y(\cdot, 0) &= 0, & \alpha(\alpha y(T) - y_\Omega), \end{aligned}$$

$$(p + \kappa u + \mu + \lambda - \nu)(x, t) = 0,$$

$$\begin{aligned} u(x, t) &\geq 0, \quad \nu(x, t) \geq 0, \quad u(x, t)\nu(x, t) = 0, \\ u(x, t) &\leq b(x, t), \quad \lambda(x, t) \geq 0, \quad (u(x, t) - b(x, t))\lambda(x, t) = 0, \end{aligned}$$

$$\begin{aligned} u(x, t) - c(x, t) - \gamma(x, t)y(x, t) &\leq 0, \quad \mu(x, t) \geq 0, \\ (u(x, t) - c(x, t) - \gamma(x, t)y(x, t))\mu(x, t) &= 0 \end{aligned}$$

must be satisfied for almost all $(x, t) \in \Sigma = \Gamma \times (0, T)$.

REFERENCES

- [1] N. ARADA AND J. P. RAYMOND, *Optimal control problems with mixed control-state constraints*, SIAM J. Control Optim., 39 (2000), pp. 1391–1407.
- [2] M. BERGOUNIOUX AND F. TRÖLTZSCH, *Optimal control of linear bottleneck problems*, ESAIM Control Optim. Calc. Var., 3 (1998), pp. 235–250.
- [3] M. BERGOUNIOUX AND F. TRÖLTZSCH, *Optimal control of semilinear parabolic equations with state-constraints of bottleneck type*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 595–608.
- [4] E. CASAS, J.-P. RAYMOND, AND H. ZIDANI, *Pontryagin's principle for local solutions of control problems with mixed control-state constraints*, SIAM J. Control Optim., 39 (2000), pp. 1182–1203.
- [5] E. CASAS, F. TRÖLTZSCH, AND A. UNGER, *Second order sufficient optimality conditions for some state-constrained control problems of semilinear elliptic equations*, SIAM J. Control Optim., 38 (2000), pp. 1369–1391.
- [6] I. EKELAND AND R. TEMAM, *Analyse Convexe et Problèmes Variationnels*, Dunod, Gauthier-Villars, Paris, 1974.
- [7] N. LEVINSON, *A class of continuous linear programming problems*, J. Math. Anal. Appl., 16 (1966), pp. 73–83.
- [8] D. G. LUENBERGER, *Optimization by Vector Space Methods*, Wiley, New York, 1969.
- [9] J.-P. RAYMOND AND F. TRÖLTZSCH, *Second order sufficient optimality conditions for nonlinear parabolic control problems with state constraints*, Discrete Contin. Dynam. Systems, 6 (2000), pp. 431–450.
- [10] R. T. ROCKAFELLAR, *Extension of Fenchel's duality theorem for convex functions*, Duke Math. J., 33 (1966), pp. 81–90.
- [11] R. T. ROCKAFELLAR, *Duality and stability in extremum problems involving convex functions*, Pacific J. Math., 21 (1967), pp. 167–187.
- [12] R. T. ROCKAFELLAR, *Convex functions and duality in optimization problems and dynamics*, in Mathematical Systems Theory and Economics, Vol. II, Lect. Notes in Oper. Res. and Math. Econ. 12, Springer, Berlin, 1969, pp. 117–141.
- [13] A. RÖSCH AND F. TRÖLTZSCH, *Sufficient second-order optimality conditions for a parabolic optimal control problem with pointwise control-state constraints*, SIAM J. Control Optim., 42 (2003), pp. 138–154.
- [14] F. TRÖLTZSCH, *A minimum principle and a generalized bang-bang principle for a distributed optimal control problem with constraints on the control and the state*, ZAMM Z. Angew. Math. Mech., 59 (1979), pp. 737–739.
- [15] W. F. TYNDALL, *A duality theorem for a class of continuous linear programming problems*, J. SIAM, 13 (1965), pp. 644–666.

CONVERGENCE CONDITIONS AND KRYLOV SUBSPACE-BASED CORRECTIONS FOR PRIMAL-DUAL INTERIOR-POINT METHOD*

SANJAY MEHROTRA[†] AND ZHIFENG LI[†]

Abstract. We present convergence conditions for a generic primal-dual interior-point algorithm with multiple corrector directions. The corrector directions can be generated by any approach. The search direction is obtained by combining predictor and corrector directions through a small linear program. We also propose a new approach to generate corrector directions. This approach generates directions using information from an appropriately defined Krylov subspace. We propose efficient implementation strategies for our approach that follow the analysis of this paper. Numerical experiments illustrating the features of the proposed approach and its practical usefulness are reported.

Key words. linear program, primal-dual interior-point method, inexact search direction, Krylov subspace, convergence

AMS subject classifications. 90C05, 90C06, 90C30, 90C51

DOI. 10.1137/S1052623403431494

1. Introduction. Implementations based on primal-dual interior-point methods have emerged as an efficient approach for solving large scale linear and nonlinear programs. The practical performance of these methods is improved significantly by introducing “corrector” directions to a Newton direction [10, 12, 16, 18]. The corrector directions are computed from an already factored matrix to reduce overall matrix factorizations in the algorithm. The predictor-corrector scheme builds on the philosophy that once an expensive factorization step is performed, it should be used to the best possible extent before doing a refactorization. Methods built on this scheme were proposed by Karmarkar et al. [15]; Mehrotra [18]; Gondzio [10]; and Jarre and Wechs [12]. The method in Karmarkar et al. [15] is based on a power-series expansion of a particular parameterization of a primal-dual trajectory toward a point on the central path. Mehrotra [17, 18] explicitly introduced the predictor-corrector concepts and proposed strategies based on adaptively generating information from fixed point iterations and/or a higher order Taylor expansion. Gondzio [10] proposed improvements based on centrality corrections, and more recently Jarre and Wechs [12] have given improved heuristics for generating the corrector steps. Jarre and Wechs [12] also proved that their approach of generating and combining corrector directions gives a convergent algorithm.

Since there can be many different ways to generate the corrector directions, it is important to consider conditions that ensure the convergence of a predictor-corrector interior-point algorithm which allows generic computations of corrector directions. This is addressed in the first part of this paper. We give these conditions in the form of a linear program. The conditions are derived from a refinement of the analysis appearing in Kojima, Megiddo, and Mizuno [14] and Jarre and Wechs [12].

We also propose a new approach for generating corrector directions. The idea is to incorporate information generated from an iterative scheme to improve the

*Received by the editors July 9, 2003; accepted for publication (in revised form) March 20, 2004; published electronically April 8, 2005. The work of both authors was supported in part by NSF grants DMI-9908038 and DMI-0200151 and ONR grant N00014-01-1-0048/P00002.

<http://www.siam.org/journals/siopt/15-3/43149.html>

[†]Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208 (mehrotra@iems.nwu.edu, zhifeng@iems.nwu.edu).

performance of a direct method-based implementation for unstructured problems. Although efforts have been made to use indirect methods (for instance, GMRES/QMR methods [5, 6] and the conjugate gradient method [11, 13, 21]), it is generally understood that for unstructured problems direct methods are more suitable [1, 3, 16]. The iterative methods are found more suitable for structured problems, particularly network flow problems [11, 13, 20, 22, 23]. The performance of iterative methods depends on the quality of preconditioner.

The proposed method for generating corrector directions is a hybrid of direct and iterative methods. The particular design is based on two observations. First, for the information from indirect methods to be effective, iterates should stay in a certain subspace. In particular, the infeasibility of the linear equations in the iterative methods should remain only in equations corresponding to the nonlinear part of the KKT conditions. The second observation is that the directions generated by iterative methods are generally “weak” and typically by themselves not good enough to give sufficient progress in the algorithm. This is particularly true when high solution accuracy is desired. As a result, iterative methods don’t give a stable implementation when accuracy requirements are high.

Our approach to generating corrector directions uses a Krylov subspace. We use an exact factorization from an “earlier iteration” to ensure the desired properties in the directions we generate. This can be viewed as choosing an easily available special preconditioner for the KKT equations. In this sense, our method is a hybrid of higher-order correction strategy and an *inexact* iterative method. A linear programming subproblem is solved at each iteration to generate a combined direction. Numerical experiments using a modification of the software package PCx [3] show the advantage of this approach for the problems with relatively dense Cholesky factors.

This paper is organized as follows. In the next section we give conditions that guarantee the convergence of a generic predictor-corrector algorithm. We present the linear programming subproblem used for generating a search direction and our modified primal-dual interior-point algorithm in this section. Convergence analysis of the proposed generic algorithm is given in section 3. In section 4 we discuss Krylov subspace-based corrector directions. Computational results are given in section 5. Concluding remarks and future research directions are discussed in section 6. Some proofs and tables containing numerical results are given in the appendix.

Notation. Lower-case letters are used for vectors, and upper-case letters are used for matrices. Superscript T denotes the transpose of a vector. For a vector denoted by a lower-case letter, the upper-case letter means the diagonal matrix whose diagonal elements are the components of the vector. For example, $X = \text{diag}(x_1, x_2, \dots, x_n)$ is the diagonal matrix associated with the vector $x = (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^n$. e denotes the vector $(1, 1, \dots, 1)^T$. The vector Euclidean norm is denoted by $\|x\| = \sqrt{x^T x}$, and $\|x\|_\infty$ is the infinity norm. The norm of a matrix A , $\max_{\|x\|=1} \|Ax\|$, is represented by $\|A\|$. The notation $x \geq 0$ (or $x \in \mathbb{R}_+^n$) implies that all components of the vector x are nonnegative. The same symbol, 0, is used to denote the number zero, the zero vector, and the zero matrix. I denotes the identity matrix.

2. Convergence conditions for a generic primal-dual predictor-corrector algorithm.

Preliminaries. We describe our ideas for the standard form linear program:

$$(1) \quad \min_{x \in \mathbb{R}^n} \{c^T x \mid Ax = b, x \geq 0\},$$

where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, and $c \in \mathbb{R}^n$. We assume that A has full row rank. The dual problem to (1) is written as

$$(2) \quad \max_{s \in \mathbb{R}^n, y \in \mathbb{R}^m} \{b^T y \mid A^T y + s = c, s \geq 0\}.$$

The KKT optimality conditions for the primal and dual problems (1) and (2) are as follows:

$$(3) \quad Ax = b,$$

$$(4) \quad A^T y + s = c,$$

$$(5) \quad \begin{aligned} Xs &= 0, \\ x, s &\geq 0. \end{aligned}$$

Equations (3) and (4) are primal and dual feasibility, respectively, and the nonlinear equation (5) is called the complementarity condition. Points $(x, y, s) \in \mathbb{R}_+^n \times \mathbb{R}^m \times \mathbb{R}_+^n$ satisfying (3), (4), and $Xs = \mu e$ are called points on the central path and are denoted by $(x(\mu), y(\mu), s(\mu))$ for some parameter μ . Primal-dual interior-point algorithms can be viewed as a variant of Newton's method applied to the system of equations

$$(6) \quad \begin{aligned} Ax &= b, \\ A^T y + s &= c, \\ Xs &= \tilde{r}, \end{aligned}$$

where $\tilde{r}$ is a suitably chosen vector to be specified later.

A linearization of (6) at (x, y, s) with $x > 0, s > 0$ yields the system of linear equations

$$(7) \quad \begin{bmatrix} X & S & 0 \\ I & 0 & A^T \\ 0 & A & 0 \end{bmatrix} \begin{bmatrix} \Delta s \\ \Delta x \\ \Delta y \end{bmatrix} = \begin{bmatrix} r \\ q \\ p \end{bmatrix},$$

where the right-hand side in (7) is given by

$$(8) \quad p := b - Ax, q := c - A^T y - s, r := \tilde{r} - Xs.$$

The predictor direction (or the primal-dual affine direction) $(\Delta x^{aff}, \Delta y^{aff}, \Delta s^{aff})$ is given by taking $\tilde{r} = 0$. A matrix factorization is required for computing the predictor direction. This factorization dominates the computational efforts in interior-point methods. Therefore, in practice it is often advantageous to use the same matrix again with appropriately chosen right-hand sides to improve on the predictor direction. The directions computed with the goal of improving the predictor direction are called the corrector directions.

LP subproblem. Since several different approaches [10, 12, 15, 18] are possible to compute corrections, it is useful to identify conditions under which a generic predictor-corrector algorithm will converge. Let $(\Delta x^0, \Delta y^0, \Delta s^0), \dots, (\Delta x^j, \Delta y^j, \Delta s^j)$ be j additional corrector directions generated by solving (7) with certain choices of (r, q, p) . These can be Mehrotra correctors, Gondzio correctors, Jarre-Wechs correctors, or other correctors such as the one described in section 4. Let the direction $(\Delta x^c, \Delta y^c, \Delta s^c)$ be computed from (7) with right-hand side $(\mu e, 0, 0)$. Here μ is an appropriately chosen parameter. Let $(\Delta x(\rho^x), \Delta y(\rho^s), \Delta s(\rho^s))$ represent a combination

of these directions:

$$\begin{aligned}
 \Delta x(\rho^x) &= \rho_{aff}^x \Delta x^{aff} + \rho_0^x \Delta x^0 + \rho_1^x \Delta x^1 + \cdots + \rho_j^x \Delta x^j + \rho_c^x \Delta x^c, \\
 \Delta y(\rho^s) &= \rho_{aff}^s \Delta y^{aff} + \rho_0^s \Delta y^0 + \rho_1^s \Delta y^1 + \cdots + \rho_j^s \Delta y^j + \rho_c^s \Delta y^c, \\
 \Delta s(\rho^s) &= \rho_{aff}^s \Delta s^{aff} + \rho_0^s \Delta s^0 + \rho_1^s \Delta s^1 + \cdots + \rho_j^s \Delta s^j + \rho_c^s \Delta s^c, \\
 p(\rho^x) &= A \Delta x(\rho^x), \quad q(\rho^s) = A^T \Delta y(\rho^s) + \Delta s(\rho^s),
 \end{aligned}
 \tag{9}$$

where $\rho^x = (\rho_{aff}^x, \rho_0^x, \dots, \rho_j^x, \rho_c^x)$ and $\rho^s = (\rho_{aff}^s, \rho_0^s, \dots, \rho_j^s, \rho_c^s)$.

The combination (9) should be such that it ensures overall convergence by taking a step of appropriate length along $(\Delta x(\rho^x), \Delta y(\rho^s), \Delta s(\rho^s))$. Toward this we consider the linear subproblem (subLP):

$$\begin{aligned}
 &\max t, \\
 (10) \quad &x + \Delta x(\rho^x) \geq 0, \\
 (11) \quad &s + \Delta s(\rho^s) \geq 0, \\
 (12) \quad &\Delta x(\rho^x) \leq c_1 \delta x, \\
 (13) \quad &\Delta s(\rho^s) \leq c_1 \delta s, \\
 (14) \quad &\|p - p(\rho^x)\|_\infty \leq (1 - \delta) \|p\|_\infty, \\
 (15) \quad &\|p(\rho^x)\|_\infty \leq \delta \|p\|_\infty, \\
 (16) \quad &\|q - q(\rho^s)\|_\infty \leq (1 - \delta) \|q\|_\infty, \\
 (17) \quad &\|q(\rho^s)\|_\infty \leq \delta \|q\|_\infty, \\
 (18) \quad &s^T \Delta x(\rho^x) + x^T \Delta s(\rho^s) \leq \delta(\beta_2 - 1) x^T s, \\
 (19) \quad &S \Delta x(\rho^x) + X \Delta s(\rho^s) \geq \delta(\beta_1 \mu e - X s), \\
 (20) \quad &c^T(x + \Delta x(\rho^x)) - b^T(y + \Delta y(\rho^s)) \leq (1 - t) \bar{\mu}, \\
 &0 \leq t \leq \delta \leq 1,
 \end{aligned}$$

where $0 < \beta_1 < \beta_2 < 1$ and $c_1 > 1$ are some parameters and $\mu = \frac{x^T s}{n}$, $\bar{\mu} = \max\{c^T x - b^T y, x^T s, \gamma_p \|Ax - b\|_\infty, \gamma_d \|A^T y + s - c\|_\infty\}$ for some $\gamma_p > 0$, $\gamma_d > 0$. The parameters β_1, β_2 , and γ satisfy additional conditions: $\beta_2 < 1 < \beta_1/\gamma$, $\beta_1 \leq \sqrt{\gamma/2n}$, and $\gamma < 1$. These conditions are used in lower bounding the step length in the proof while satisfying (20) and Lemma 3.3. The parameter γ controls closeness to centrality while taking a step. The constant c_1 is a large constant whose theoretical value is specified in (34) in Appendix A.

Conditions (10) and (11) in subLP preserve the nonnegativity of x and s . Conditions (12) and (13) ensure that the combined predictor-corrector direction remains bounded. Conditions (14)–(17) reduce the infeasibility of primal and dual problems at a sufficiently fast rate. Condition (18) ensures that the complementary gap reduces at a sufficient rate. Condition (19) is used to ensure that the combined search direction stays inside the neighborhood of the central path while giving sufficient progress. Condition (20) guarantees the reduction in the difference of primal and dual objectives along the direction $(\Delta x(\rho^x), \Delta y(\rho^s), \Delta s(\rho^s))$, while simultaneously ensuring that it is not very different from the complementarity gap and feasibility. The use of two variables t and δ allow us to use different values for ρ^x and ρ^s . This allows us to develop a practical implementation that very closely maintains the theoretical properties discussed in this and the next section. The subLP (10)–(20) is more general than a similar search finding problem introduced by Jarre and Wechs [12]. In particular, it allows the use of generic corrector directions, as well as different primal and

dual combinations while maintaining global convergence. Essentially, subLP ensures that the combined direction retains the properties of the Newton direction in [14] and only strengthens them without compromising the global convergence. We will discuss implementation aspects of these conditions in section 5.

Generic predictor-corrector method.

ALGORITHM 1.

INPUT: An initial feasible point $(x^0, y^0, s^0) > 0$, convergence parameters: $\epsilon^*, \epsilon_p, \epsilon_d > 0, \omega^*$, and algorithm parameters: $\tilde{\mu}_0 = \max\{\gamma_p \|Ax^0 - b\|_\infty, \gamma_d \|A^T y^0 + s^0 - c\|_\infty\}$ with $\gamma_p, \gamma_d > 0$,

$0 < \gamma < \beta_1 < \beta_2 < 1$ with $\beta_1 \leq \sqrt{\gamma/(2n)}$ and $\gamma \ll \min_i x_i s_i / \max\{x^T s, \tilde{\mu}_0\}$.

For $k = 0, 1, 2, \dots$

Step 1. Set $(x, y, s) = (x^k, y^k, s^k)$ and let

$$\mu = \frac{x^T s}{n}, \quad \bar{\mu} = \max\{c^T x - b^T y, x^T s, \tilde{\mu}_k\}.$$

If $\bar{\mu} < \epsilon^*$, $\|Ax - b\|_\infty < \epsilon_p$, and $\|A^T y + s - c\|_\infty < \epsilon_d$, or if $\|(x, s)\|_\infty > \omega^*$, then STOP, else go to Step 2.

Step 2. Compute the predictor direction $(\Delta x^{aff}, \Delta y^{aff}, \Delta s^{aff})$ and generate j correction directions

$$(\Delta x^0, \Delta y^0, \Delta s^0), \dots, (\Delta x^j, \Delta y^j, \Delta s^j).$$

In addition, (if needed) compute $(\Delta x^c, \Delta y^c, \Delta s^c)$, then solve subLP to obtain the search direction $(\Delta x(\rho^x), \Delta y(\rho^s), \Delta s(\rho^s))$;

Step 3. Set

$$\Delta x = \Delta x(\rho^x), \quad \Delta y = \Delta y(\rho^s), \quad \Delta s = \Delta s(\rho^s),$$

and compute

$$\bar{\alpha} := \max_{\alpha \in [0,1]} \left\{ \alpha \mid (X + \alpha \Delta X)(s + \alpha \Delta s) \geq \frac{\gamma}{n} \max\{(1 - \alpha \delta^*) \bar{\mu}, (x + \alpha \Delta x)^T (s + \alpha \Delta s)\} e \right\};$$

where δ^* is the value of δ at an optimum solution of subLP.

Step 4. Update $k \leftarrow k + 1$ and the new iterate

$$(x^{k+1}, y^{k+1}, s^{k+1}) := (x^k, y^k, s^k) + \bar{\alpha}(\Delta x, \Delta y, \Delta s),$$

$$\tilde{\mu}_{k+1} = \max \{ \gamma_p \|A^T y^{k+1} + s^{k+1} - c\|_\infty, \gamma_d \|Ax^{k+1} - b\|_\infty \}.$$

End (For)

3. Global convergence of the generic predictor-corrector method. We prove the following convergence theorem for Algorithm 1 in this section.

THEOREM 1. *Algorithm 1 stops after finitely many iterations. In particular, in Step 1, it either identifies an approximate solution of problems (1) and (2) or generates (x^k, s^k) satisfying $\|(x^k, s^k)\|_\infty > \omega^*$ for some given ω^* . Furthermore, the algorithm converges linearly.*

Our proof of Theorem 1 borrows heavily from the analysis in Jarre and Wechs [12] and Kojima, Megiddo, and Mizuno [14]. In particular, Kojima, Megiddo, and Mizuno [14] showed (for $l = 2$) in (21)–(23) that for suitably chosen parameters $\rho_{aff}(\Delta x^{aff}, \Delta y^{aff}, \Delta s^{aff}) + \rho_c(\Delta x^c, \Delta y^c, \Delta s^c)$ can be used to generate a sequence of iterates in a neighborhood defined by

$$(21) \quad x_i s_i \geq \gamma \frac{x^T s}{n} \quad \text{for } 1 \leq i \leq n,$$

$$(22) \quad x^T s \geq \gamma_p \|Ax - b\|_l \quad \text{or} \quad \|Ax - b\|_l \leq \epsilon_p,$$

$$(23) \quad x^T s \geq \gamma_d \|A^T y + s - c\|_l \quad \text{or} \quad \|A^T y + s - c\|_l \leq \epsilon_d.$$

The iterates generated in Algorithm 1 are in the neighborhood for $l = \infty$.

We divide this proof into three parts. The first part constructs a feasible solution for subLP in Step 2 of Algorithm 1. The second part shows a lower bound for the step length $\bar{\alpha}$ in Step 3 that maintains the new iterate in the desired neighborhood. The third part proves global linear convergence. These are given in the next three subsections.

Now let us assume that $\|(x, s)\|_\infty \leq \omega^*$ is satisfied throughout the algorithm. Otherwise, the algorithm terminates and the theorem holds true.

3.1. Feasible solution for the subLP. We show that $\rho^x = \rho^s = \bar{\rho} := (\frac{1}{c_1}, 0, \dots, 0, \frac{\beta_1}{c_1})$, $\bar{\delta} := \frac{1}{c_1}$, and $\bar{t} := \frac{1}{4c_1}$ is a feasible solution to the subLP. The constant c_1 is defined in (34) while proving Lemma 3.1. Note that $c_1(\Delta x(\bar{\rho}), \Delta y(\bar{\rho}), \Delta s(\bar{\rho}))$ is the Newton direction in Kojima, Megiddo, and Mizuno [14]. The construction here is similar to Jarre and Wechs except for small differences in parameter selection.

LEMMA 3.1. *For the choice of c_1 given in (34) and $\bar{\rho}$ defined as above, $\Delta x(\bar{\rho})$ and $\Delta s(\bar{\rho})$ satisfy*

$$\|X^{-1}\Delta x(\bar{\rho})\|, \quad \|S^{-1}\Delta s(\bar{\rho})\| \leq 1.$$

Proof. See the appendix. $\square$

Feasibility of conditions (10)–(13), (14)–(17), and (18)–(19). From Lemma 3.1 it is easy to see that (10)–(13) hold for $(\Delta x(\bar{\rho}), \Delta y(\bar{\rho}), \Delta s(\bar{\rho}))$. The inequalities (14)–(17) are feasible because $p(\bar{\rho}) = \frac{1}{c_1} A \Delta x^{aff} = \frac{1}{c_1} p$, and similarly $q(\bar{\rho}) = \frac{1}{c_1} q$. Conditions (18) and (19) are satisfied because $X \Delta s(\bar{\rho}) + S \Delta x(\bar{\rho}) = \frac{1}{c_1}(\beta_1 \mu e - Xs)$ and $\beta_2 > \beta_1$.

Feasibility of condition (20). The following result was proved in Jarre and Wechs [12].

LEMMA 3.2. *The vectors $\Delta x(\bar{\rho})$ and $\Delta s(\bar{\rho})$ satisfy*

$$c^T(x + \Delta x(\bar{\rho})) - b^T(y + \Delta y(\bar{\rho})) \leq \frac{1}{2c_1}(x^T s + u^T \Lambda^2 u) + \left(1 - \frac{1}{c_1}\right)(c^T x - b^T y),$$

where $u := c_1(X \Delta s(\bar{\rho}) + S \Delta x(\bar{\rho})) + Xs$ and $\Lambda := X^{-1/2} S^{-1/2}$.

Using the definition of $\bar{\mu}$, $u = \beta_1 \mu e$, and $\beta_1 \leq \sqrt{\gamma/2n}$ from Lemma 3.2 we obtain

$$\begin{aligned} c^T(x + \Delta x(\bar{\rho})) - b^T(y + \Delta y(\bar{\rho})) &\leq \left(1 - \frac{1}{2c_1}\right) \bar{\mu} + \frac{1}{2c_1} u^T \Lambda^2 u \\ &= \left(1 - \frac{1}{2c_1}\right) \bar{\mu} + \frac{1}{2c_1} \beta_1^2 \mu^2 e^T \Lambda^2 e \\ &\leq \left(1 - \frac{1}{2c_1}\right) \bar{\mu} + \frac{1}{2c_1} \frac{\gamma}{2n} \frac{n \mu^2}{\gamma \max\{\mu, \tilde{\mu}/n\}} \\ &\quad (\text{by (33) in Appendix A}) \\ &\leq \left(1 - \frac{1}{4c_1}\right) \bar{\mu}. \end{aligned}$$

3.2. A lower bound for $\bar{\alpha}$ in Step 3. In this section we give a lower bound for the step size taken along a direction obtained from an optimal solution of subLP. Note that although we do not require $\rho^x = \rho^s$, the present analysis uses equal steps for the combined primal and dual search directions. The ability to take different steps for primal and dual directions is more desirable from an implementation perspective.

LEMMA 3.3. *The step length $\bar{\alpha}$ in Step 3 of Algorithm 1 is at least $\frac{\beta_1 - \gamma\beta_2}{(n+\gamma)c_1^2}$, $\beta_2 < 1 < \beta_1/\gamma$.*

Proof. Following Jarre and Wechs [12] we consider two cases.

Case I. $(1 - \alpha\delta)\tilde{\mu} \geq (x + \alpha\Delta x(\rho^x))^T(s + \alpha\Delta s(\rho^s))$. In this case,

$$\begin{aligned} & (X + \alpha\Delta X(\rho^x))(s + \alpha\Delta s(\rho^s)) - \frac{\gamma}{n}(1 - \alpha\delta)\tilde{\mu}e \\ &= Xs + \alpha \underbrace{(X\Delta s(\rho^x) + S\Delta x(\rho^s))}_{r(\rho):=} + \alpha^2\Delta X(\rho^x)\Delta s(\rho^s) - \frac{\gamma}{n}(1 - \alpha\delta)\tilde{\mu}e \\ &= (1 - \alpha\delta) \underbrace{(Xs - \frac{\gamma}{n}\tilde{\mu}e)}_{\geq 0} + \alpha(r(\rho) + \delta Xs) + \alpha^2\Delta X(\rho^x)\Delta s(\rho^s) \\ &\geq \alpha\delta\beta_1\mu e + \alpha^2X^{-1}\Delta X(\rho^x)(XS)S^{-1}\Delta s(\rho^s) \quad (\text{from (19)}) \\ &\geq \alpha\delta\beta_1\mu e - \alpha^2n\mu c_1^2\delta^2e \quad (\text{by (10)–(13)}). \end{aligned}$$

Since $\delta < 1$, the above inequality holds when $0 \leq \alpha \leq \frac{\beta_1}{nc_1^2} \leq \frac{\beta_1}{nc_1^2\delta}$.

Case II. $(1 - \alpha\bar{\rho})\tilde{\mu} < (x + \alpha\Delta x)^T(s + \alpha\Delta s)$. In this case,

$$\begin{aligned} & (X + \alpha\Delta X(\rho^x))(s + \alpha\Delta s(\rho^s)) - \frac{\gamma}{n}(x + \alpha\Delta x(\rho^x))^T(s + \alpha\Delta s(\rho^s))e \\ &= Xs + \alpha(X\Delta s(\rho^s) + S\Delta x(\rho^x)) + \alpha^2\Delta X(\rho^x)\Delta s(\rho^s) \\ &\quad - \frac{\gamma}{n}(x^Ts + \alpha(x^T\Delta s(\rho^s) + s^T\Delta x(\rho^x)) + \alpha^2\Delta x(\rho^x)^T\Delta s(\rho^s))e \\ &= (1 - \alpha\delta) \underbrace{(Xs - \gamma\mu e)}_{\geq 0} + \alpha(r(\rho) + \delta Xs) - \alpha\frac{\gamma}{n}((r(\rho) + \delta Xs)^Te)e \\ &\quad + \alpha^2(\Delta X(\rho^x)\Delta s(\rho^s) - \frac{\gamma}{n}(\Delta x^T(\rho^x)\Delta s(\rho^s))e) \\ &\geq \alpha\delta\beta_1\mu e - \delta\alpha\frac{\gamma}{n}\beta_2(x^Ts)e + \alpha^2(X^{-1}\Delta x(\rho^x))(XS)(S^{-1}\Delta s(\rho^s)) \\ &\quad - \frac{\gamma}{n}(X^{-1}\Delta x(\rho^x))^T(XS)(S^{-1}\Delta s(\rho^s))e \quad (\text{from (18) and (19)}) \\ &\geq \alpha\delta\mu(\beta_1 - \gamma\beta_2)e - \alpha^2(n + \gamma)\delta^2c_1^2\mu e \quad (\text{by (10)–(13)}). \end{aligned}$$

Hence $\bar{\alpha}$ has the lower bound $0 < \bar{\alpha} = \frac{\beta_1 - \gamma\beta_2}{(n+\gamma)c_1^2} \leq \frac{\beta_1 - \gamma\beta_2}{\delta(n+\gamma)c_1^2}$, because $\beta_2 < 1 < \beta_1/\gamma$ and $\delta < 1$. $\square$

3.3. Convergence. From conditions (14)–(17),

$$\begin{aligned} \|Ax^{k+1} - b\|_\infty &= \|-p^k + \alpha p^k(\rho^x)\|_\infty \leq \|p^k - p^k(\rho^x)\|_\infty + (1 - \alpha)\|p^k(\rho^x)\|_\infty \\ &\leq (1 - \alpha\delta)\|p^k\|_\infty, \end{aligned}$$

and similarly $\|A^Ty^{k+1} + s^{k+1} - c\|_\infty \leq (1 - \alpha\delta)\|q^k\|_\infty$. Hence, $\tilde{\mu}_k \leq (1 - \alpha\delta)\tilde{\mu}_{k-1}$. Because

$$\bar{\alpha} \geq \hat{\alpha} := \frac{\beta_1 - \gamma\beta_2}{(n + \gamma)c_1^2} \quad \text{and} \quad \delta \geq \hat{\delta} := \frac{1}{4c_1},$$

$\tilde{\mu}_k$ is bounded above by

$$(24) \quad \tilde{\mu}_k \leq (1 - \hat{\alpha}\hat{\delta})^k \tilde{\mu}_0.$$

Condition (20) of the subLP guarantees that

$$c^T x^{k+1} - b^T y^{k+1} \leq (1 - \hat{\alpha}\hat{\delta}) \bar{\mu}_k.$$

From

$$Ax = b - p, \quad A^T y + s = c - q,$$

we have

$$(25) \quad x^{k+1T} s^{k+1} = c^T x^{k+1} - b^T y^{k+1} - q^{k+1T} x^{k+1} + p^{k+1T} y^{k+1}.$$

Note that

$$\|y^{k+1}\| = \|(AA^T)^{-1} A(c - q^{k+1} - s^{k+1})\| \leq \|(AA^T)^{-1} A\|(\|c\| + \tilde{\mu}_{k+1}/\gamma_d + \omega^*) \leq c_3 \omega^*.$$

Inserting these estimates into (25) yields

$$\bar{\mu}_{k+1} \leq \bar{\mu}_k(1 - \hat{\alpha}\hat{\delta}) + \omega^*(1 - \hat{\alpha}\hat{\delta})\tilde{\mu}_k/\gamma_p + c_3\omega^*(1 - \hat{\alpha}\hat{\delta})\tilde{\mu}_k/\gamma_d \leq \bar{\mu}_k(1 - \hat{\alpha}\hat{\delta}) + c_4(1 - \hat{\alpha}\hat{\delta})^{k+1},$$

where $c_4 = \frac{2c_3\omega^*\mu_0}{\min\{\gamma_p, \gamma_d\}}$. By Proposition 1 in [12] we have constants $c_M < 1$ and $M > 0$ such that

$$\bar{\mu}_k \leq M(c_M)^k.$$

4. Generating search directions from Krylov subspace. In this section we present a new method for generating corrector directions. The basic idea is to blend information from an iterative method with information generated from a direct method. The analysis of the previous section was motivated to develop stable implementation based on this approach for generating corrector directions.

The standard approach for solving the linear system (7) in interior-point algorithms is to use direct methods based on a factorization of some matrix. Usually, this is done by first reducing the linear system resulting from the KKT optimality conditions to a symmetric positive definite linear system that is then solved by sparse Cholesky factorization [3, 7]. An alternative is to solve the symmetric indefinite 2×2 -block system by means of a sparse adaption of the Bunch–Parlett decomposition [2]. In both cases, a substantial amount of fill-in may occur.

The fill-ins during factorization in direct methods are one of the motivations for using iterative methods. When working with the normal equations, we may use a preconditioned CG method [13]. However, if the augmented or full system is used, Krylov subspace algorithms are useful [5]. Many iterative Krylov subspace algorithms have been developed to solve large sparse linear systems of equations—for instance, the Lanczos method, GMRES, and QMR [8, 9]. The power basis used in the definition of Krylov subspaces is, in general, ill-conditioned and requires preconditioning. When using an iterative method, linear systems are solved to a low relative accuracy.

The approach we develop in this section is a hybrid of direct and indirect methods for corrector directions. First, a tentative future iterate is generated from existing predictor and corrector directions. A Krylov subspace is generated at this tentative

iterate. The Krylov subspace corrector directions are appropriately translated and combined with the predictor and possibly other corrector directions. This process can be repeated if beneficial.

To illustrate the idea, let H_k be the coefficient matrix in (7) at the k th iteration, z be the solution vector, and $\bar{b}$ be the right-hand side. That is,

$$(26) \quad H_k = \begin{bmatrix} X^k & S^k & 0 \\ I & 0 & A^T \\ 0 & A & 0 \end{bmatrix}, \quad z = \begin{bmatrix} \Delta s \\ \Delta x \\ \Delta y \end{bmatrix}, \quad \bar{b} = \begin{bmatrix} r \\ q \\ p \end{bmatrix}.$$

Then the linear system (7) is

$$(27) \quad H_k z = \bar{b}.$$

This system is solved by a direct method, e.g., Cholesky factorization of normal equations. Let (x^+, y^+, s^+) be a new iterate which is tentatively generated by using some search direction

$$x^+ = x^k + \alpha_x^+ d_x, \quad y^+ = y^k + \alpha_s^+ d_y, \quad s^+ = s^k + \alpha_s^+ d_s,$$

where α_x^+ and α_s^+ are appropriate steps and (d_x, d_y, d_s) satisfies

$$A d_x = p, \quad A^T d_y + d_s = q.$$

At the “tentative iterate,” in a standard interior-point algorithm we should solve

$$(28) \quad H^+ z = \bar{b}^+, \quad \bar{b}^+ = [r^+, p^+, q^+]^T,$$

to generate subsequent iterates. In the proposed approach we use this system to generate corrector directions via a Krylov subspace. We do this by preconditioning H^+ with H_k^{-1} .

DEFINITION 1. *The Krylov subspace of the above linear equation (28) with preconditioner H_k is defined to be*

$$(29) \quad K_j(H_k, H^+, \bar{b}^+) := \text{span}\{b^1, \Theta b^1, \Theta^2 b^1, \dots, \Theta^j b^1\},$$

where $b^1 = H_k^{-1} \bar{b}^+$ and $\Theta = I - H_k^{-1} H^+$.

Let $b^{i+1} = \Theta^i b^1$ for $i = 1, 2, \dots$. Then these vectors of the Krylov subspace have a useful property described in the following proposition.

PROPOSITION 1. *The vectors $b^1, b^2, \dots, b^j$ satisfy*

$$\begin{bmatrix} 0 & I & A^T \\ A & 0 & 0 \end{bmatrix} b^1 = \begin{bmatrix} q^+ \\ p^+ \end{bmatrix}, \quad \begin{bmatrix} 0 & I & A^T \\ A & 0 & 0 \end{bmatrix} b^i = 0 \text{ for } i = 2, \dots.$$

Proof. The proof of first equality is from the definition of H_k . The second equality follows from observing that $H_k b^i = (H_k - H^+) \Theta^{i-1} b^1$ and noting that H_k and H^+ are the same in the rows under consideration. $\square$

Remark 1. We can generate at most n linearly independent Krylov subspace vectors. This is true because at most n linearly independent vectors are possible that would satisfy the equations in Proposition 1.

Remark 2. In practice, we want to generate only a few Krylov subspace vectors so checking linear independence is not an issue.

Remark 3. A standard implementation of interior-point method using iterative methods will use the Krylov subspace vectors to generate an approximate direction at (x^+, y^+, s^+) , and move to the next iteration. This approach in our experience results in a less stable implementation with generally inferior performance.

5. Implementation and numerical experiments. In this section we first discuss modifications for the search direction conditions given in sublet. We follow this with a discussion on practical implementation strategies that use these conditions. Computational results based on an implementation of Krylov subspace corrector directions are discussed subsequently. The discussion in this section assumes that Δx^0 is the Mehrotra corrector (obtained by solving (7) with the right-hand side of $(\mu e - Xs - \Delta x^{aff} \Delta s^{aff}, 0, 0)$), $\Delta x^i = b^i$, where $b^i, i = 1, \dots$, are Krylov vectors in the previous section, and

$$(d_x, d_y, d_s) := (\Delta x^{aff}, \Delta y^{aff}, \Delta s^{aff}) + (\Delta x^0, \Delta y^0, \Delta s^0)$$

in primal and dual spaces. The steps α_x^+, α_s^+ are computed using the Mehrotra step length heuristic as implemented in PC [3].

Under these assumptions conditions (14)–(17) in the subLP simplify to

$$(30) \quad \delta \leq \rho_{aff}^x + (1 - \alpha_x^+) \rho_1^x \leq 1 \text{ and } \delta \leq \rho_{aff}^s + (1 - \alpha_s^+) \rho_1^s \leq 1.$$

Modification to subLP. In our implementation we replace (30) with the following conditions:

$$.1\delta \leq \rho_{aff}^x + (1 - \alpha_x^+) \rho_1^x \leq 1 \text{ and } .1\delta \leq \rho_{aff}^s + (1 - \alpha_s^+) \rho_1^s \leq 1,$$

and

$$\lambda[\rho_{aff}^x + \rho_1^x(1 - \alpha_x^+)] + (1 - \lambda)[\rho_{aff}^s + \rho_1^s(1 - \alpha_s^+)] \geq \delta,$$

where the parameter λ is chosen in $[0, 1]$. The weight factor λ balances reduction of the primal and dual infeasibility. Definition of λ follows Jarre and Wechs [12]; $\lambda := \frac{\gamma_p \|p\|}{\gamma_p \|p\| + \gamma_d \|q\|}$, where γ_p and γ_d are some parameters. This modification gave better computational performance for several problems. This is the only modification we made to subLP in our implementation.

Implementation strategy. Each major iteration starts with a fresh matrix factorization used to generate the Mehrotra corrector and Krylov directions as described in section 4. The computational results discussed below are based on a prespecified number of Krylov vectors. In order to compute the combined direction we first solve subLP heuristically without including (computing) $(\Delta x^c, \Delta y^c, \Delta s^c)$, and ignoring conditions (12), (13), and (19). If the optimal objective value of this subLP is larger than a threshold (.1), we scale the combined direction by $1/(\rho_{aff}^x + (1 - \alpha_x^+) \rho_1^x)$ and $1/(\rho_{aff}^s + (1 - \alpha_s^+) \rho_1^s)$ (ensuring feasibility of linear equality constraint for a full step), and compute a step using the Mehrotra step length heuristic [18]. If the new point generated in this way satisfies the centrality neighborhood conditions (21)–(23) and it satisfies $\bar{\mu}^{k+1} \leq \frac{1}{2} \bar{\mu}^k$, then we accept it and go to the new iteration.

Otherwise (if t^* is less than the threshold, or the new iterate computed above fails to satisfy neighborhood conditions, or $\bar{\mu}^{k+1} > \frac{1}{2} \bar{\mu}^k$), the direction $(\Delta x^c, \Delta y^c, \Delta s^c)$ and conditions (12), (13), and (19) are added to the subLP and a new combined direction is computed. We use the Mehrotra step length heuristic for the combined direction. If it is successful then we update the iterates. Otherwise, we compute an equal step for primal and dual combined directions as suggested in Step 3 of Algorithm 1 according to the theory. The implementation strategy is explained in Figure 1.

We now describe reasons for the above implementation strategy. The analysis of Algorithm 1 in section 3 assumes that equal step length is computed in both

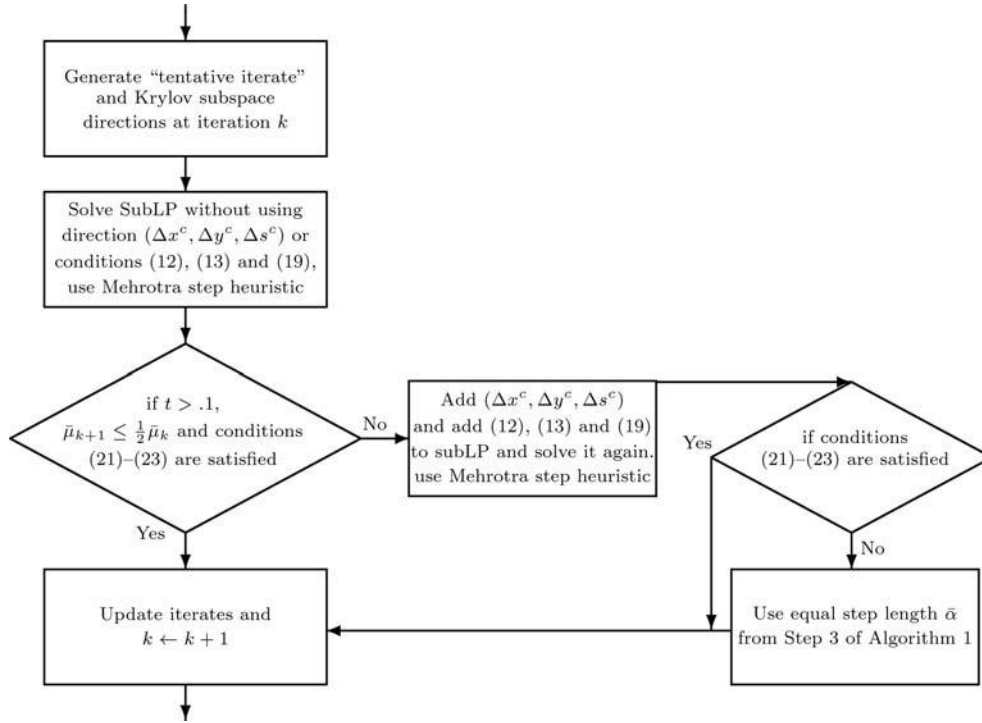


FIG. 1. Flow chart of implementation of Algorithm 1.

primal and dual spaces. It is well known that the practical performance of these methods is improved significantly by using different steps in primal and dual spaces. With the above implementation strategy we are ensuring the theoretical convergence properties of the algorithm while allowing different steps when possible. Also note that trial points generated using the Mehrotra step length heuristic are typically not at the boundary of the central path neighborhood. This generally works better than an implementation which always forces iterates to be at the boundary of the central path neighborhood.

Conditions (12) and (13) are unlikely to be active for most of the problems because of the large value of c_1 defined in (34). Condition (19) is used in the worst case analysis, and sometimes rules out good search directions. Consequently, always maintaining feasibility of constraints (19) while solving subLP generally gives an inferior performance. There are a few exceptions to this, which are mentioned during the discussion on computational results below. Also, the direction $(\Delta x^c, \Delta y^c, \Delta s^c)$ can be ignored in many problems because the Mehrotra corrector contains some centering information.

Solutions of subLP. For a few Krylov directions the subLP has a small number of variables, but a much larger number of constraints. We used a combination of primal and dual simplex methods to solve the subLP. An in-house research implementation of the primal and dual simplex method (with full pricing) was used. This was done as follows.

We first identified a “good candidate set” of constraints from subLP and used these to define our first small LP. Conditions (14)–(18) and (20) are always in subLP and a subset of (10), (11) is used to build the subLP. A heuristic to guess the set

of constraints most likely to be active is used to build the initial subLP. Subsequent subLPs always include (14)–(18) and a subset of (10), (11). A subset of conditions (12), (13), and (19) are also included in subLPs if they are needed. Since 0 is a feasible solution for subLP, after adding slacks to this problem we immediately have a feasible basis. Using this as an initial basis we solved the small LP to optimality. The optimal solution provides a feasible basis for the dual of subLP (we call it primal). Using this as a starting primal feasible basis, the primal simplex method was used. A small linear program was constructed by including at most $\sqrt{n}$ most violated inequalities of subLP corresponding to the primal feasible basis, where n is the number of variables for the original problem we are trying to solve. The small linear program was solved to optimality (tolerance 10^{-8}), and the process was repeated until subLP had no violated (tolerance 10^{-2}) inequality. Interestingly, no significant difference in the solution times and the number of iterations was observed for results with smaller optimality tolerance (10^{-8}).

Computational results. We implemented our algorithm in PCx release 1.1. The numerical results for the Netlib problems are reported in the attached tables. “PCx” represents the PCx version with only predictor and Mehrotra’s corrector. “PCxH” represents PCx version with predictor and Mehrotra’s corrector, and adaptively adding Gondzio’s higher order corrections according to the ratio “cost for back solve” and “cost for factorization.” No subLP is solved in this implementation. We denote the modified algorithm from this paper by “PCx0-4” for zero, one, two, three, and four Krylov subspace vectors. PCx0 uses the same (predictor and corrector) directions as PCx, except it solves a subLP to generate the combined direction. “PCx2F” and “PCx4F” denote the results with two and four Krylov vectors and full sets of constraints and centering direction in all the subLP, respectively. All versions are run on scaled problems.

The numerical experiments were performed on a Sun Sparc Ultra-5_10 with 256MB RAM. The parameters used were $\gamma = 10^{-4}$, $\gamma_p = \gamma_d = 100$, $\beta_1 = 1.5\gamma$, $\beta_2 = 0.5$, and $c_1 = O(n)$. These parameters have generally worked well. Note that γ controls the width of the central path neighborhood. Larger values of γ (we tried 10^{-2} or 10^{-3}) generally result in slightly worse performance. However, for values of $\gamma = 10^{-2}$ and 10^{-3} the implementation had additional numerical breakdowns for problems *df1001* and *greenbea*, respectively, if at an iteration we satisfied exactly the centrality constraints (21) and then we reduced γ to 0.99γ . This has helped occasionally with numerical stability and better performance. The following termination criterions were used for all runs. Optimal termination occurs when the current iterate satisfies the following test:

$$\begin{aligned}\frac{\|p\|}{1 + \|b^T\|} &\leq \text{prifeastol}, \\ \frac{\|q\|}{1 + \|c\|} &\leq \text{dualfeastol}, \\ \frac{\mu}{1 + |c^T x|} &\leq \text{opttol},\end{aligned}$$

where *prifeastol*, *dualfeastol*, and *opttol* are three tolerances whose default values are 10^{-8} , 10^{-8} , and 10^{-10} , respectively. If $\frac{\|x\|_\infty}{\|x_0\|_\infty} \geq 10^6$ or $\frac{\|s\|_\infty}{\|s_0\|_\infty} \geq 10^6$, where (x_0, s_0) is an initial point, then we declare unbounded variables found. These stopping criterion are subjective; however, they have worked well in our case.

The results in Tables 1 and 2 give the number of iterations (matrix factoriza-

tions) needed to achieve the indicated accuracy. The column “density” is the density of nonzero entries in the matrix $A(S^{-1}X)A^T$, on which Cholesky factorization is performed. The column “dg” is defined to be $\lfloor -\log_{10}(\frac{|c^T x - b^T s|}{1+|c^T x|}) \rfloor$, which measures the accuracy of the difference between primal and dual objective values upon termination. The column “It.” is the number of iterations. The column “#sLP” reports the average number of times small subLP were constructed and solved. The column “#C” reports the number of times a full subLP was solved as a result of either $t < .1$ or failure in the Mehrotra step length heuristic to produce an iterate in the central path neighborhood.

Several situations where PCx reports “Unknown” status are now traced to either numerical difficulties or situations where the solutions are getting large, suggesting that the problem is either infeasible or has an unbounded optimal face. These are indicated in Tables 1 and 2.

We can draw several conclusions from the results in Tables 1 and 2. First, there is no significant difference in the performance of PCx and PCx0. This suggests that by using the suggested strategy we can implement primal-dual methods while ensuring global convergence. For most problems, PCx0 took fewer (by one or two) iterations; however, it took significantly more iterations for problem *pilot4*. For this problem the performance improved when a full set of constraints were used in subLP in all the iterations (see results in Table 3). Unfortunately, adding a full set of constraints makes the performance worse for other problems. This can be seen from the results for PCx2F and PCx4F in Tables 1 and 2.

We conclude that adding one Krylov direction saves about 20%, two Krylov directions saves about 25%, three Krylov directions saves about 30%, and four Krylov directions saves about 37% number of matrix factorizations. This decrease in factorization is in addition to the decrease already obtained from using the Mehrotra corrector. We experimented with adding more Krylov directions, and our computational results suggest that the incremental decrease in the number of iterations as a result of additional Krylov directions is not significant.

Since for a large number of Netlib test problems the cost of Cholesky factorization is not much more than the cost for a forward and back solve, improvement in CPU time is not expected on these problems. However, problems for which time to compute Cholesky factorization dominates the total time improvements in CPU times are expected. Computation times for these problems are reported in Table 4. The best improvement is for problem *pds-20*, where the CPU time reduces by approximately 30%.

The results on infeasible Netlib problems are given in Table 5. Interestingly, in this case no significant reduction in the number of matrix factorizations is observed. We think that this is likely because H_{k-1} may no longer be a good preconditioner for H_k . For problems *ceria3d* and *cplex2*, we encountered numerical problems before detecting infeasibility or unbounded variables.

Finally, we comment on the time it takes to find the composite direction from the solution of subLP (with partial or full sets of constraints). In our current implementation, for problems where the Cholesky factorization is inexpensive, this time may be as much as 20% to 30% of the total CPU time on average. However, for problems where the Cholesky factorization is expensive (Table 4), this time is typically less than 8% of the total CPU time on average. While finding the composite direction, the calculations are typically dominated by the effort in finding violated inequalities,

TABLE 2
Numerical experiments with *Netlib* problems (continued).

Problems	Rows	Cols	Density (%)	PCx	PCxH	PCx0	PCx1	PCx2	PCx3	PCx4	PCx2F	PCx4F
				It.	dg	It.	dg	It.	dg	It.	dg	It.
plowt	722	2789	6.20	46	6	46	7	(35.8,48.33)	7	(31.9,74.10)	6	(27.10,3)
plowt8	2030	4883	21.86	34	6	26	6	(34.6,18.18)	6	(25.5,4.4)	6	(19.6,37.0)
recipe	471	180	11.96	9	14	9	14	(9.2,33.0)	9	(7.4,0)	6	27
scagr25	91	500	2.46	18	12	18	12	(16.4,81.3)	7	(13.5,1)	11	9
scagr7	129	140	8.26	14	8	14	8	(14.3,64.0)	9	(11.4,86.1)	9	15
scfwm1	330	457	9.19	17	8	17	8	(18.3,89.2)	12	(15.4,17.0)	10	13
scfwm2	660	914	4.82	20	11	20	11	(19.4,21.3)	8	(14.5,07.1)	7	12
scfwm3	990	1371	3.26	20	10	20	10	(21.5,81.6)	10	(15.4,27.1)	8	11
scfwm3	490	1169	4.81	22	7	22	7	(21.4,24.1)	8	(15.4,33.1)	7	12
scscl1	77	760	45.65	9	15	9	15	(9.2,67.0)	8	(15.4,87.0)	7	11
scscl6	147	1350	22.87	12	8	12	8	(8.3,38.0)	11	(15.4,87.0)	7	20
scscl8	397	2750	7.20	11	7	11	7	(12.3,58.0)	10	(10.5,9.1)	9	10
actap1	3090	4880	5.68	16	12	16	12	(16.4,63.5)	6	(11.4,82.1)	9	11
actap2	1090	1880	2.10	14	14	14	14	(13.5,54.3)	11	(11.6,1)	12	13
actap3	1480	2480	1.55	15	13	15	13	(13.4,62.4)	7	(12.3,83.0)	8	12
seba	515	1028	53.29	14	12	12	11	(14.5,14.4)	10	(11.4,82.1)	13	13
share1b	117	225	21.70	19	9	19	9	(19.4,26.3)	8	(15.4,2.0)	9	13
share2b	96	79	21.54	17	12	17	12	(17.6,12.8)	7	(14.6,93.6)	8	17
shell	536	1775	3.15	21	10	21	10	(20.4,40.2)	7	(18.5,5.2)	7	11
ship04l	402	2118	5.85	13	12	13	12	(12.4,1)	8	(10.4,3.0)	8	18
ship04s	402	1458	7.15	13	8	13	8	(13.3,92.1)	8	(11.4,36.0)	7	13
ship12l	1151	5427	2.79	16	6	16	6	(14.3,79.2)	7	(13.4,92.1)	9	14
ship12s	1151	2763	4.04	13	12	13	12	(12.3,17.0)	9	(10.4,3.0)	10	13
sierra	1227	2036	1.66	21	8	21	8	(19.2,63.0)	7	(15.4,2.0)	8	17
stair	356	467	22.88	14	8	13	7	(14.4,71.2)	9	(13.4,62.1)	11	13
standata	359	1075	4.54	13	8	13	8	(13.3,39.0)	15	(11.4,0)	15	12
standmps	467	1075	4.20	26	13	26	13	(22.5,55.6)	11	(20.5,85.4)	12	15
stocfor1	117	111	14.49	12	10	12	10	(11.3,18.0)	12	(9.4,0)	12	12
stocfor2	2157	2031	1.11	20	6	20	6	(20.4,94.0)	6	(15.5,07.0)	6	16
truss	1000	8806	10.60	20	8	16	6	(20.3,85.4)	8	(16.3,56.0)	8	12
tuff	333	587	20.96	20	12	18	12	(19.4,47.3)	11	(15.4,53.1)	10	18
vpbase	198	203	18.09	11	8	11	8	(11.2,82.0)	15	(9.3,1.0)	12	13
woodlp	244	2594	79.06	21	11	18	10	(19.4,42.2)	9	(17.4,47.2)**	14	10
woodw	1098	8405	11.83	31	9	31	9	(30.6,83.18)	8	(25.7,28.11)	6	19
cre-a	3516	4067	0.70	24	6	24	6	(24.5,92.13)	6	(19.6,26.6)	6	19
cre-b	9648	72447	1.72	40	6	36	5	(40.7,53.29)	5	(32.8,59.21)	6	24
cre-c	3068	3678	0.96	26	7	26	7	(25.6,12)	7	(20.5,3.5)	7	12
cre-d	8926	69980	2.49	40	5	36	6	(39.7,69.28)	5	(32.8,44.20)	6	20
ken-07	2426	3602	0.90	15	10	15	10	(15.2,67.0)	9	(12.3,92.0)	9	13
ken-11	1469	21349	0.19	21	6	21	6	(21.4,33.6)	5	(16.3,63.0)	5	16
ken-13	2863	42659	0.11	26	5	26	5	(28.6,07.15)	5	(19.3,74.0)	5	22
ken-18	1051	154699	0.06	30	5	28	5	(30.5,6.15)	4	(24.5,83.9)	4	28
osa-07	1118	23949	4.74	25	6	25	6	(18.6,94.8)	7	(18.8,11.8)	9	13
osa-14	2337	52460	2.25	27	7	27	7	(19.8,26.9)	7	(18.9,44.10)	6	14
osa-30	4350	100024	1.21	27	7	27	7	(24.1,13.15)	10	(17.8,59.7)	6	15
osa-60	232966	0.49	31	7	31	7	31	(22.7,77.12)	5	(19.9,84.11)	6	17
pds-02	2953	7535	1.26	24	6	24	6	(25.5,72.12)	6	(19.6,37.6)	7	19
pds-06	9881	28655	1.39	37	5	30	5	(35.6,8.22)	5	(28.6,96.14)	5	20
pds-10	1655	48763	1.37	41	5	33	5	(41.7,1.30)	5	(34.7,71.18)	5	26
pds-20	3387	105728	1.35	55	4	40	4	(58.8,88.53)	4	(42.9,02.31)	4	29
Total			2287	2183		2183		(175.2,618.425)		(158.7,634.4,318)	1889	1791
Average			22.64	21.61		21.61		(17.9,5.44,4.5)		(15.7,6.28,3.15)	18.7	17.7

* = Terminated with unknown status. ** = Numerical instability detected. † = Infeasibility was detected. ‡ = Unbounded variables found. § = Infeasibility detected in preprocess.

TABLE 3
Netlib problems which need strong centering components.

Program	PCx		PCxH		PCx2		PCX0F		PCx2F		PCx4		PCx4F	
Problems	It. digit		It. digit		It. digit		It. digit		It. digit		It. digit		It. digit	
pilot4	47	7	43	7	65	7	58	7	44	10	53	9	39	8

TABLE 4
Netlib problems showing computational savings.

Program			Density	PCx		PCxH		PCx4	
Problems	Rows	Cols	(%)	It.	Sec.	It.	Sec.	It.	Sec.
bnl2	2324	3489	4.1	35	11.43	30	10.69	29	9.517
cycle	1903	2857	5.4	23	3.743	21	3.856	13	3.663
d2q06c	2171	5167	5.9	29	17.86	25	16.67	18	16.11
d6cube	415	6184	67.2	18	5.581	16	5.583	11	5.655
df001	6071	12230	9.1	58	2090	44	1622	31	1181
israel	174	142	75.3	21	0.857	19	0.841	12	0.809
maros-r7	3136	9408	23.0	18	65.49	13	51.83	11	48.53
pilot	1441	3652	21.3	36	47.01	30	42.25	22	36.27
pilot87	2030	4883	21.8	34	153.5	26	125.1	19	99.32
seba	515	1028	53.2	14	4.769	12	4.350	9	3.793
pds-06	9881	28655	1.3	37	206.3	30	181.3	23	171.0
pds-10	16558	48763	1.3	41	1162	33	995.3	27	884.7
pds-20	3387	105728	1.3	55	12933	40	9917	34	8726

which requires a dense matrix vector product. This work grows as $O(nks)$, where k is the number of Krylov directions and s is the number of times we check for violated inequalities before finding the optimal solution in subLP. The computational results suggest that s is typically between 5 and 7 on average.

6. Conclusion. In this paper, we gave conditions for achieving linear convergence for a generic primal-dual interior-point method with corrections. A small linear program is formulated to generate the search direction. We also proposed an approach for generating correction directions by generating a Krylov subspace. Numerical results indicate that this method is promising. We described an implementation strategy balancing the need for efficiency with theoretical convergence.

In our proposed method, it is also possible to include directions from the previous iterations $(\Delta x_i, \Delta y_i, \Delta s_i)$ with $i = 1, \dots, k-1$ in the subLP. In this case, the search direction can be computed in the following way:

$$\begin{aligned}\Delta x(\theta, \rho) &= \theta_1 \Delta x_1 + \dots + \theta_{k-1} \Delta x_{k-1} + \Delta x(\rho), \\ \Delta y(\theta, \rho) &= \theta_1 \Delta y_1 + \dots + \theta_{k-1} \Delta y_{k-1} + \Delta y(\rho), \\ \Delta s(\theta, \rho) &= \theta_1 \Delta s_1 + \dots + \theta_{k-1} \Delta s_{k-1} + \Delta s(\rho),\end{aligned}$$

where $(\Delta x(\rho), \Delta y(\rho), \Delta s(\rho))$ is given in (9). Also one can use successive tentative solutions and generate more Krylov vectors at these tentative iterates. The practical value of both of these ideas is a topic of further research.

Appendix (Proof of Lemma 1). Define

$$(\Delta x, \Delta y, \Delta s) := c_1(\Delta x(\bar{\rho}), \Delta y(\bar{\rho}), \Delta s(\bar{\rho})).$$

Then Lemma 1 is equivalent to showing $\|X^{-1}\Delta x\|, \|S^{-1}\Delta s\| \leq c_1$. Note that $(\Delta x, \Delta y, \Delta s)$ is a solution of (7) for $p = b - Ax$, $q = c - A^T y - s$, $r = \beta_1 \mu e - Xs$. Let

$$D := X^{1/2} S^{-1/2} \text{ and } \Pi_R := DA^T(AD^2 A^T)^{-1} AD$$

TABLE 5
Numerical experiments with infeasible Netlib problems. Terminated with either infeasibility detected or unbound variables found.

Problems	Rows	Cols	Density (%)	PCx It.	PCxH It.	PCx0 It.	PCx1 It.	PCx2 It.	PCx3 It.	PCx4 It.	PCx2F It.	PCx4F It.	dg
bgdgl	348	407	5.182	6	1	6	0	6	0	7	0	6	0
bgetam	400	688	19.14	7	0	7	0	7	0	7	0	7	0
bgprtr	20	34	33.5	6	1	6	0	7	2	6	1	6	0
box1	231	261	4.22	4	0	4	0	6	1	6	0	6	0
ceria3d	3576	824	1.13	12	1	12	0	10	2	10**	0	11	0
chemcom	288	720	10.55	8	0	8	0	8	0	8	0	8	0
cplex1	3005	3221	0.2	5	0	5	0	4	0	4	0	5	0
cplex2	224	221	8.15	32**	6	32**	5	20**	5	18**	5	22**	6
ex72a	197	215	3.68	4	0	4	0	5	0	5	0	5	0
ex73a	193	211	3.72	4	0	4	0	6	0	6	0	6	0
forest6	66	95	9.275	11	2	11	0	13	1	13	1	12	0
galenet	8	68	5	0	5	0	6	5	0	5	0	6	0
gosh	3792	10733	3.216	14	0	13	0	12	0	11	0	10	0
gran §	2658	2520	2.513	10	0	10	0	9	0	8	0	8	0
greenbea	2504	5405	67.9	5	0	5	0	6	0	5	0	5	0
itest2	9	4	48	5	0	5	0	6	0	6	0	6	0
itest6	11	8	80.93	32	0	24	0	22	0	28	0	26	0
klein1	54	54	69.93	30	0	18	0	56	0	67	0	46	0
klein2	477	54	67.86	35	0	30	0	82	0	80	0	72	0
klein3	994	88	1.61	8	0	8	0	7	0	6	0	6	0
mondou2	312	604	10.64	24	0	24	0	19	0	12	0	14	0
pang	361	460	17.96	32	2	31	1	37	1	38	1	37	0
pilot4i	410	1000	9.3	35	1	35	1	29	1	27	1	28	1
qual	323	464	8.37	9	0	9	0	8	0	8	0	7	0
reactor	318	637	9.26	20	0	20	0	24	1	26	0	24	0
refinery	323	464	9.30	32	1	31	0	31	1	27	1	25	0
voll	323	464											
woodinfe §	35	89											

* = Terminated with unknown status. ** = Numerical instability detected. † = Infeasibility was detected. ‡ = Unbounded variables found. § = Infeasibility detected in preprocess.

be the orthogonal projection onto the range of DA^T . Let $\Pi_N := I - \Pi_R$ be the orthogonal projection onto the null space of AD . Let $\bar{p}$ be the minimal norm solution of $A\bar{p} = p$. We can verify that

$$(31) \quad D^{-1}\Delta x = v + \Pi_N \Lambda r,$$

$$(32) \quad D\Delta s = -v + \Pi_R \Lambda r,$$

where $v = \Pi_R D^{-1}\bar{p} - \Pi_N Dq$ and $\Lambda = X^{-1/2}S^{-1/2}$. From (31), (32) and using $X^{-1}D = S^{-1}D^{-1} = \Lambda$, we obtain

$$X^{-1}\Delta x = \Lambda(v + \Pi_N \Lambda r),$$

$$S^{-1}\Delta s^0 = \Lambda(\Pi_R \Lambda r - v).$$

From Step 3 of Algorithm 1 and the assumption that we have not terminated, we get

$$(33) \quad \|\Lambda^2\| \leq \frac{1}{\gamma \max\{\mu, \tilde{\mu}/n\}},$$

$$\|D^{-1}\| = \|D^{-1}S^{-1}S\| \leq \|\Lambda\|\|S\| \leq \frac{\omega^*}{\sqrt{\gamma \max\{\mu, \tilde{\mu}/n\}}},$$

$$\|D\| = \|DX^{-1}X\| \leq \|\Lambda\|\|X\| \leq \frac{\omega^*}{\sqrt{\gamma \max\{\mu, \tilde{\mu}/n\}}}.$$

It follows that

$$\|v\| \leq \frac{c_2 \omega^* \tilde{\mu}}{\sqrt{\gamma \max\{\mu, \tilde{\mu}/n\}}}$$

for some suitably chosen $c_2 \geq 0$, which depends on the condition number of A , γ_p , and γ_d . This is because $\|\Pi_R\|, \|\Pi_N\|$ is only dependent on the condition number of A . We can also bound

$$\|\Pi_N \Lambda r\|, \|\Pi_R \Lambda r\| \leq \|\Lambda\|\|r\| \leq \frac{\beta_1 n^{1/2} \mu + n^{3/2} \mu}{\sqrt{\gamma \max\{\mu, \tilde{\mu}/n\}}}.$$

It implies that $X^{-1}\Delta x$, $S^{-1}\Delta s$ have the following upper bound:

$$(34) \quad \begin{aligned} \|X^{-1}\Delta x\|, \|S^{-1}\Delta s\| &\leq \|\Lambda\|(\|v\| + \|\Lambda r\|) \\ &\leq \frac{c_2 \omega^* \tilde{\mu} + \beta_1 n^{1/2} \mu + n^{3/2} \mu}{\gamma \max\{\mu, \tilde{\mu}/n\}} \\ &\leq \frac{c_2 \omega^* n + \beta_1 n^{1/2} + n^{3/2}}{\gamma} =: c_1. \quad \square \end{aligned}$$

Acknowledgments. We thank two anonymous referees for their careful reading of the first version of this paper and their suggestions for improvements.

REFERENCES

- [1] I. ADLER, N. K. KARMARKAR, M. G. C. RESENDE, AND G. VEIGA, *An implementation of Karmarkar's algorithm for linear programming*, Math. Programming, 44 (1989), pp. 297–335.
- [2] J. R. BUNCH AND B. N. PARLETT, *Direct methods for solving symmetric indefinite systems of linear equations*, SIAM J. Numer. Anal., 8 (1971), pp. 639–655.
- [3] J. CZYZYK, S. MEHROTRA, M. WAGNER, AND S. J. WRIGHT, *PCx: An interior-point code for linear programming*, Optim. Methods Soft., 11/12 (1999), pp. 397–430.
- [4] I. S. DUFF, *The solution of augmented systems*, Pitman Res. Notes Math. Ser. Longman Sci. Tech., Harlow, UK, 1994.
- [5] R. W. FREUND, AND F. JARRE, *A QMR-based interior-point algorithm for solving linear programs*, Math. Programming, 76 (1997), pp. 183–210.
- [6] R. W. FREUND, F. JARRE, AND S. MIZUO, *Convergence of a class of inexact interior-point algorithms for linear programs*, Math. Oper. Res., 24 (1999), pp. 50–71.
- [7] R. FOURER AND S. MEHROTRA, *Solving symmetric indefinite systems in an interior-point method for linear programming*, Math. Programming, 62 (1993), pp. 15–39.
- [8] R. W. FREUND AND N. M. NACHTIGAL, *QMR: A quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 60 (1991), pp. 315–339.
- [9] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, 1996.
- [10] J. GONDZIO, *Multiple centrality corrections in a primal-dual method for linear-programming*, Comput. Optim. Appl., 6 (1996), pp. 137–156.
- [11] J. J. JUDICE, J. M. PATRICIO, L. F. PORTUGAL, M. G. C. RESENDE, AND G. VEIGA, *A Study of Preconditioners for Network Interior Point Methods*, AT&T Labs Research Technical Report, AT&T Shannon Laboratory, April 2001.
- [12] F. JARRE AND M. WECHS, *Extending Mehrotra's corrector for linear programs*, Advanced Modeling and Optimization, 1 (1999), pp. 38–60.
- [13] J. KORZAK, *Convergence analysis of inexact infeasible interior-point algorithm for solving linear programming problems*, SIAM J. Optim., 11 (2000), pp. 133–148.
- [14] M. KOJIMA, N. MEGIDDO, AND S. MIZUNO, *A primal-dual infeasible interior-point algorithm for linear programming*, Math. Programming, 61 (1993), pp. 263–280.
- [15] N. K. KARMARKAR, J. C. LAGARIAS, L. SLUTSMAN, AND P. WANG, *Power series variants of Karmarkar-type algorithms*, AT&T Tech. J., 68 (1989), pp. 20–36.
- [16] I. J. LUSTIG, R. E. MARSTEN AND D. F. SHANNO, *On implementing Mehrotra's predictor-corrector interior-point method for linear programming*, SIAM J. Optim., 2 (1992), pp. 435–449.
- [17] S. MEHROTRA, *Generalized Predictor-Corrector Methods and Their Performance*, TR91-17, Dept. of IE/MS, Northwestern University, Evanston, IL.
- [18] S. MEHROTRA, *On the implementation of a primal-dual interior-point method*, SIAM J. Optim. 2 (1992), pp. 575–601.
- [19] S. MEHROTRA, *Asymptotic convergence in a generalized predictor-corrector method*, Math. Programming, 74 (1996), pp. 11–28.
- [20] S. MEHROTRA AND J. WANG, *Conjugate gradient based implementation of interior point methods for network flow problems*, in Linear and Nonlinear Conjugate Gradient-Related Methods, AMS Summer Conference Proceedings, L. Adams and J. L. Nazareth eds., SIAM, Philadelphia, 1996, pp. 124–142.
- [21] S. MIZUNO AND F. JARRE, *Global and polynomial-time convergence of an infeasible-interior-point algorithm using inexact computation*, Math. Program., 84 (1999), pp. 105–122.
- [22] L. PORTUGAL, M. RESENDE, G. VEIGA, AND J. JUDICE, *An efficient implementation of an infeasible primal-dual network flow method*, Networks, 35 (2000), pp. 91–108.
- [23] M. RESENDE AND G. VEIGA, *An efficient implementation of a network interior point method*, in Network Flows and Matching: First DIMACS Implementation Challenge, D. S. Johnson and C. C. McGeoch, eds., DIMACS Ser. Discrete Math. Theoret. Comput. Sci., AMS, Providence, RI, 1993, pp. 299–348.
- [24] R. J. VANDERBEL, *LOQO User's Manual*, Technical Report SOR 92-5, Program in Statistics and Operations Research, Princeton University, Princeton, NJ, 1992.
- [25] S. VAVASIS, *Stable numerical algorithms for equilibrium systems*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1108–1131.
- [26] S. J. WRIGHT, *Stability of augmented system factorizations in interior-point methods*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 191–222.

QUADRATIC CONVERGENCE OF A NONSMOOTH NEWTON-TYPE METHOD FOR SEMIDEFINITE PROGRAMS WITHOUT STRICT COMPLEMENTARITY*

CHRISTIAN KANZOW[†] AND CHRISTIAN NAGEL[†]

Abstract. We consider a Newton-type method for the solution of semidefinite programs. This Newton-type method is based on a semismooth reformulation of the semidefinite program as a nonsmooth system of equations. We establish local quadratic convergence of this method under a linear independence assumption and a weak nondegeneracy condition that implies uniqueness of the optimal solution but does not imply strict complementarity.

Key words. semidefinite programs, Newton's method, quadratic convergence, nondegeneracy, strict complementarity

AMS subject classifications. 90C22, 90C46

DOI. 10.1137/S1052623403431147

1. Introduction. A semidefinite program is a minimization problem of the form

$$(1.1) \quad \min_{X \in \mathcal{S}^{n \times n}} C \bullet X \quad \text{s.t.} \quad A_i \bullet X = b_i \quad (i = 1, \dots, m), \quad X \succeq 0,$$

where $C, A_1, \dots, A_m \in \mathcal{S}^{n \times n}$ are symmetric matrices and $b \in \mathbb{R}^m$ is a given vector (the notation used here is standard in the semidefinite literature and will be defined at the end of this section).

Using some results from the corresponding duality theory, it is not difficult to see that, under a Slater-type condition, the semidefinite program (1.1) has a solution if and only if the *optimality conditions*

$$(1.2) \quad \begin{aligned} \sum_{i=1}^m \lambda_i A_i + S &= C, \\ A_i \bullet X &= b_i \quad \forall i = 1, \dots, m, \\ X \succeq 0, \quad S \succeq 0, \quad XS &= 0, \end{aligned}$$

have a solution. Many of the standard and efficient interior-point methods for the solution of semidefinite programs are based on this reformulation, see, e.g., [6, 2, 23, 21, 12, 15]. However, to the best of our knowledge, none of these interior-point methods is known to be locally superlinearly or quadratically convergent if the solution of (1.2) does not satisfy the strict complementarity condition.

Another approach is based on a reformulation of the optimality conditions (1.2) as a nonsmooth system of equations. This idea leads to the class of semismooth and smoothing methods; see [3, 18, 10, 19, 11]. One iteration of these methods is sometimes more expensive than one iteration of an interior-point method; however, the class of semismooth and smoothing methods has the advantage that the matrices

*Received by the editors July 10, 2003; accepted for publication (in revised form) May 4, 2004; published electronically April 8, 2005. This research was partially supported by the DFG (Deutsche Forschungsgemeinschaft).

<http://www.siam.org/journals/siopt/15-3/43114.html>

[†]University of Würzburg, Institute of Applied Mathematics and Statistics, Am Hubland, 97074 Würzburg, Germany (kanzow@mathematik.uni-wuerzburg.de, nagel@mathematik.uni-wuerzburg.de).

X and S do not have to belong to the cone of positive definite matrices. Hence they are of particular interest for warm or hot starts. Moreover, the main ideas of these methods can, in principle, be easily extended to nonlinear semidefinite programs.

In order to prove local fast convergence of these semismooth or smoothing methods, however, strict complementarity is also needed among some further assumptions. One exception is a result in [19], where local quadratic convergence of a nonsmooth Newton method is established for semidefinite complementarity problems without assuming strict complementarity, but using a positive definiteness assumption which is never satisfied when this result is applied to semidefinite programs. In the revision of the paper [19] (see [20]), the authors use an approach similar to ours and specialize their results to a Newton-type method applied to semidefinite programs, but then they have to assume strict complementarity in order to prove fast local convergence of their method.

The aim of this paper is now to have a closer look at the local convergence behavior of a nonsmooth Newton-type method for the solution of the optimality conditions (1.2). It turns out that we can prove local quadratic convergence of this method under a linear independence condition and a weak nondegeneracy condition that implies uniqueness of the solution but does not imply strict complementarity.

The organization of the paper is as follows. In section 2 we give a formal statement of our nonsmooth Newton method and present some background material. A linearized system is solved at each iteration of this method. We present a reformulation of this linearized system in the more standard matrix-vector form in section 3. Based on this reformulation, we present our local convergence analysis in section 4. Some numerical results illustrating the quadratic convergence behavior are given in section 5. We then close with some final remarks in section 6.

We use the following notation: For two matrices $A, B \in \mathbb{R}^{n \times n}$, we set

$$A \bullet B := \text{tr}(AB^T),$$

where $\text{tr}(C) := \sum_{i=1}^n c_{ii}$ denotes the trace of a matrix $C \in \mathbb{R}^{n \times n}$. It is easy to see that $\bullet$ defines a scalar product on the set of matrices $\mathbb{R}^{n \times n}$. We further write $\mathcal{S}^{n \times n}$ for the set of symmetric matrices in $\mathbb{R}^{n \times n}$, while $A \succeq 0$ and $A \succ 0$ indicate that A is a symmetric positive semidefinite and symmetric positive definite matrix, respectively. If $A \succeq 0$, we denote by $A^{1/2}$ the unique symmetric positive semidefinite square root of A . Moreover, for any matrix $A \in \mathcal{S}^{n \times n}$ we set $|A| = (A^2)^{1/2}$. Finally, if $E \succ 0$ is a given symmetric positive definite matrix, the corresponding *Lyapunov operator* L_E is defined by $L_E[X] := EX + XE$ ($X \in \mathcal{S}^{n \times n}$). Then it is well known (see [8]) that the resulting *Lyapunov equation* $L_E[X] = H$ has a unique solution for each symmetric $H \in \mathcal{S}^{n \times n}$, and we denote this solution by $L_E^{-1}[H]$.

2. Nonsmooth Newton method. In order to formulate the optimality conditions (1.2) as a system of equations, let us introduce the function

$$(2.1) \quad \phi(X, S) := X + S - |X - S| \quad (X, S \in \mathcal{S}^{n \times n}).$$

The mapping ϕ is usually called the *minimum function*. Note that it is not differentiable everywhere. Nevertheless, it has a number of interesting properties which we collect in the following result.

PROPOSITION 2.1. *Let ϕ be defined by (2.1). Then the following statements hold for any two matrices $X, S \in \mathcal{S}^{n \times n}$:*

(a) ϕ satisfies the equivalence

$$(2.2) \quad \phi(X, S) = 0 \iff X \succeq 0, \quad S \succeq 0, \quad XS = 0.$$

(b) If $E := |X - S|$ is nonsingular, then ϕ is continuously differentiable (in the sense of Fréchet) with

$$\nabla \phi(X, S)(U, V) = U + V - L_E^{-1}[L_{X-S}[U - V]].$$

(c) The matrix $E = |X - S| = ((X - S)^2)^{1/2}$ is nonsingular (or, equivalently, positive definite) if and only if the mapping $(A, B) \mapsto |A - B|$ is continuously differentiable at $(A, B) = (X, S)$.

For a proof of part (a) we refer to [22, 10], part (b) may be found in [3, 10], and part (c) follows from [18].

Part (a) may be used in order to reformulate the optimality conditions (1.2) as the nonsmooth system of equations

$$(2.3) \quad \Theta(X, \lambda, S) = 0,$$

where

$$\Theta : \mathcal{S}^{n \times n} \times \mathbb{R}^m \times \mathcal{S}^{n \times n} \rightarrow \mathcal{S}^{n \times n} \times \mathbb{R}^m \times \mathcal{S}^{n \times n}$$

is defined by

$$\Theta(X, \lambda, S) := \begin{pmatrix} \sum_{i=1}^m \lambda_i A_i + S - C \\ A_i \bullet X - b_i \quad (i = 1, \dots, m) \\ \phi(X, S) \end{pmatrix}.$$

A nonsmooth Newton method applied to the system (2.3) is an iterative method of the form

$$Z^{k+1} := Z^k + \Delta Z^k \quad \forall k = 0, 1, 2, \dots$$

where we used the abbreviation

$$Z^k := (X^k, \lambda^k, S^k),$$

and where ΔZ^k is a solution of the linearized equation $H^k \Delta Z^k = -\Theta(Z^k)$ with H^k being an element of the B-subdifferential of Θ in Z^k , i.e., $H^k \in \partial_B \Theta(Z^k)$, where

$$\partial_B \Theta(W) := \{H \mid \exists \{W^k\} \subset D_\Theta : W^k \rightarrow W, \nabla \Theta(W^k) \rightarrow H\}$$

and where D_Θ denotes the set of points at which Θ is differentiable; cf. [16]. Note that this set is always nonempty because Θ is locally Lipschitz. Moreover, we have $\partial_B \Theta(W^k) = \{\nabla \Theta(W^k)\}$ whenever Θ is continuously differentiable at W^k .

We formally state our method in the following algorithm.

ALGORITHM 2.2.

(S.0) Choose $Z^0 := (X^0, \lambda^0, S^0) \in \mathcal{S}^{n \times n} \times \mathbb{R}^m \times \mathcal{S}^{n \times n}$, $\varepsilon \geq 0$, and set $k := 0$.

(S.1) If $\|\Theta(Z^k)\| \leq \varepsilon$: STOP.

(S.2) Choose $H^k \in \partial_B \Theta(Z^k)$ and find a solution $\Delta Z^k = (\Delta X^k, \Delta \lambda^k, \Delta S^k)$ of the linearized system $H^k \Delta Z = -\Theta(Z^k)$.

(S.3) Set $Z^{k+1} := Z^k + \Delta Z^k$, $k \leftarrow k + 1$, and go to (S.1).

We note that the solution $\Delta Z^k = (\Delta X^k, \Delta \lambda^k, \Delta S^k)$ of the linearized equation in (S.2) automatically produces symmetric matrices ΔX^k and ΔS^k , see [3, 10]. The main local convergence result is as follows.

THEOREM 2.3. *Let $W^* := (X^*, \lambda^*, S^*)$ be a solution of the optimality conditions (1.2) and assume that all elements $H^* \in \partial_B \Theta(W^*)$ are invertible. Then Algorithm 2.2 is locally quadratically convergent.*

Proof. According to [18], the minimum function ϕ is strongly semismooth (see also [17, 16] for the definition and some properties of strongly semismooth mappings). Hence Θ is strongly semismooth. The local quadratic convergence result therefore follows from a general theorem in [16]; see also the recent books [4, 5] for more details on this subject. $\square$

In view of Theorem 2.3, the aim of this paper is to find suitable conditions under which the elements of the B-subdifferential $\partial_B \Theta(W^*)$ are invertible. This will be done in section 4 after some preliminary discussions in the following section.

3. Matrix-vector formulation of Newton system. Throughout this section, we assume that the mapping Θ is continuously differentiable at a current point $(X, \lambda, S) \in \mathcal{S}^{n \times n} \times \mathbb{R}^m \times \mathcal{S}^{n \times n}$. According to Proposition 2.1(b), (c), this assumption is satisfied if and only if the matrix $X - S$ is nonsingular. Later, in section 4, we will drop this assumption.

Since Θ is continuously differentiable at (X, λ, S) , Newton's method applied to $\Theta(X, \lambda, S) = 0$ has to solve the linearized equation

$$(3.1) \quad \nabla \Theta(X, \lambda, S)(\Delta X, \Delta \lambda, \Delta S) = -\Theta(X, \lambda, S)$$

at the current point. For our analysis in section 4, it will be useful to write this linearized system in the usual matrix-vector format. We therefore present such a formulation in this section.

Since Θ is continuously differentiable at (X, λ, S) , the minimum function ϕ is continuously differentiable at (X, S) . In view of Proposition 2.1(c), this means that the matrix

$$(3.2) \quad E := |X - S| = ((X - S)^2)^{1/2}$$

is positive definite. Consequently, the corresponding Lyapunov operator L_E is invertible. Defining the residuals

$$\begin{aligned} R_C &:= C - \sum_{j=1}^m \lambda_j A_j - S, \\ r_{b,i} &:= b_i - A_i \bullet X \quad (i = 1, \dots, m), \\ r_b &:= (r_{b,1}, \dots, r_{b,m})^T, \end{aligned}$$

the Newton system (3.1) can be rewritten as

$$(3.3) \quad \sum_{j=1}^m \Delta \lambda_j A_j + \Delta S = R_C,$$

$$(3.4) \quad A_i \bullet \Delta X = r_{b,i} \quad (i = 1, \dots, m),$$

$$(3.5) \quad \nabla \phi(X, S)(\Delta X, \Delta S) = -\phi(X, S).$$

In order to reformulate this system in the usual matrix-vector format, we need to transform matrices into vectors. For a general (not necessarily symmetric) matrix $A \in \mathbb{R}^{n \times n}$, this can be done by using the mapping $\text{vec} : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n^2}$ defined by

$$\text{vec}(A) := (a_{11}, a_{21}, \dots, a_{n1}, a_{12}, a_{22}, \dots, a_{n2}, \dots, a_{nn})^T \in \mathbb{R}^{n^2};$$

i.e., vec stacks the columns of A into a vector of length n^2 . For a symmetric matrix, we are not interested in all entries of A . It suffices to consider the lower triangular part of A , and the corresponding transformation can be done using the mapping $\text{svec} : \mathcal{S}^{n \times n} \rightarrow \mathbb{R}^{\frac{n(n+1)}{2}}$ defined by

$$\text{svec}(A) := (a_{11}, \sqrt{2}a_{21}, \dots, \sqrt{2}a_{n1}, a_{22}, \sqrt{2}a_{32}, \dots, \sqrt{2}a_{n2}, \dots, a_{nn})^T \in \mathbb{R}^{\frac{n(n+1)}{2}}.$$

The reason for the $\sqrt{2}$ factor in front of all nondiagonal elements is due to the fact that this is consistent with the inner product, i.e.,

$$(3.6) \quad A \bullet B = \text{svec}(A)^T \text{svec}(B) \quad \forall A, B \in \mathcal{S}^{n \times n}.$$

Having introduced vec and svec , the next question is how an ordinary matrix product can be expressed in terms of vec and svec . To this end, let us define the *Kronecker product* of two (not necessarily symmetric) matrices $G, K \in \mathbb{R}^{n \times n}$ by

$$G \otimes K := [g_{ij}K] = \begin{pmatrix} g_{11}K & \cdots & g_{1n}K \\ \vdots & \ddots & \vdots \\ g_{n1}K & \cdots & g_{nn}K \end{pmatrix} \in \mathbb{R}^{n^2 \times n^2}.$$

Then it can easily be verified that

$$(G \otimes K) \text{vec}(H) = \text{vec}(KHG^T) \quad (H \in \mathbb{R}^{n \times n}).$$

Similarly, we define the *symmetric Kronecker product* by

$$(3.7) \quad (G \otimes_s K) \text{svec}(H) := \frac{1}{2} \text{svec}(KHG^T + GHK^T) \quad (H \in \mathcal{S}^{n \times n}).$$

Some properties of the symmetric Kronecker product are summarized in the following result. The proofs of these properties are elementary. In fact, statements (a), (b), and (c) can be found in [2, 21], whereas part (d) is just a reformulation of statement (c).

LEMMA 3.1. *The symmetric Kronecker product $\otimes_s$ defined by (3.7) has the following properties:*

- (a) $G \otimes_s K = K \otimes_s G$.
- (b) If G and K are symmetric positive definite, then so is $G \otimes_s K$.
- (c) If G, K are two commuting symmetric matrices with eigenvalues $\sigma_1, \dots, \sigma_n$ and $\mu_1, \dots, \mu_n$, respectively, and if $q_1, \dots, q_n$ denotes a common set of orthonormal eigenvectors, then the $n(n+1)/2$ eigenvalues of $G \otimes_s K$ are given by

$$\frac{1}{2}(\sigma_i \mu_j + \mu_i \sigma_j) \quad (1 \leq j \leq i \leq n)$$

with the corresponding set of orthonormal eigenvectors v_{ij} ($1 \leq j \leq i \leq n$) defined by

$$v_{ij} := \begin{cases} \text{svec}(q_i q_i^T) & \text{if } i = j, \\ \frac{1}{\sqrt{2}} \text{svec}(q_i q_j^T + q_j q_i^T) & \text{if } j < i. \end{cases}$$

(d) If G, K are two commuting symmetric matrices with simultaneous spectral decompositions $G = QD_GQ^T$ and $K = QD_KQ^T$ for some orthogonal matrix $Q \in \mathbb{R}^{n \times n}$ and diagonal matrices $D_G, D_K \in \mathbb{R}^{n \times n}$, then $G \otimes_s K = (Q \otimes_s Q)(D_G \otimes_s D_K)(Q \otimes_s Q)^T$ is a spectral decomposition of $G \otimes_s K$ (in particular, $Q \otimes_s Q$ is an orthogonal matrix).

We now consider the Newton system (3.1), i.e., we consider the system (3.3)–(3.5). The first two equations (3.3) and (3.4) may be reformulated in matrix-vector notation in exactly the same way as described in [21], resulting in the two equations

$$(3.8) \quad \mathcal{A}^T \Delta \lambda + \text{svec}(\Delta S) = \text{svec}(R_C)$$

and

$$(3.9) \quad \mathcal{A} \text{svec}(\Delta X) = r_b,$$

respectively, where

$$(3.10) \quad \mathcal{A} := (\text{svec}(A_1), \dots, \text{svec}(A_m))^T \in \mathbb{R}^{m \times \frac{n(n+1)}{2}}.$$

Hence it remains to consider the third block (3.5). Using Proposition 2.1(b) and the definition of E from (3.2), (3.5) can be rewritten as

$$(3.11) \quad \Delta X + \Delta S - L_E^{-1}[L_{X-S}[\Delta X - \Delta S]] = -\phi(X, S).$$

Applying the Lyapunov operator L_E to both sides of (3.11) yields

$$L_E[\Delta X] + L_E[\Delta S] - L_{X-S}[\Delta X - \Delta S] = -L_E[\phi(X, S)].$$

Rearranging terms gives

$$L_{E-(X-S)}[\Delta X] + L_{E+(X-S)}[\Delta S] = -L_E[\phi(X, S)].$$

Using the notation

$$(3.12) \quad A_E := E - (X - S) \quad \text{and} \quad B_E := E + (X - S),$$

this equation may be rewritten as

$$(3.13) \quad L_{A_E}[\Delta X] + L_{B_E}[\Delta S] = -L_E[\phi(X, S)].$$

Applying $\frac{1}{2} \text{svec}$ to both sides then gives

$$\frac{1}{2} \text{svec}(L_{A_E}[\Delta X]) + \frac{1}{2} \text{svec}(L_{B_E}[\Delta S]) = -\frac{1}{2} \text{svec}(L_E[\phi(X, S)]).$$

Using the definition (3.7) of svec , we have the identity

$$\frac{1}{2} \text{svec}(L_A[H]) = \frac{1}{2} \text{svec}(AHI + IHA) = (I \otimes_s A) \text{svec}(H)$$

for all symmetric matrices $A, H \in \mathcal{S}^{n \times n}$. Setting

$$(3.14) \quad \mathcal{E} := I \otimes_s A_E \quad \text{and} \quad \mathcal{F} := I \otimes_s B_E,$$

we therefore get

$$(3.15) \quad \mathcal{E} \operatorname{svec}(\Delta X) + \mathcal{F} \operatorname{svec}(\Delta S) = -(I \otimes_s E) \operatorname{svec}(\phi(X, S)).$$

Since E and I are both positive definite, it follows from Lemma 3.1(b) that the matrix $(I \otimes_s E)$ is also positive definite and therefore, in particular, nonsingular. Hence (3.15) can be rewritten as

$$(3.16) \quad (I \otimes_s E)^{-1} \mathcal{E} \operatorname{svec}(\Delta X) + (I \otimes_s E)^{-1} \mathcal{F} \operatorname{svec}(\Delta S) = -\operatorname{svec}(\phi(X, S)).$$

Summarizing our discussion, we obtain the following result as a consequence of (3.8), (3.9), and (3.16).

THEOREM 3.2. *Let (X, λ, S) be given such that $X - S$ is nonsingular. Then the triple $(\Delta X, \Delta \lambda, \Delta S) \in \mathcal{S}^{n \times n} \times \mathbb{R}^m \times \mathcal{S}^{n \times n}$ satisfies the Newton system (3.1) if and only if the vector $(\operatorname{svec}(\Delta X), \Delta \lambda, \operatorname{svec}(\Delta S))$ satisfies the linear system of equations*

$$(3.17) \quad \begin{pmatrix} 0 & \mathcal{A}^T & \mathcal{I} \\ \mathcal{A} & 0 & 0 \\ (I \otimes_s E)^{-1} \mathcal{E} & 0 & (I \otimes_s E)^{-1} \mathcal{F} \end{pmatrix} \begin{pmatrix} \operatorname{svec}(\Delta X) \\ \Delta \lambda \\ \operatorname{svec}(\Delta S) \end{pmatrix} = \begin{pmatrix} \operatorname{svec}(R_C) \\ r_b \\ -\operatorname{svec}(\phi(X, S)) \end{pmatrix}.$$

Note that the linear system (3.17) looks very similar to the one obtained for interior-point methods in [21]; however, the reader should be careful because the matrices $\mathcal{E}$ and $\mathcal{F}$ have a different meaning here.

4. Local convergence analysis. Now let (X^*, λ^*, S^*) be a solution of the optimality conditions (1.2). In order to motivate our approach, we first assume that the following conditions hold; in particular, we assume that strict complementarity holds at this solution.

Assumption 4.1. Let (X^*, λ^*, S^*) be a solution of the optimality conditions (1.2).

(A.1) (Linear independence).

The matrices $A_1, \dots, A_m$ are linearly independent.

(A.2) (Nondegeneracy).

The following implication holds for any triple $(\Delta X, \Delta \lambda, \Delta S)$:

$$\left. \begin{aligned} \sum_{i=1}^m \Delta \lambda_i A_i + \Delta S &= 0, \\ A_i \bullet \Delta X &= 0 \quad (i = 1, \dots, m), \\ X^* \Delta S + \Delta X S^* &= 0 \end{aligned} \right\} \implies (\Delta X, \Delta S) = (0, 0).$$

(A.3) (Strict complementarity).

$X^* + S^* \succ 0$.

These conditions are quite standard in order to prove local fast convergence of several methods for the solution of semidefinite programs; see, for example, [2, 12, 3, 10]. Interestingly, it turns out that Assumption (A.3) is automatically satisfied if (A.1) and (A.2) hold. This was conjectured by one of the referees and, independently, by Florian Jarre [9]. Since this result is certainly of some interest, but not (explicitly) needed for our analysis, we present a proof in the appendix.

We next give some vector formulations of the conditions from Assumption 4.1.

LEMMA 4.2. *Let (X^*, λ^*, S^*) be a solution of the optimality conditions (1.2). Then the following statements hold:*

(a) *Assumption (A.1) is equivalent to the full rank of the matrix $\mathcal{A}$ from (3.10).*

(b) *Assumption (A.2) is equivalent to the following implication:*

$$\left. \begin{aligned} \mathcal{A}^T \Delta \lambda + \text{svec}(\Delta S) &= 0, \\ \mathcal{A} \text{svec}(\Delta X) &= 0, \\ (I \otimes_s X^*) \text{svec}(\Delta S) + (I \otimes_s S^*) \text{svec}(\Delta X) &= 0 \end{aligned} \right\} \implies \begin{cases} \text{svec}(\Delta X) = 0, \\ \text{svec}(\Delta S) = 0. \end{cases}$$

(c) *Assumption (A.2) is equivalent to the following implication:*

$$\left. \begin{aligned} \mathcal{A}^T \Delta \lambda + \text{svec}(\Delta S) &= 0, \\ \mathcal{A} \text{svec}(\Delta X) &= 0, \\ V^*(I \otimes_s D_{S^*})(V^*)^T \text{svec}(\Delta X) + \dots \\ V^*(I \otimes_s D_{X^*})(V^*)^T \text{svec}(\Delta S) &= 0 \end{aligned} \right\} \implies \begin{cases} \text{svec}(\Delta X) = 0, \\ \text{svec}(\Delta S) = 0. \end{cases}$$

Here, $X^* = QD_{X^*}Q^T$ and $S^* = QD_{S^*}Q^T$ denotes the simultaneous spectral decomposition of the two commuting matrices X^* , S^* , and $V^* := Q \otimes_s Q$.

Proof. (a) This follows directly from the definition of the matrix $\mathcal{A}$.

(b) By applying the svec operator, we obviously have

$$\sum_{i=1}^m \Delta \lambda_i A_i + \Delta S = 0 \iff \mathcal{A}^T \Delta \lambda + \text{svec}(\Delta S) = 0$$

and

$$A_i \bullet \Delta X = 0 \ (i = 1, \dots, m) \iff \mathcal{A} \text{svec}(\Delta X) = 0.$$

Hence it remains to show the equivalence

$$(4.1) \quad X^* \Delta S + \Delta X S^* = 0 \iff (I \otimes_s X^*) \text{svec}(\Delta S) + (I \otimes_s S^*) \text{svec}(\Delta X) = 0.$$

To this end, first assume that $X^* \Delta S + \Delta X S^* = 0$. By taking the transpose and recalling that $\Delta S, \Delta X$ are automatically symmetric, we obtain $\Delta S X^* + S^* \Delta X = 0$. Adding these two equations, applying $\frac{1}{2} \text{svec}$ to the resulting equation and using (3.7) gives the right-hand side formulation in (4.1). Conversely, assume that

$$(I \otimes_s X^*) \text{svec}(\Delta S) + (I \otimes_s S^*) \text{svec}(\Delta X) = 0$$

holds. In view of (3.7), this is equivalent to

$$X^* \Delta S + \Delta S X^* + S^* \Delta X + \Delta X S^* = 0.$$

Proceeding as in the proof of [13, Lemma 6.2] (in this reference, strict complementarity is assumed, but it is easy to see that this assumption can be avoided here), it follows that $X^* \Delta S + \Delta X S^* = 0$, so that (4.1) holds.

(c) This statement follows immediately from part (b) by applying Lemma 3.1(d). $\square$

Part (b) of Lemma 4.2 will be used in our subsequent nonsingularity result, whereas the reformulation given in statement (c) is presented here in order to have a better comparison between the nondegeneracy condition from (A.2) and the one to be introduced later.

We next show how our previous results may be used in order to prove the nonsingularity of the matrix from Theorem 3.2 at a strictly complementary solution.

THEOREM 4.3. *Let (X^*, λ^*, S^*) be a solution of the optimality conditions (1.2) satisfying Assumptions (A.1)–(A.3). Furthermore, define the matrices (cf. (3.12), (3.14))*

$$\begin{aligned} E^* &:= |X^* - S^*|, \\ \mathcal{E}^* &:= I \otimes_s A_{E^*} \quad \text{with } A_{E^*} := E^* - (X^* - S^*), \\ \mathcal{F}^* &:= I \otimes_s B_{E^*} \quad \text{with } B_{E^*} := E^* + (X^* - S^*). \end{aligned}$$

Then $(I \otimes_s E^*)^{-1}$ exists, and the matrix

$$\begin{pmatrix} 0 & \mathcal{A}^T & \mathcal{I} \\ \mathcal{A} & 0 & 0 \\ (I \otimes_s E^*)^{-1} \mathcal{E}^* & 0 & (I \otimes_s E^*)^{-1} \mathcal{F}^* \end{pmatrix}$$

(cf. (3.17)) is nonsingular.

Proof. Since (A.3) holds, the matrix E^* is positive definite. Hence $I \otimes_s E^*$ is also positive definite by Lemma 3.1(b). Therefore, the inverse $(I \otimes_s E^*)^{-1}$ exists, and we are done if we are able to show that the matrix

$$\begin{pmatrix} 0 & \mathcal{A}^T & \mathcal{I} \\ \mathcal{A} & 0 & 0 \\ \mathcal{E}^* & 0 & \mathcal{F}^* \end{pmatrix}$$

is nonsingular. To this end, let $(\text{svec}(\Delta X), \Delta \lambda, \text{svec}(\Delta S))$ be a given triple with

$$\begin{pmatrix} 0 & \mathcal{A}^T & \mathcal{I} \\ \mathcal{A} & 0 & 0 \\ \mathcal{E}^* & 0 & \mathcal{F}^* \end{pmatrix} \begin{pmatrix} \text{svec}(\Delta X) \\ \Delta \lambda \\ \text{svec}(\Delta S) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Blockwise this becomes

$$\begin{aligned} \mathcal{A}^T \Delta \lambda + \text{svec}(\Delta S) &= 0, \\ \mathcal{A} \text{svec}(\Delta X) &= 0, \\ \mathcal{E}^* \text{svec}(\Delta X) + \mathcal{F}^* \text{svec}(\Delta S) &= 0. \end{aligned} \tag{4.2}$$

Since $X^* S^* = 0$, the two matrices X^* , S^* commute. Hence there is an orthogonal matrix $Q^* \in \mathbb{R}^{n \times n}$ and two diagonal matrices $D_{X^*}, D_{S^*} \succeq 0$ such that

$$X^* = Q^* D_{X^*} (Q^*)^T \quad \text{and} \quad S^* = Q^* D_{S^*} (Q^*)^T; \tag{4.3}$$

cf. [7]. This implies

$$X^* S^* = 0 \iff D_{X^*} D_{S^*} = 0.$$

We therefore get

$$\begin{aligned} E^* &= |X^* - S^*| \\ &= ((X^* - S^*)^2)^{1/2} \\ &= Q((D_{X^*} - D_{S^*})^2)^{1/2} Q^T \\ &= Q(D_{X^*} + D_{S^*}) Q^T \\ &= X^* + S^*. \end{aligned}$$

This yields

$$A_{E^*} = E^* - (X^* - S^*) = 2S^* \quad \text{and} \quad B_{E^*} = E^* + (X^* - S^*) = 2X^*.$$

Consequently, we obtain

$$\mathcal{E}^* = I \otimes_s A_{E^*} = 2(I \otimes_s S^*) \quad \text{and} \quad \mathcal{F}^* = I \otimes_s B_{E^*} = 2(I \otimes_s X^*).$$

Hence the last line from (4.2) may be rewritten as

$$(I \otimes_s X^*) \text{svec}(\Delta S) + (I \otimes_s S^*) \text{svec}(\Delta X) = 0.$$

Together with the first two lines from (4.2), it now follows from Assumption (A.2) and Lemma 4.2(b) that $\text{svec}(\Delta X) = 0$ and $\text{svec}(\Delta S) = 0$. This, in turn, implies $\mathcal{A}^T \Delta \lambda = 0$ by (4.2). Since the columns of $\mathcal{A}^T$ are linearly independent by Lemma 4.2(a), we finally get $\Delta \lambda = 0$. $\square$

The previous result may also be obtained in a different way from [3, 10]; however, the formulation given here is more convenient for our subsequent generalization.

If (A.3) does not hold at $W^* = (X^*, \lambda^*, S^*)$, then the matrix E^* is not positive definite any longer. Hence $I \otimes_s E^*$ might be singular. Nevertheless, it is possible to generalize the previous discussion. To this end, we recall that any element $H^* \in \partial_B \Theta(W^*)$ may be obtained by taking a sequence $W^k := (X^k, \lambda^k, S^k)$ converging to W^* such that Θ is differentiable at each W^k and such that $H^* = \lim_{k \rightarrow \infty} \nabla \Theta(W^k)$. In view of Proposition 2.1(b), (c), however, Θ is continuously differentiable at W^k if and only if $X^k - S^k$ is nonsingular. But this is precisely the assumption we used in our analysis of section 3. Hence, using the corresponding notation (cf. (3.2), (3.12), and (3.14))

$$\begin{aligned} E^k &:= |X^k - S^k|, \\ \mathcal{E}^k &:= I \otimes_s A_{E^k} \quad \text{with } A_{E^k} := E^k - (X^k - S^k), \\ \mathcal{F}^k &:= I \otimes_s B_{E^k} \quad \text{with } B_{E^k} := E^k + (X^k - S^k) \end{aligned}$$

it follows from Theorem 3.2 that the usual matrix formulation of the Jacobian $\nabla \Theta(W^k)$ is given by

$$(4.4) \quad \begin{pmatrix} 0 & \mathcal{A}^T & \mathcal{I} \\ \mathcal{A} & 0 & 0 \\ (I \otimes_s E^k)^{-1} \mathcal{E}^k & 0 & (I \otimes_s E^k)^{-1} \mathcal{F}^k \end{pmatrix}.$$

We have to find the limiting elements of a sequence of matrices of this form. We therefore have to take a closer look at the last block row, i.e., at the limiting behavior of the two matrices

$$(I \otimes_s E^k)^{-1} \mathcal{E}^k = (I \otimes_s E^k)^{-1} (I \otimes_s A_{E^k})$$

and

$$(I \otimes_s E^k)^{-1} \mathcal{F}^k = (I \otimes_s E^k)^{-1} (I \otimes_s B_{E^k}).$$

To this end, let

$$X^k - S^k = Q^k \Pi^k (Q^k)^T \quad \text{with } \Pi^k = \text{diag}(\pi_1^k, \dots, \pi_n^k)$$

be a spectral decomposition of the symmetric matrix $X^k - S^k$, and let

$$|\Pi^k| = \text{diag}(|\pi_1^k|, \dots, |\pi_n^k|).$$

Then it is not difficult to see that

$$\begin{aligned} E^k &= Q^k |\Pi^k| (Q^k)^T, \\ A_{E^k} &= Q^k (|\Pi^k| - \Pi^k) (Q^k)^T, \\ B_{E^k} &= Q^k (|\Pi^k| + \Pi^k) (Q^k)^T. \end{aligned}$$

Let q_i^k be the i th column of Q^k and define the orthogonal matrix V^k by

$$V^k := (\dots, v_{ij}^k, \dots)_{1 \leq j \leq i \leq n}$$

with the columns $v_{ij}^k \in \mathbb{R}^{n(n+1)/2}$ given by

$$v_{ij}^k := \begin{cases} \text{svec}(q_i^k (q_i^k)^T) & \text{if } i = j, \\ \frac{1}{\sqrt{2}} \text{svec}(q_i^k (q_j^k)^T + q_j^k (q_i^k)^T) & \text{if } j < i. \end{cases}$$

Then Lemma 3.1(c) yields

$$\begin{aligned} I \otimes_s E^k &= V^k \text{diag} \left(\dots, \frac{1}{2} (|\pi_i^k| + |\pi_j^k|), \dots \right) (V^k)^T, \\ I \otimes_s A_{E^k} &= V^k \text{diag} \left(\dots, \frac{1}{2} (|\pi_i^k| + |\pi_j^k| - \pi_i^k - \pi_j^k), \dots \right) (V^k)^T, \\ I \otimes_s B_{E^k} &= V^k \text{diag} \left(\dots, \frac{1}{2} (|\pi_i^k| + |\pi_j^k| + \pi_i^k + \pi_j^k), \dots \right) (V^k)^T. \end{aligned}$$

Note that all diagonal elements π_i^k of Π^k are nonzero since E^k was assumed to be nonsingular, so that $|\pi_i^k| + |\pi_j^k| \neq 0$ for all $1 \leq j \leq i \leq n$. We therefore obtain

$$\begin{aligned} (I \otimes_s E^k)^{-1} (I \otimes_s A_{E^k}) &= V^k \text{diag} \left(\dots, \frac{|\pi_i^k| + |\pi_j^k| - \pi_i^k - \pi_j^k}{|\pi_i^k| + |\pi_j^k|}, \dots \right) (V^k)^T \\ &=: V^k \Sigma_-^k (V^k)^T, \\ (I \otimes_s E^k)^{-1} (I \otimes_s B_{E^k}) &= V^k \text{diag} \left(\dots, \frac{|\pi_i^k| + |\pi_j^k| + \pi_i^k + \pi_j^k}{|\pi_i^k| + |\pi_j^k|}, \dots \right) (V^k)^T \\ &=: V^k \Sigma_+^k (V^k)^T. \end{aligned}$$

Using the above identities we can rewrite the coefficient matrix from (4.4) as

$$\begin{pmatrix} 0 & \mathcal{A}^T & \mathcal{I} \\ \mathcal{A} & 0 & 0 \\ V^k \Sigma_-^k (V^k)^T & 0 & V^k \Sigma_+^k (V^k)^T \end{pmatrix}.$$

Subsequencing if necessary, we may assume without loss of generality that the orthogonal (and therefore bounded) matrix sequence $\{Q^k\}$ (and therefore $\{\Pi^k\}$) converge with

$$Q^* := \lim_{k \rightarrow \infty} Q^k \quad \text{and} \quad \Pi^* := \lim_{k \rightarrow \infty} \Pi^k.$$

Clearly, Q^* is again an orthogonal matrix such that the identity $X^* - S^* = Q^* \Pi^* (Q^*)^T$ holds, i.e., Q^* , Π^* correspond to a spectral decomposition of the symmetric matrix $X^* - S^*$. We also note that the corresponding subsequence of V^k converges to

$$(4.5) \quad V^* := (\dots, v_{ij}^*, \dots)_{1 \leq j \leq i \leq n},$$

where

$$(4.6) \quad v_{ij}^* := \begin{cases} \text{svec}(q_i^* (q_i^*)^T) & \text{if } i = j, \\ \frac{1}{\sqrt{2}} \text{svec}(q_i^* (q_j^*)^T + q_j^* (q_i^*)^T) & \text{if } j < i, \end{cases}$$

and q_i^* denotes the i th column of Q^* ; cf. Lemma 3.1(c). Furthermore, it is not difficult to see that all diagonal elements of the matrices Σ_-^k and Σ_+^k belong to the interval $[0, 2]$ and that $\Sigma_-^k + \Sigma_+^k = 2I$. In particular, we may also assume without loss of generality that

$$\Sigma_-^k \rightarrow \Sigma_-^* \quad \text{and} \quad \Sigma_+^k \rightarrow \Sigma_+^*$$

for some diagonal matrices Σ_-^* , Σ_+^* . Then we have

$$(4.7) \quad \begin{aligned} (I \otimes_s E^k)^{-1} (I \otimes_s A_{E^k}) &\rightarrow V^* \Sigma_-^* (V^*)^T, \\ (I \otimes_s E^k)^{-1} (I \otimes_s B_{E^k}) &\rightarrow V^* \Sigma_+^* (V^*)^T. \end{aligned}$$

In order to see the precise structure of the diagonal matrices Σ_-^* and Σ_+^* , let us write $\Pi^* = \text{diag}(\pi_1^*, \dots, \pi_n^*)$ and define the index sets

$$(4.8) \quad \begin{aligned} \alpha &:= \{i \mid \pi_i^* > 0\} = \{i \mid \lambda_i(X^*) > 0, \lambda_i(S^*) = 0\}, \\ \beta &:= \{i \mid \pi_i^* = 0\} = \{i \mid \lambda_i(X^*) = 0, \lambda_i(S^*) = 0\}, \\ \gamma &:= \{i \mid \pi_i^* < 0\} = \{i \mid \lambda_i(X^*) = 0, \lambda_i(S^*) > 0\}, \end{aligned}$$

where $\lambda_i(X^*)$ and $\lambda_i(S^*)$ denote the eigenvalues of X^* and S^* , respectively. Then an easy calculation shows that

$$\frac{|\pi_i^k| + |\pi_j^k| - \pi_i^k - \pi_j^k}{|\pi_i^k| + |\pi_j^k|} \rightarrow \sigma_{ij}^- \quad \text{with } \sigma_{ij}^- \begin{cases} > 0 & \text{if } i \in \gamma \text{ or } j \in \gamma, \\ \in [0, 2] & \text{if } i \in \beta \text{ and } j \in \beta, \\ 0 & \text{otherwise} \end{cases}$$

and, similarly,

$$\frac{|\pi_i^k| + |\pi_j^k| + \pi_i^k + \pi_j^k}{|\pi_i^k| + |\pi_j^k|} \rightarrow \sigma_{ij}^+ \quad \text{with } \sigma_{ij}^+ \begin{cases} > 0 & \text{if } i \in \alpha \text{ or } j \in \alpha, \\ \in [0, 2] & \text{if } i \in \beta \text{ and } j \in \beta, \\ 0 & \text{otherwise.} \end{cases}$$

In particular, we have

$$\begin{aligned} \sigma_{ij}^- &= 0 \quad \text{if } (i, j) \in (\alpha \times \alpha) \cup (\alpha \times \beta) \cup (\beta \times \alpha), \\ \sigma_{ij}^+ &= 0 \quad \text{if } (i, j) \in (\beta \times \gamma) \cup (\gamma \times \beta) \cup (\gamma \times \gamma). \end{aligned}$$

Further note that the limiting elements are not uniquely defined for all index pairs $(i, j) \in \beta \times \beta$, but that at least one of the elements σ_{ij}^- , σ_{ij}^+ is always positive.

Summarizing this discussion, we get the following result.

THEOREM 4.4. *Let (X^*, λ^*, S^*) be a solution of the optimality conditions (1.2), let $X^* - S^* = Q^* \Pi^* (Q^*)^T$ be a spectral decomposition of $X^* - S^*$, and let the index sets α, β, γ be defined as in (4.8). Then every element from the B -subdifferential $\partial_B \Theta(X^*, \lambda^*, S^*)$ can, in matrix-vector notation, be written as*

$$\begin{pmatrix} 0 & \mathcal{A}^T & \mathcal{I} \\ \mathcal{A} & 0 & 0 \\ V^* \Sigma_-^* (V^*)^T & 0 & V^* \Sigma_+^* (V^*)^T \end{pmatrix}$$

with V^* being the matrix from (4.5), (4.6) and with diagonal matrices

$$\begin{aligned} \Sigma_-^* &= \text{diag}(\dots, \sigma_{ij}^-, \dots)_{1 \leq j \leq i \leq n}, \\ \Sigma_+^* &= \text{diag}(\dots, \sigma_{ij}^+, \dots)_{1 \leq j \leq i \leq n} \end{aligned}$$

having the following properties:

$$(4.9) \quad \begin{aligned} \sigma_{ij}^-, \sigma_{ij}^+ &\geq 0, \quad \sigma_{ij}^- + \sigma_{ij}^+ > 0 \quad \text{if } (i, j) \in \beta \times \beta, \\ \sigma_{ij}^- &= 0 \quad \text{if } (i, j) \in (\alpha \times \alpha) \cup (\alpha \times \beta) \cup (\beta \times \alpha), \\ \sigma_{ij}^+ &= 0 \quad \text{if } (i, j) \in (\beta \times \gamma) \cup (\gamma \times \beta) \cup (\gamma \times \gamma), \end{aligned}$$

while all other diagonal entries are positive.

Motivated by these considerations, we next state a slightly modified nondegeneracy condition.

Assumption 4.5. Let (X^*, λ^*, S^*) be a solution of the optimality conditions (1.2), $X^* - S^* = Q^* \Pi^* (Q^*)^T$ be a spectral decomposition of $X^* - S^*$, α, β, γ the index sets from (4.8), and define the corresponding matrix V^* as in (4.5), (4.6).

(A.4) Given diagonal matrices

$$\begin{aligned} \Sigma_- &= \text{diag}(\dots, \sigma_{ij}^-, \dots)_{1 \leq j \leq i \leq n}, \\ \Sigma_+ &= \text{diag}(\dots, \sigma_{ij}^+, \dots)_{1 \leq j \leq i \leq n} \end{aligned}$$

satisfying (4.9) and having positive diagonal entries elsewhere, the following implication holds for any triple $(\text{svec}(\Delta X), \Delta \lambda, \text{svec}(\Delta S))$:

$$\left. \begin{aligned} \mathcal{A}^T \Delta \lambda + \text{svec}(\Delta S) &= 0 \\ \mathcal{A} \text{svec}(\Delta X) &= 0, \\ V^* \Sigma_- (V^*)^T \text{svec}(\Delta X) + V^* \Sigma_+ (V^*)^T \text{svec}(\Delta S) &= 0 \end{aligned} \right\} \implies \begin{cases} \text{svec}(\Delta X) = 0, \\ \text{svec}(\Delta S) = 0. \end{cases}$$

In order to see the difference between Assumption (A.4) and the nondegeneracy condition from Assumption (A.2), we first note that the precise values of the diagonal matrices Σ_- and Σ_+ from Assumption (A.4) are unimportant. The only important thing is whether a diagonal element is zero or positive. Taking this into account and using Lemma 4.2(c), it follows that Assumption (A.2) corresponds to the case where $\Sigma_- = I \otimes_s D_{S^*}$ and $\Sigma_+ = I \otimes_s D_{X^*}$, where we used the notation from Lemma 4.2. Hence, if strict complementarity holds, Assumptions (A.2) and (A.4) are equivalent. However, without strict complementarity, Assumption (A.2) cannot hold (see the appendix), whereas Example 5.1 shows that Assumption (A.4) might still be satisfied. We summarize these observations in the following result.

THEOREM 4.6. *Let $W^* = (X^*, \lambda^*, S^*)$ be a solution of the optimality conditions (1.2) satisfying (A.1). Then Assumption (A.4) is strictly weaker than Assumption (A.2). More precisely, Assumptions (A.2) and (A.4) are equivalent if strict complementarity holds, whereas Assumption (A.4) may hold without strict complementarity in contrast to Assumption (A.2).*

Our modified nondegeneracy condition from Assumption (A.4) will now be used in order to prove the following result.

THEOREM 4.7. *Let $W^* = (X^*, \lambda^*, S^*)$ be a solution of the optimality conditions (1.2) satisfying (A.1) and (A.4). Then all elements $H^* \in \partial_B \Theta(W^*)$ are invertible.*

Proof. In view of our construction, each element $H^* \in \partial_B \Theta(W^*)$ has, in the usual matrix representation, the form of a matrix as given in Theorem 4.4. Taking into account that the two diagonal matrices Σ_-^* and Σ_+^* from Theorem 4.4 satisfy the conditions of Assumption (A.4), we can now follow the technique of proof from Theorem 4.3 in order to see that this matrix is indeed nonsingular under (A.1) and (A.4). $\square$

As a direct consequence of Theorems 2.3 and 4.7, we obtain the following result.

THEOREM 4.8. *Let (X^*, λ^*, S^*) be a solution of the optimality conditions (1.2) satisfying (A.1) and (A.4). Then Algorithm 2.2 is locally quadratically convergent.*

The next result also follows from Theorem 4.7 together with a result from [16].

COROLLARY 4.9. *Let (X^*, λ^*, S^*) be a solution of the optimality conditions (1.2) satisfying (A.1) and (A.4). Then (X^*, λ^*, S^*) is the unique solution of (1.2).*

5. Illustrative examples. In this section, we illustrate our theory developed in the previous section by using two examples. The first example is taken from [1] and contains a semidefinite program with a unique solution not satisfying strict complementarity, but where our nondegeneracy condition from Assumption (A.4) holds.

Example 5.1. Let $n = 3, m = 3$ with $b = (1, 0, 0)^T$ and

$$C = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then Assumption (A.1) is obviously satisfied. Moreover, the corresponding semidefinite program has the unique solution

$$X^* = \text{diag}(1, 0, 0), \quad \lambda^* = (0, 0, 0)^T, \quad S^* = \text{diag}(0, 0, 1)$$

which does not satisfy the strict complementarity condition from (A.3). Moreover, it is not difficult to see that Assumption (A.2) is also violated. We now want to verify, however, that Assumption (A.4) holds. To this end, let

$$X^* - S^* = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix} = Q^* \Pi^* (Q^*)^T \quad \text{with } Q^* := I \in \mathbb{R}^{3 \times 3}$$

a spectral decomposition of $X^* - S^*$. Then

$$V^* = (v_{11}^*, v_{21}^*, v_{31}^*, v_{22}^*, v_{32}^*, v_{33}^*) = I \in \mathbb{R}^{6 \times 6}.$$

Let Σ_- , Σ_+ be two diagonal matrices as described in Assumption (A.4). Noting that the index sets from (4.8) are given by $\alpha = \{1\}$, $\beta = \{2\}$, $\gamma = \{3\}$, it follows that

$$(5.1) \quad \begin{aligned} \Sigma_- &= \text{diag}(0, 0, \sigma_{31}^-, \sigma_{22}^-, \sigma_{32}^-, \sigma_{33}^-), \\ \Sigma_+ &= \text{diag}(\sigma_{11}^+, \sigma_{21}^+, \sigma_{31}^+, \sigma_{22}^+, 0, 0) \end{aligned}$$

TABLE 1
Numerical results for Example 5.1.

k	$\ \Theta(W^k)\ $	$\ W^k - W^*\ $
0	0.4308903418625942	0.4931582715081313
1	0.0221433416930840	0.0571178343308808
2	0.0003449019008887	0.0066730984503571
3	0.0000000147308449	0.0000778721196135
4	0.0000000000000004	0.0000000105401872

for certain numbers satisfying

$$(5.2) \quad \begin{aligned} \sigma_{31}^-, \sigma_{32}^-, \sigma_{33}^- &> 0, \\ \sigma_{11}^+, \sigma_{21}^+, \sigma_{31}^+ &> 0, \\ \sigma_{22}^-, \sigma_{22}^+ &\geq 0, \quad \sigma_{22}^- + \sigma_{22}^+ > 0. \end{aligned}$$

Now let $\Delta X, \Delta S \in \mathcal{S}^{n \times n}$ and $\Delta \lambda \in \mathbb{R}^m$ satisfy the equations

$$(5.3) \quad \begin{aligned} \mathcal{A}^T \Delta \lambda + \text{svec}(\Delta S) &= 0, \quad \mathcal{A} \text{svec}(\Delta X) = 0, \\ V^* \Sigma_- (V^*)^T \text{svec}(\Delta X) + V^* \Sigma_+ (V^*)^T \text{svec}(\Delta S) &= 0, \end{aligned}$$

where

$$\mathcal{A} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{2} & 1 & 0 & 0 \\ 0 & \sqrt{2} & 0 & 0 & 0 & 1 \end{pmatrix};$$

cf. (3.10). The system of equations (5.3) is equivalent to

$$\begin{aligned} \Delta \lambda_1 + \Delta S_{11} &= 0, & \Delta \lambda_3 + \Delta S_{21} &= 0, \\ \Delta \lambda_2 + \Delta S_{31} &= 0, & \Delta \lambda_2 + \Delta S_{22} &= 0, \\ \Delta S_{32} &= 0, & \Delta \lambda_3 + \Delta S_{33} &= 0, \\ \Delta X_{11} &= 0, & 2\Delta X_{31} + \Delta X_{22} &= 0, \\ 2\Delta X_{21} + \Delta X_{33} &= 0, & \sigma_{11}^+ \Delta S_{11} &= 0, \\ \sigma_{21}^+ \Delta S_{21} &= 0, & \sigma_{31}^- \Delta X_{31} + \sigma_{31}^+ \Delta S_{31} &= 0, \\ \sigma_{22}^- \Delta X_{22} + \sigma_{22}^+ \Delta S_{22} &= 0, & \sigma_{32}^- \Delta X_{32} &= 0, \\ \sigma_{33}^- \Delta X_{33} &= 0. \end{aligned}$$

Depending on whether $\sigma_{22}^- > 0$ or $\sigma_{22}^+ > 0$, it is now an easy exercise to verify that these conditions imply $\Delta X = 0$ and $\Delta S = 0$, so that Assumption (A.4) holds.

In order to see that we really get fast local convergence, we applied Algorithm 2.2 to Example 5.1. To this end, we take $X^0 := X^*$, $\lambda^0 := (0.1, 0.1, 0.1)^T$, $S^0 := C - \sum_{i=1}^3 \lambda_i^0 A_i$ as a starting point (which may be viewed as a perturbation of the solution). The corresponding numerical results are given in Table 1. The columns of that table contain the norm of Θ at the current iterate $W^k = (X^k, \lambda^k, S^k)$ and the distance of W^k to the solution $W^* = (X^*, \lambda^*, S^*)$. The results from Table 1 clearly indicate the local quadratic convergence of our method.

The second example is taken from [12]. Also this example has a unique solution not satisfying strict complementarity. Here, however, Assumption (A.4) does not hold.

Example 5.2. Let $n = 3$, $m = 2$ with $b = (-1, 0)^T$ and

$$C = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad A_1 = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 & 0.5 \\ 0 & -0.5 & 0 \\ 0.5 & 0 & -1 \end{pmatrix}.$$

Assumption (A.1) obviously holds for this example. Moreover, the corresponding semidefinite program has the unique solution

$$X^* = \text{diag}(1, 0, 0), \quad \lambda^* = (0, 0)^T, \quad S^* = \text{diag}(0, 0, 1),$$

so that strict complementarity is violated. Similar to the previous example, we get $V^* = I$, $\alpha = \{1\}$, $\beta = \{2\}$, $\gamma = \{3\}$. Hence, if Σ_- , Σ_+ denote two matrices as described in Assumption (A.4), they satisfy (5.1), (5.2). Now let $(\Delta X, \Delta \lambda, \Delta S)$ any triple such that (5.3) holds with

$$A = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/\sqrt{2} & -0.5 & 0 & -1 \end{pmatrix}.$$

Componentwise, this may be rewritten as

$$\begin{aligned} -\Delta \lambda_1 + \Delta S_{11} &= 0, & \Delta S_{21} &= 0, \\ \frac{1}{\sqrt{2}} \Delta \lambda_2 + \sqrt{2} \Delta S_{31} &= 0, & -\frac{1}{2} \Delta \lambda_2 + \Delta S_{22} &= 0, \\ \Delta S_{32} &= 0, & -\Delta \lambda_2 + \Delta S_{33} &= 0, \\ -\Delta X_{11} &= 0, & \Delta X_{31} - \frac{1}{2} \Delta X_{22} - \Delta X_{33} &= 0, \\ \sigma_{11}^+ \Delta S_{11} &= 0, & \sigma_{21}^+ \Delta S_{21} &= 0, \\ \sigma_{31}^- \Delta X_{31} + \sigma_{31}^+ \Delta S_{31} &= 0, & \sigma_{22}^- \Delta X_{22} + \sigma_{22}^+ \Delta S_{22} &= 0, \\ \sigma_{32}^- \Delta X_{32} &= 0, & \sigma_{33}^- \Delta X_{33} &= 0 \end{aligned}$$

with certain elements σ_{ij}^- , σ_{ij}^+ satisfying (5.2). This is a homogenous linear system with 14 equations in 14 unknowns. Since the two equations $\Delta S_{21} = 0$ and $\sigma_{21}^+ \Delta S_{21} = 0$ are linearly dependent, this system has a nontrivial solution. Hence Assumption (A.4) does not hold in this case.

We also applied our nonsmooth Newton-type method to Example 5.2. The starting point was chosen as for the previous example with the only exception that λ^0 has two instead of three components. The corresponding numerical results are given in Table 2, beginning with iteration $k = 8$ (in the first iteration, our method goes far away from the solution and then comes back slowly to the region of interest). Table 2 indicates that our nonsmooth Newton method is only linearly convergent for this example. Since this example does not satisfy Assumption (A.4), this illustrates the importance of our modified nondegeneracy condition from (A.4).

6. Final remarks. We have shown that strict complementarity is not needed for a Newton-type method to be locally quadratically convergent when applied to a suitable reformulation of semidefinite programs. In order to obtain such a result, however, we had to introduce a modified nondegeneracy condition. This nondegeneracy condition was pointed out to be different from the one stated in Assumption (A.2), which was taken from Kojima, Shida, and Shindoh [12]. On the other hand, Alizadeh,

TABLE 2
Numerical results for Example 5.2.

k	$\ \Theta(W^k)\ $	$\ W^k - W^*\ $
8	0.0017244772625384	0.1029363671058705
9	0.0007359141856123	0.0710397847450809
10	0.0003522003507112	0.0487517513261899
11	0.0001655934643123	0.0335283142195323
12	0.0000783186561310	0.0230606035395086
13	0.0000370435896294	0.0158644262001553
14	0.0000175306616160	0.0109146131676810
15	0.0000082976229090	0.0075094716404199
16	0.0000039278073645	0.0051667563091444
17	0.0000018593651802	0.0035549233885654
18	0.0000008802133548	0.0024459312061810
19	0.0000004166919843	0.0016829029486340
20	0.0000001972627994	0.0011579083735320
21	0.0000000933847737	0.0007966901704891
22	0.0000000442080280	0.0005481577137891
23	0.0000000209273826	0.0003771581497793
24	0.0000000099063616	0.0002595043690389

Haerberly, and Overton [2] used another nondegeneracy condition which is known to be equivalent to the one from [12] if strict complementarity holds but is different without strict complementarity. Unfortunately, we do not know in how far the nondegeneracy condition from [2] is related to our Assumption (A.4).

Appendix A. Discussion of Assumptions (A.1)–(A.3). Here we show that Assumptions (A.1) and (A.2) together imply strict complementarity.

THEOREM A.1. *Let (X^*, λ^*, S^*) be a solution of the optimality conditions (1.2) satisfying Assumptions (A.1) and (A.2). Then $X^* + S^* \succ 0$, i.e., the solution satisfies the strict complementarity condition from Assumption (A.3).*

Proof. In view of (A.1) and Lemma 4.2(a), (b), we may rewrite Assumption (A.2) in the following way:

$$\left. \begin{aligned} \mathcal{A}^T \Delta \lambda + \text{svec}(\Delta S) &= 0, \\ \mathcal{A} \text{svec}(\Delta X) &= 0, \\ (I \otimes_s X^*) \text{svec}(\Delta S) + (I \otimes_s S^*) \text{svec}(\Delta X) &= 0 \end{aligned} \right\} \implies \begin{cases} \text{svec}(\Delta X) = 0, \\ \Delta \lambda = 0, \\ \text{svec}(\Delta S) = 0. \end{cases}$$

This is equivalent to saying that the matrix

$$\begin{pmatrix} 0 & \mathcal{A}^T & \mathcal{I} \\ \mathcal{A} & 0 & 0 \\ \mathcal{G} & 0 & \mathcal{H} \end{pmatrix}$$

with

$$\mathcal{G} := I \otimes_s S^*, \quad \mathcal{H} := I \otimes_s X^*$$

is nonsingular. Taking into account that X^* and S^* commute, it follows from standard properties of the symmetric Kronecker product (see [21]) that $\mathcal{G}$ and $\mathcal{H}$ commute,

too. Hence there is a simultaneous spectral decomposition, i.e., there is an orthogonal matrix $U \in \mathbb{R}^{\frac{n(n+1)}{2} \times \frac{n(n+1)}{2}}$ and diagonal matrices $D_X, D_S \in \mathbb{R}^{\frac{n(n+1)}{2} \times \frac{n(n+1)}{2}}$ such that

$$\mathcal{G} = UD_S U^T \quad \text{and} \quad \mathcal{H} = UD_X U^T.$$

We therefore have

$$\begin{pmatrix} 0 & \mathcal{A}^T & \mathcal{I} \\ \mathcal{A} & 0 & 0 \\ \mathcal{G} & 0 & \mathcal{H} \end{pmatrix} = \begin{pmatrix} \mathcal{I} & 0 & 0 \\ 0 & \mathcal{I} & 0 \\ 0 & 0 & U \end{pmatrix} \begin{pmatrix} 0 & \mathcal{A}^T & U \\ \mathcal{A}U & 0 & 0 \\ D_S & 0 & D_X \end{pmatrix} \begin{pmatrix} U^T & 0 & 0 \\ 0 & \mathcal{I} & 0 \\ 0 & 0 & U^T \end{pmatrix}.$$

Hence

$$(A.1) \quad \begin{pmatrix} 0 & \mathcal{A}^T & U \\ \mathcal{A}U & 0 & 0 \\ D_S & 0 & D_X \end{pmatrix}$$

is nonsingular in view of Assumptions (A.1) and (A.2).

Now suppose that (A.3) does not hold. Then there is a component $r \in \{1, \dots, n\}$ such that $\lambda_r(X^*) = 0$ and $\lambda_r(S^*) = 0$, where $\lambda_i(\cdot)$ denotes the i th eigenvalue of a given matrix. Since D_X and D_S are diagonal matrices with the eigenvalues of $\mathcal{H}$ and $\mathcal{G}$ on their diagonal, it follows from Lemma 3.1(c) that

$$\begin{aligned} D_X &= \frac{1}{2} \text{diag}(\dots, \lambda_j(X^*) + \lambda_i(X^*), \dots)_{1 \leq j \leq i \leq n}, \\ D_S &= \frac{1}{2} \text{diag}(\dots, \lambda_j(S^*) + \lambda_i(S^*), \dots)_{1 \leq j \leq i \leq n}. \end{aligned}$$

In particular, taking the pair $(j, i) = (r, r)$, we see that D_X and D_S have a zero entry in the same diagonal element. This implies that the matrix from (A.1) has a zero row, a contradiction to the nonsingularity of this matrix in view of Assumptions (A.1) and (A.2). $\square$

Acknowledgments. The authors would like to thank Florian Jarre for some insightful discussions regarding Assumptions (A.1)–(A.3). They also thank the referees for pointing out several improvements in the presentation of the material.

REFERENCES

- [1] F. ALIZADEH, J.-P. HAEBERLY, AND M.L. OVERTON, *Complementarity and nondegeneracy in semidefinite programming*, Math. Programming, 77 (1997), pp. 111–128.
- [2] F. ALIZADEH, J.-P. A. HAEBERLY, AND M.L. OVERTON, *Primal-dual interior-point methods for semidefinite programming: Convergence rates, stability and numerical results*, SIAM J. Optim., 8 (1998), pp. 746–768.
- [3] X. CHEN AND P. TSENG, *Non-interior continuation methods for solving semidefinite complementarity problems*, Math. Program., 95 (2003), pp. 431–474.
- [4] F. FACCHINEI AND J.-S. PANG, *Finite-Dimensional Variational Inequalities and Complementarity Problems I*, Springer-Verlag, New York, 2003.
- [5] F. FACCHINEI AND J.-S. PANG, *Finite-Dimensional Variational Inequalities and Complementarity Problems II*, Springer-Verlag, New York, 2003.
- [6] C. HELMBERG, F. RENDL, R. VANDERBEI, AND H. WOLKOWICZ, *An interior-point method for semidefinite programming*, SIAM J. Optim., 6 (1996), pp. 342–361.
- [7] R.A. HORN AND C.R. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

- [8] R.A. HORN AND C.R. JOHNSON, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.
- [9] F. JARRE, *private communication*, December 2003.
- [10] C. KANZOW AND C. NAGEL, *Semidefinite programs: New search directions, smoothing-type methods, and numerical results*, SIAM J. Optim., 13 (2002), pp. 1–23. (Corrigendum in SIAM J. Optim., SIAM J. Optim., 14 (2004), pp. 936–937.)
- [11] C. KANZOW AND C. NAGEL, *Some structural properties of a Newton-type method for semidefinite programs*, J. Optim. Theory Appl., 122 (2004), pp. 219–226.
- [12] M. KOJIMA, M. SHIDA, AND S. SHINDOH, *Local convergence of predictor-corrector infeasible-interior-point algorithms for SDPs and SDLCPs*, Math. Programming, 80 (1998), pp. 129–160.
- [13] M. KOJIMA, M. SHIDA, AND S. SHINDOH, *A predictor-corrector interior-point algorithm for the semidefinite linear complementarity problem using the Alizadeh-Haeberly-Overton search direction*, SIAM J. Optim., 9 (1999), pp. 444–465.
- [14] R.D.C. MONTEIRO, *Polynomial convergence of primal-dual algorithms for semidefinite programming based on the Monteiro and Zhang family of directions*, SIAM J. Optim., 8 (1998), pp. 797–812.
- [15] Y. NESTEROV AND M.J. TODD, *Primal-dual interior-point methods for self-scaled cones*, SIAM J. Optim., 8 (1998), pp. 324–364.
- [16] L. QI, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res., 18 (1993), pp. 227–244.
- [17] L. QI AND J. SUN, *A nonsmooth version of Newton's method*, Math. Programming, 58 (1993), pp. 353–367.
- [18] D. SUN AND J. SUN, *Semismooth matrix valued functions*, Math. Oper. Res., 27 (2002), pp. 150–169.
- [19] D. SUN, J. SUN, AND L. QI, *From Strong Semismoothness of the Squared Smoothing Matrix Function to Semidefinite Complementarity Problems*, Applied Mathematics Report AMR 00/20, School of Mathematics, University of New South Wales, Sydney, Australia, 2000.
- [20] J. SUN, D. SUN, AND L. QI, *A squared smoothing Newton method for nonsmooth matrix equations and its applications in semidefinite optimization problems*, SIAM J. Optim., 14 (2004), pp. 783–806 (revision of [19]).
- [21] M.J. TODD, K.C. TOH, AND R.H. TÜTÜNCÜ, *On the Nesterov-Todd direction in semidefinite programming*, SIAM J. Optim., 8 (1998), pp. 769–796.
- [22] P. TSENG, *Merit functions for semi-definite complementarity problems*, Math. Programming, 83 (1998), pp. 159–185.
- [23] Y. ZHANG, *On extending some primal-dual interior-point algorithms from linear programming to semidefinite programming*, SIAM J. Optim., 8 (1998), pp. 365–386.

AN ALGORITHM FOR DEGENERATE NONLINEAR PROGRAMMING WITH RAPID LOCAL CONVERGENCE*

STEPHEN J. WRIGHT†

Abstract. This paper describes and analyzes an algorithmic framework for solving nonlinear programming problems in which strict complementarity conditions and constraint qualifications are not necessarily satisfied at a solution. The framework is constructed from three main algorithmic ingredients. The first is any conventional method for nonlinear programming that produces estimates of the Lagrange multipliers at each iteration; the second is a technique for estimating the set of active constraint indices; the third is a stabilized Lagrange–Newton algorithm with rapid local convergence properties. Results concerning rapid local convergence and global convergence of the proposed framework are proved. The approach improves on existing approaches in that less restrictive assumptions are needed for convergence and/or the computational workload at each iteration is lower.

Key words. nonlinear programming, degeneracy, superlinear convergence

AMS subject classifications. 90C30, 90C55, 65K05, 49M05

DOI. 10.1137/030601235

1. Introduction. Consider the following nonlinear programming problem with inequality constraints:

$$(1.1) \quad \text{NLP:} \quad \min_z \phi(z) \quad \text{subject to } g(z) \leq 0,$$

where $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ are twice continuously differentiable functions. In this paper we describe an algorithmic framework for this problem that converges superlinearly to a solution z^* under mild assumptions on the functions ϕ and g in the neighborhood of z^* . Rapid convergence can be proved even for difficult problems, such as those with complementarity constraints or with equality constraints of the form $h(x) = 0$ that have been split into two inequalities $h(x) \leq 0$, $h(x) \geq 0$. The use of less-than-ideal formulation techniques such as the latter is well known in the case of linear programming, for which production software contains presolvers or preprocessors that remove many such infelicities before they can cause difficulties for the underlying algorithm. Since for nonlinear programming it is often difficult to detect and remedy poor formulations at the level of the modeling language or the mathematical formulation, we should aim to design algorithms that perform as well as possible in a variety of difficult circumstances.

Optimality conditions for (1.1) can be derived from the Lagrangian for (1.1), which is

$$(1.2) \quad \mathcal{L}(z, \lambda) = \phi(z) + \lambda^T g(z),$$

where $\lambda \in \mathbb{R}^m$ is the vector of Lagrange multipliers. The Karush–Kuhn–Tucker

*Received by the editors October 16, 2003; accepted for publication (in revised form) June 9, 2004; published electronically April 8, 2005. This research was supported in part by the National Science Foundation under grants MTS-0086559, ACI-0196485, and EIA-0127857.

<http://www.siam.org/journals/siopt/15-3/60123.html>

†Computer Sciences Department, University of Wisconsin, 1210 W. Dayton Street, Madison, WI 53706 (swright@cs.wisc.edu).

(KKT) conditions are satisfied at a point (z^*, λ^*) if

$$(1.3a) \quad \nabla_z \mathcal{L}(z^*, \lambda^*) = \nabla \phi(z^*) + \sum_{i=1}^m \lambda_i^* \nabla g_i(z^*) = 0,$$

$$(1.3b) \quad 0 \geq g(z^*) \perp \lambda^* \geq 0,$$

where the notation “ $\perp$ ” signifies that $g(z^*)^T \lambda^* = 0$. We call a point z^* at which these relations are satisfied a *KKT point* or *stationary point*. The set $\mathcal{A}^*$ of active constraints at z^* is defined as follows:

$$(1.4) \quad \mathcal{A}^* = \{i = 1, 2, \dots, m \mid g_i(z^*) = 0\}.$$

In this paper, we develop an algorithmic framework (“Framework INEQ”) that exhibits local superlinear convergence under mild assumptions. It dispenses with constraint qualifications altogether and does not require strict complementarity. It assumes only that a KKT point exists and that a certain second-order sufficient condition is satisfied at this point. Framework INEQ incorporates an “outer strategy” whose purpose is to ensure good global convergence properties and an “equality constrained phase” (the “EQ phase”) which is activated when the set of active constraints has apparently been identified. We do not dwell on the outer strategy further; techniques involving merit functions, filters, and a variety of algorithms for generating steps can be used for this purpose. Rather, the focus of this paper is on the EQ phase and Framework INEQ. In the EQ phase, a “stabilized Lagrange–Newton” algorithm (Algorithm SLN) is applied to the optimality conditions for the subproblem obtained by enforcing the apparently active constraints as equalities. Algorithm SLN is described in section 4.

The cost of each iteration in the EQ phase of Framework INEQ is essentially the same as in most nonlinear programming algorithms that use second derivatives, except for the solution at some iterations of an ancillary linear program. In particular, computation of the step requires solution of a single linear system whose structure and sparsity are almost identical to the well-known “augmented system” matrix that arises in many interior-point and sequential quadratic programming approaches. Our approach includes various checks (which are inexpensive, apart from the linear program mentioned above) to ensure that the algorithm does not get trapped at a nonoptimal point because of incorrect identification of the active set.

The remainder of the paper is structured as follows. In the remainder of this section, we summarize the previous literature on superlinearly convergent algorithms based on weaker-than-usual assumptions, and compare it to the approach of this paper. In section 2, we give some background on optimality conditions for (1.1) and for the equality constrained nonlinear optimization problem, focusing in particular on the critical cone that is used in the specification of the second-order condition. Section 3 discusses computable estimates of the distance of a given point (z, λ) to the primal-dual solution set for (1.1) in the vicinity of a point satisfying the second-order condition. Algorithm SLN for the equality constrained problem is presented and analyzed in section 4, and we comment on the similarity of the linear systems solved at each iteration of this approach to the corresponding systems for the quadratic penalty and augmented Lagrangian methods. We present Framework INEQ for solving (1.1) in section 5, along with local and global convergence results. Finally, we discuss some consequences of the second-order assumption in section 6, and outline the behavior of the approach on a particular degenerate problem.

1.1. Previous work. Other work on algorithms that retain rapid local convergence properties under weakened conditions has been performed by Fischer [5, 6], Wright [16, 15], Hager [8], and Izmailov and Solodov [11]. Wright [15] proposed a stabilized sequential quadratic programming method (“stabilized SQP”) in which a linear-quadratic minimax problem is solved at each iteration. Rapid local convergence is proved for the case in which the Mangasarian–Fromovitz constraint qualification (MFCQ) and strict complementarity hold. This approach also required a starting point that is not too close to the “relative boundary” of the primal-dual solution set. By using a more restrictive second-order condition, Hager [8] dispensed with the constraint qualifications and strict complementarity assumptions. He proved rapid convergence provided that the primal-dual starting point is located in a neighborhood of a particular primal-dual solution satisfying his second-order condition. Fischer [5] proposed an algorithm in which an additional quadratic program is solved between iterations of SQP in order to adjust the Lagrange multiplier estimate. He proved superlinear convergence under conditions that are weaker than the standard nondegeneracy assumptions, but stronger than the ones made in this paper. In [6], Fischer described a more general framework for generalized equations that includes the stabilized SQP approach as a special case. By using assumptions similar to those of Hager [8] (in particular, the stronger second-order condition), he proves local superlinear convergence.

Wright [16] described superlinear local convergence properties of a class of inexact SQP methods and showed that the stabilized SQP approach of [15, 8] and Fischer’s method [5] could be expressed as members of this class. This paper also introduced a modification of standard SQP that enforced only a subset of the linearized constraints—those in a “strictly active working set”—and permitted slight violations of the unenforced constraints. Still it achieved superlinear convergence under weaker-than-usual conditions. More recently, Wright [17] described a constraint identification procedure that distinguishes between weakly active constraints (those for which the Lagrange multiplier component is zero, for all Lagrange multiplier vectors satisfying the KKT conditions) and strongly active constraints. Rapid local convergence of a stabilized SQP approach based on this identification technique is proved under MFCQ and a certain second-order condition.

The approach of Izmailov and Solodov [11] is similar to the one described here in that it uses explicit estimates of the active constraint set $\mathcal{A}^*$ and crossover to an equality constrained phase at the final approach to the solution. The second-order assumption of [11] is the same as those used here, although we also consider situations in which these conditions hold uniformly on a compact subset of the set of multipliers satisfying the KKT conditions. The main difference between Framework INEQ and the approach of [11] is that the latter requires computation of a matrix that projects (approximately) onto the kernel of the active constraint Jacobian. This projection can be quite expensive to compute in practice; the singular-value-decomposition approach presented in [11] is not practical for large problems as it yields a dense matrix in general even when the constraint Jacobian is sparse. By contrast, the linear system to be solved at each iteration of our equality constrained phase is quite similar to those that are solved at each iteration of a conventional SQP or primal-dual interior-point method. Another difference is that Framework INEQ includes checks that detect errors in the estimation of the active set. These checks are essential to preventing convergence to a nonoptimal point and in yielding a global convergence result.

We believe that by choosing an appropriate outer strategy, the approach of this paper can be incorporated readily into algorithms (and software) with good global

convergence properties, thereby enhancing the performance of such algorithms in problems that exhibit degeneracy at the solution.

2. Assumptions, notation, and basic results. We now review the optimality conditions for (1.1) and its equality constrained counterpart, and outline the assumptions and notation that are used in subsequent sections. We treat the inequality constrained problem (1.1) in section 2.1 and the equality constrained problem in section 2.2.

2.1. The inequality constrained problem. Recall the KKT conditions (1.3). The set of “optimal” Lagrange multipliers λ^* is denoted by $\mathcal{S}_\lambda$, and the primal-dual optimal set is denoted by $\mathcal{S}$. Specifically, we have

$$(2.1) \quad \mathcal{S}_\lambda \stackrel{\text{def}}{=} \{\lambda^* \mid \lambda^* \text{ satisfies (1.3)}\}, \quad \mathcal{S} \stackrel{\text{def}}{=} \{z^*\} \times \mathcal{S}_\lambda.$$

Given the optimal active set $\mathcal{A}^*$, defined in (1.4), we define several related index sets. For any optimal multiplier $\lambda^* \in \mathcal{S}_\lambda$, we define the set $\mathcal{A}_+^*(\lambda^*)$ to be the “support” of λ^* , that is,

$$\mathcal{A}_+^*(\lambda^*) = \{i \in \mathcal{A}^* \mid \lambda_i^* > 0\}.$$

We define $\mathcal{A}_+^*$ (without argument) as

$$(2.2) \quad \mathcal{A}_+^* \stackrel{\text{def}}{=} \cup_{\lambda^* \in \mathcal{S}_\lambda} \mathcal{A}_+^*(\lambda^*);$$

this set contains the indices of the *strongly active* constraints. Its complement in $\mathcal{A}^*$ is denoted by $\mathcal{A}_0^*$, that is,

$$\mathcal{A}_0^* \stackrel{\text{def}}{=} \mathcal{A}^* \setminus \mathcal{A}_+^*.$$

This set $\mathcal{A}_0^*$ contains the *weakly active* constraint indices, those indices $i \in \mathcal{A}^*$ such that $\lambda_i^* = 0$ for all $\lambda^* \in \mathcal{S}_\lambda$. It is easy to show, using convexity of $\mathcal{S}_\lambda$, that there exists a particular vector $\hat{\lambda} \in \mathcal{S}_\lambda$ such that

$$(2.3) \quad \hat{\lambda}_i = 0 \text{ for all } i \notin \mathcal{A}_+^*; \quad \hat{\lambda}_i > 0 \text{ for all } i \in \mathcal{A}_+^*.$$

We define the following second-order sufficient condition at some point $(z^*, \lambda^*) \in \mathcal{S}$: There is a scalar $v > 0$ such that

$$(2.4) \quad \begin{aligned} & w^T \nabla_{zz}^2 \mathcal{L}(z^*, \lambda^*) w \geq v \|w\|^2, \\ \text{for all } w \text{ such that } & \begin{aligned} & \nabla g_i(z^*)^T w = 0 \quad \text{for all } i \in \mathcal{A}_+^*, \\ & \nabla g_i(z^*)^T w \leq 0 \quad \text{for all } i \in \mathcal{A}_0^*. \end{aligned} \end{aligned}$$

Condition 2s.1 in [16, Section 3] assumes that (2.4) holds for *all* $\lambda^* \in \mathcal{S}_\lambda$, for a fixed $v > 0$. In other works (see Hager and Gowda [9] and Fischer [6, Lemma 5]), the condition on $\nabla_{zz}^2 \mathcal{L}(z^*, \lambda^*)$ in (2.4) is assumed to hold at just one multiplier $\lambda^* \in \mathcal{S}_\lambda$, but for all w satisfying the condition

$$(2.5) \quad \begin{aligned} & \nabla g_i(z^*)^T w = 0 \quad \text{for all } i \in \mathcal{A}_+^*(\lambda^*), \\ & \nabla g_i(z^*)^T w \leq 0 \quad \text{for all } i \in \mathcal{A}^* \setminus \mathcal{A}_+^*(\lambda^*). \end{aligned}$$

Since $\mathcal{A}_+^*(\lambda^*) \subset \mathcal{A}_+^*$ for all $\lambda^* \in \mathcal{S}_\lambda$, we might expect that the set of vectors w defined by (2.5) is in general a superset of the set defined in (2.4). In fact, the two sets are identical, as the following simple result shows.

LEMMA 2.1. *The direction sets defined in (2.4) and (2.5) coincide, for any vector $\lambda^* \in \mathcal{S}_\lambda$.*

Proof. It is clear that the direction set in (2.4) is a subset of (2.5), because $\mathcal{A}_+^*(\lambda^*) \subset \mathcal{A}_+^*$. To prove the reverse inclusion, we use the multiplier $\hat{\lambda} \in \mathcal{S}_\lambda$ defined in (2.3). For w satisfying (2.5), we have from (1.3) that

$$\sum_{i \in \mathcal{A}_+^*(\lambda^*)} \lambda_i^* \nabla g_i(z^*) = -\nabla \phi(z^*) = \sum_{i \in \mathcal{A}_+^*} \hat{\lambda}_i \nabla g_i(z^*).$$

Taking inner products of these expressions with w , and using the first relation in (2.5), we have

$$0 = \sum_{i \in \mathcal{A}_+^*} \hat{\lambda}_i w^T \nabla g_i(z^*).$$

Then, using the fact that $\hat{\lambda}_i > 0$ and $w^T \nabla g_i(z^*) \leq 0$ for all $i \in \mathcal{A}_+^*$, it follows that in fact $w^T \nabla g_i(z^*) = 0$ for all $i \in \mathcal{A}_+^*$. We also have from (2.5) that $w^T \nabla g_i(z^*) \leq 0$ for all $i \in \mathcal{A}_0^* = \mathcal{A}^* \setminus \mathcal{A}_+^*$. Hence, w satisfies the conditions in (2.4), as claimed. $\square$

A less restrictive second-order condition, stated in terms of a quadratic growth condition of the objective $\phi(z)$ in a feasible neighborhood of z^* , is discussed by Bonnans and Ioffe [3] and Anitescu [1].

Our standing assumption for the problem (1.1) in this paper is as follows.

ASSUMPTION 1. *The KKT conditions (1.3) hold at z^* (that is, $\mathcal{S} \neq \emptyset$) and the functions ϕ and g are twice Lipschitz continuously differentiable in a neighborhood of z^* .*

In the following result, our claim that z^* is a strict local minimizer means that there exists a neighborhood of z^* such that $f(z^*) \leq f(z)$ for all z in this neighborhood with $g(z) \leq 0$, and that this inequality is strict if $z \neq z^*$.

THEOREM 2.2. *Suppose that Assumption 1 holds and that the second-order condition (2.4) is satisfied at some $\lambda^* \in \mathcal{S}_\lambda$. Then z^* is a strict local minimizer of (1.1).*

Proof. For the proof, see Robinson [14, Theorem 2.2]. $\square$

2.2. The equality constrained problem. Consider now the nonlinear programming problem with equality constraints defined by

$$(2.6) \quad \min_z \phi(z) \quad \text{subject to } h(z) = 0,$$

where $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$ and $h : \mathbb{R}^n \rightarrow \mathbb{R}^p$ are twice continuously differentiable functions. We denote the Lagrangian for (2.6) as follows:

$$(2.7) \quad \bar{\mathcal{L}}(z, \gamma) \stackrel{\text{def}}{=} \phi(z) + \gamma^T h(z).$$

The KKT conditions for (2.6) at z^* are that there exists a vector $\gamma^* \in \mathbb{R}^p$ such that

$$(2.8) \quad \nabla_z \bar{\mathcal{L}}(z^*, \gamma^*) = \nabla \phi(z^*) + \sum_{i=1}^p \gamma_i^* \nabla h_i(z^*) = 0.$$

The second-order sufficient condition is satisfied at some multiplier γ^* from (2.8) if there exists $v > 0$ such that

$$(2.9) \quad \begin{aligned} & w^T \nabla^2 \bar{\mathcal{L}}(z^*, \gamma^*) w \geq v \|w\|^2, \\ & \text{for all } w \text{ such that } \nabla h_i(z^*)^T w = 0, \quad i = 1, 2, \dots, p. \end{aligned}$$

Finally, we define

$$(2.10) \quad \bar{\mathcal{S}}_\gamma \stackrel{\text{def}}{=} \{\gamma^* \mid \gamma^* \text{ satisfies (2.8)}\}, \quad \bar{\mathcal{S}} \stackrel{\text{def}}{=} \{z^*\} \times \bar{\mathcal{S}}_\gamma.$$

ASSUMPTION 2. *The KKT conditions (1.3) are satisfied at z^* (that is, $\bar{\mathcal{S}} \neq \emptyset$) and the functions ϕ and h are twice Lipschitz continuously differentiable in a neighborhood of z^* .*

Similarly to Theorem 2.2, we have that z^* is a strict local solution of (2.6) when this assumption is satisfied and the second-order condition (2.9) holds for some multiplier γ^* satisfying (2.8).

2.3. Notation. We define the distance of a vector $x \in \mathbb{R}^r$ to a set $X \in \mathbb{R}^r$ by

$$\text{dist}(x, X) = \inf_{\bar{x} \in X} \|\bar{x} - x\|,$$

where here and elsewhere, $\|\cdot\|$ denotes the Euclidean norm unless a subscript specifically indicates otherwise.

We use order notation in the following (fairly standard) way: If two matrix, vector, or scalar quantities M and A are functions of a common quantity, we write $M = O(\|A\|)$ if there is a constant β such that $\|M\| \leq \beta\|A\|$ whenever $\|A\|$ is sufficiently small.

For a vector $\lambda \in \mathbb{R}^m$, we denote by $\lambda_{\mathcal{A}}$ the subvector whose components are λ_i , $i \in \mathcal{A}$. We denote by $g_{\mathcal{A}}(z)$ the subvector of the vector function $g(z)$ whose components are $g_i(z)$, $i \in \mathcal{A}$.

3. Distance-to-solution estimates. In this section we describe simple formulae for estimating distance to the solution of (1.1) (or (2.6)) from a given primal-dual point (z, λ) (or (z, γ)). We are interested in two-sided estimates; that is, those that are bounded both above and below by the true distance to the solution. These estimates are used by the methods of sections 4 and 5 both to decide on acceptability of a step and as a parameter in the step computation.

We aim for results that hold in a neighborhood of a compact subset of $\mathcal{S}_\lambda$ (or $\bar{\mathcal{S}}_\gamma$) at which the second-order condition is satisfied.

The following result, which makes use of Assumption 1 and (2.4) at a particular $\lambda^* \in \mathcal{S}_\lambda$, gives a practical estimate of $\text{dist}((z, \lambda), \mathcal{S})$. We use $\min(\lambda, -g(z))$ here to denote the vector whose i th component is $\min(\lambda_i, -g_i(z))$.

THEOREM 3.1. *Suppose that Assumption 1 holds at z^* , and that the second-order condition (2.4) is satisfied at some $\lambda^* \in \mathcal{S}_\lambda$. Then there are positive quantities $\delta_0(\lambda^*)$ and $\beta(\lambda^*)$ such that for all (z, λ) with $\|(z, \lambda) - (z^*, \lambda^*)\| \leq \delta_0(\lambda^*)$, the quantity $\eta(z, \lambda)$ defined by*

$$(3.1) \quad \eta(z, \lambda) \stackrel{\text{def}}{=} \left\| \begin{bmatrix} \nabla_z \mathcal{L}(z, \lambda) \\ \min(\lambda, -g(z)) \end{bmatrix} \right\|_1$$

satisfies

$$\eta(z, \lambda) \in [1/\beta(\lambda^*), \beta(\lambda^*)] \text{dist}((z, \lambda), \mathcal{S}).$$

Proof. One side of the result, namely,

$$\eta(z, \lambda) \leq \beta(\lambda^*) \text{dist}((z, \lambda), \mathcal{S}),$$

is relatively trivial; for a proof see Wright [16]. The other inequality, namely,

$$\text{dist}((z, \lambda), \mathcal{S}) \leq \beta(\lambda^*)\eta(z, \lambda)$$

has been examined by a number of authors under different assumptions; see, for example, Facchinei, Fischer, and Kanzow [4, Theorem 3.6], Wright [16, Theorem A.1], Hager and Gowda [9], and Fischer [6]. Under the assumptions used here, the result follows from Theorem 2 of Fischer [6], with $\Sigma_0 = \{(z^*, \lambda^*)\}$ in the notation of [6], where we use [9, Lemma 2] to justify [6, Assumption 1], and also Lemma 2.1 above to verify that the critical direction set of [9, Lemma 2] is the same as the direction set defined in (2.4). $\square$

We now extend Theorem 3.1 to the case in which (2.4) holds for all λ^* in a compact subset of $\mathcal{S}_\lambda$.

THEOREM 3.2. *Suppose that Assumption 1 holds for the problem (1.1) at z^* , and that the second-order condition (2.4) holds for all $\lambda^* \in \mathcal{V}_\lambda$, where $\mathcal{V}_\lambda$ is a compact subset of $\mathcal{S}_\lambda$. Let us define $\mathcal{V} = \{z^*\} \times \mathcal{V}_\lambda$. Then there are positive constants $\delta_1(\mathcal{V}_\lambda)$ and $\beta(\mathcal{V}_\lambda)$ such that for all (z, λ) satisfying*

$$(3.2) \quad \text{dist}((z, \lambda), \mathcal{V}) \leq \delta_1(\mathcal{V}_\lambda)$$

we have

$$\eta(z, \lambda) \in [1/\beta(\mathcal{V}_\lambda), \beta(\mathcal{V}_\lambda)]\text{dist}((z, \lambda), \mathcal{S}),$$

where $\eta(z, \lambda)$ is defined in (3.1).

Proof. From Theorem 3.1, we have for any particular $\lambda^* \in \mathcal{V}_\lambda$ that there are positive constants $\delta_1(\lambda^*)$ and $\beta(\lambda^*)$ such that for any (z, λ) with $\|(z, \lambda) - (z^*, \lambda^*)\|_2 \leq \delta_1(\lambda^*)$ we have

$$\eta(z, \lambda) \in [1/\beta(\lambda^*), \beta(\lambda^*)]\text{dist}((z, \lambda), \mathcal{S}).$$

The open balls

$$\mathcal{B}(\lambda^*) \stackrel{\text{def}}{=} \{(z, \lambda) \mid \|(z, \lambda) - (z^*, \lambda^*)\|_2 < \delta_1(\lambda^*)\} \quad \text{for all } \lambda^* \in \mathcal{V}_\lambda$$

form an open cover of $\mathcal{V}$. Since $\mathcal{V}$ is compact, there is a finite subcover, indexed by a finite set of multipliers, say $\{\lambda_1^*, \lambda_2^*, \dots, \lambda_M^*\}$. We choose $\delta_1(\mathcal{V}_\lambda) > 0$ such that

$$\{(z, \lambda) \mid \text{dist}((z, \lambda), \mathcal{V}) \leq \delta_1(\mathcal{V}_\lambda)\} \subset \cup_{i=1}^M \mathcal{B}(\lambda_i^*),$$

and set $\beta(\mathcal{V}_\lambda) = \max_{i=1,2,\dots,M} \beta(\lambda_i^*)$ to obtain the result. $\square$

An active set identification strategy for (1.1) follows immediately from the estimate of Theorem 3.2. Similarly to Facchinei, Fischer, and Kanzow [4, equation (2.5)], we choose a constant $\tau \in (0, 1)$ and define

$$(3.3) \quad \mathcal{A}(z, \lambda) \stackrel{\text{def}}{=} \{i = 1, 2, \dots, m \mid g_i(z) \geq -\eta(z, \lambda)^\tau\}.$$

Given a set $\mathcal{V}$ defined as in Theorem 3.2, we have, for all (z, λ) with $\text{dist}((z, \lambda), \mathcal{V})$ sufficiently small, that

$$\begin{aligned} i \in \mathcal{A}^* &\Rightarrow -g_i(z) \leq O(\|z - z^*\|) \leq \|z - z^*\|^\tau / \beta(\mathcal{V}_\lambda)^\tau \leq \eta(z, \lambda)^\tau, \\ i \notin \mathcal{A}^* &\Rightarrow g_i(z) \leq (1/2)g_i(z^*) < -\eta(z, \lambda)^\tau, \end{aligned}$$

so that $\mathcal{A}(z, \lambda) = \mathcal{A}^*$. We formalize this result as follows.

THEOREM 3.3. *Suppose that Assumption 1 holds at z^* and that the second-order condition (2.4) holds for all $\lambda^* \in \mathcal{V}_\lambda$, where $\mathcal{V}_\lambda$ is a compact subset of $\mathcal{S}_\lambda$. Let us define $\mathcal{V} = \{z^*\} \times \mathcal{V}_\lambda$. Then there exists $\delta_2(\mathcal{V}_\lambda) > 0$ such that for all (z, λ) with $\text{dist}((z, \lambda), \mathcal{V}) \leq \delta_2(\mathcal{V}_\lambda)$, we have $\mathcal{A}(z, \lambda) = \mathcal{A}^*$.*

We now use the results of Hager and Gowda [9] to obtain distance-to-solution estimates for the equality constrained problem (2.6), for multipliers γ in some neighborhood of a compact subset of $\bar{\mathcal{S}}_\gamma$, which we denote by $\bar{\mathcal{V}}_\gamma$. We also introduce the notation $\bar{\mathcal{V}} \stackrel{\text{def}}{=} \{z^*\} \times \bar{\mathcal{V}}_\gamma$ to denote the primal-dual extension of $\bar{\mathcal{V}}_\gamma$.

THEOREM 3.4. *Consider the problem (2.6), and suppose that (2.8) and (2.9) hold at (z^*, γ^*) for some fixed $v > 0$ and all $\gamma^* \in \bar{\mathcal{V}}_\gamma \subset \bar{\mathcal{S}}_\gamma$, where $\bar{\mathcal{V}}_\gamma$ is compact. Then there are positive constants $\bar{\delta}_1(\bar{\mathcal{V}}_\gamma)$ and $\bar{\beta}(\bar{\mathcal{V}}_\gamma)$ such that for all (z, γ) with $\text{dist}((z, \gamma), \bar{\mathcal{V}}) \leq \bar{\delta}_1(\bar{\mathcal{V}}_\gamma)$, we have*

$$(3.4) \quad \bar{\eta}(z, \gamma) \stackrel{\text{def}}{=} \left\| \begin{bmatrix} \nabla_z \bar{\mathcal{L}}(z, \gamma) \\ h(z) \end{bmatrix} \right\|_1 \in [1/\bar{\beta}(\bar{\mathcal{V}}_\gamma), \bar{\beta}(\bar{\mathcal{V}}_\gamma)] \text{dist}((z, \gamma), \bar{\mathcal{S}}).$$

Proof. From Hager and Gowda [9, Lemma 2, Theorem 3], we have for any particular $\gamma^* \in \bar{\mathcal{V}}_\gamma$ that there are positive constants $\bar{\delta}_1(\gamma^*)$ and $\bar{\beta}(\gamma^*)$ such that for any (z, γ) with $\|(z, \gamma) - (z^*, \gamma^*)\|_2 \leq \bar{\delta}_1(\gamma^*)$ we have

$$\bar{\eta}(z, \gamma) \in [1/\bar{\beta}(\gamma^*), \bar{\beta}(\gamma^*)] \text{dist}((z, \gamma), \bar{\mathcal{S}}).$$

The result now follows from the same compactness argument as in the proof of Theorem 3.2. $\square$

4. An algorithm for the equality constrained problem. In this section, we describe a locally convergent algorithm for the equality constrained problem (2.6), and then discuss its relationship to other methods which require the solution of a similar linear system at each iteration.

4.1. Algorithm SLN and its convergence properties. We obtain the step from a point (z, γ) by the following process. First, define μ as follows:

$$(4.1) \quad \mu \stackrel{\text{def}}{=} \left\| \begin{bmatrix} \nabla_z \bar{\mathcal{L}}(z, \gamma) \\ h(z) \end{bmatrix} \right\|_1 = \bar{\eta}(z, \gamma).$$

Next, solve the following linear system:

$$(4.2) \quad \begin{bmatrix} \nabla_{zz}^2 \bar{\mathcal{L}}(z, \gamma) & \nabla h(z) \\ -\nabla h(z)^T & \mu I \end{bmatrix} \begin{bmatrix} \Delta z \\ \Delta \gamma \end{bmatrix} = \begin{bmatrix} -\nabla_z \bar{\mathcal{L}}(z, \gamma) \\ h(z) \end{bmatrix}.$$

We call the approach “stabilized Lagrange-Newton,” since it is based on solving the Lagrange-Newton equations that constitute the first-order optimality conditions.

ALGORITHM SLN.

Given starting point (z^0, γ^0)

for $k = 0, 1, 2, \dots$

 Define μ_k by setting $(z, \gamma) = (z^k, \gamma^k)$ in (4.1);

 Solve (4.2) with $(z, \gamma) = (z^k, \gamma^k)$ and $\mu = \mu_k$ to obtain $(\Delta z^k, \Delta \gamma^k)$;

 Set $(z^{k+1}, \gamma^{k+1}) \leftarrow (z^k, \gamma^k) + (\Delta z^k, \Delta \gamma^k)$;

end (for)

LEMMA 4.1. *Suppose that Assumption 2 holds and that (2.9) holds for constant $v > 0$ and all $\gamma^* \in \bar{\mathcal{V}}_\gamma$, where $\bar{\mathcal{V}}_\gamma$ is a compact subset of $\bar{\mathcal{S}}_\gamma$. Then there exist positive constants $\delta_3(\bar{\mathcal{V}}_\gamma)$, $\chi_1(\bar{\mathcal{V}}_\gamma)$, and $\chi_2(\bar{\mathcal{V}}_\gamma)$ with $\delta_3(\bar{\mathcal{V}}_\gamma) \leq \min(1, \bar{\delta}_1(\bar{\mathcal{V}}_\gamma))$ such that for all (z, γ) with*

$$(4.3) \quad \text{dist}((z, \gamma), \bar{\mathcal{V}}) \leq \delta_3(\bar{\mathcal{V}}_\gamma),$$

where $\bar{\mathcal{V}} = \{z^*\} \times \bar{\mathcal{V}}_\gamma$, the step $(\Delta z, \Delta \gamma)$ obtained from (4.1), (4.2) satisfies the following relations:

$$(4.4) \quad \|(\Delta z, \Delta \gamma)\|_1 \leq \chi_1(\bar{\mathcal{V}}_\gamma) \bar{\eta}(z, \gamma),$$

$$(4.5) \quad \bar{\eta}(z + \Delta z, \gamma + \Delta \gamma) \leq \chi_2(\bar{\mathcal{V}}_\gamma) \bar{\eta}(z, \gamma)^2.$$

Proof. We omit a detailed proof of this result, since it follows closely the proofs of Theorems 3.2 and 4.1 in Wright [15]. It is based on a singular value decomposition of the constraint Jacobian $\nabla h(z^*)$, which induces a partitioning of both Δz and $\Delta \gamma$ into two orthogonal components. The uniform positive definiteness of $\nabla_{zz}^2 \bar{\mathcal{L}}(z, \gamma)$ on the null space of $\nabla h(z^*)^T$ is crucial to the proof, as is the distance-to-solution result, Theorem 3.4. $\square$

THEOREM 4.2. *Suppose that Assumption 2 holds. Let $\hat{\mathcal{V}}_\gamma$ be a compact subset of $\bar{\mathcal{S}}_\gamma$, with $\hat{\mathcal{V}} = \{z^*\} \times \hat{\mathcal{V}}_\gamma$, and define*

$$(4.6) \quad \bar{\mathcal{V}}_\gamma = \{\gamma \in \bar{\mathcal{S}}_\gamma \mid \text{dist}(\gamma, \hat{\mathcal{V}}_\gamma) \leq \epsilon\}$$

for some $\epsilon > 0$. Suppose that (2.9) holds for all $\gamma^* \in \bar{\mathcal{V}}_\gamma$ and fixed $v > 0$. Then there exists a positive constant $\delta_5(\hat{\mathcal{V}}_\gamma, \epsilon)$ such that if (z^0, γ^0) satisfies

$$(4.7) \quad \text{dist}((z^0, \gamma^0), \hat{\mathcal{V}}) \leq \delta_5(\hat{\mathcal{V}}_\gamma, \epsilon),$$

then the sequence $\{(z^k, \gamma^k)\}$ generated in Algorithm SLN converges Q -quadratically to a point $(z^*, \gamma^*) \in \bar{\mathcal{S}}$ and satisfies the relations (4.4) and (4.5) at each iteration.

Proof. This proof makes extensive use of the quantities $\bar{\beta}(\bar{\mathcal{V}}_\gamma)$, $\delta_3(\bar{\mathcal{V}}_\gamma)$, $\chi_1(\bar{\mathcal{V}}_\gamma)$, and $\chi_2(\bar{\mathcal{V}}_\gamma)$ defined in Theorem 3.4 and Lemma 4.1. For convenience of notation during this proof, we drop the explicit dependence of the quantities $\delta_3(\bar{\mathcal{V}}_\gamma)$, $\chi_1(\bar{\mathcal{V}}_\gamma)$, $\chi_2(\bar{\mathcal{V}}_\gamma)$, $\bar{\beta}(\bar{\mathcal{V}}_\gamma)$, $\delta_5(\hat{\mathcal{V}}_\gamma, \epsilon)$ on the sets $\bar{\mathcal{V}}_\gamma$ and $\hat{\mathcal{V}}_\gamma$ and the scalar ϵ .

We show that the theorem holds for the following choice of δ_5 :

$$(4.8) \quad \delta_5 \stackrel{\text{def}}{=} \min \left(\frac{1}{4\chi_2\bar{\beta}}, \frac{\min(\delta_3, \epsilon)}{4}, \frac{\min(\delta_3, \epsilon)}{4\chi_1\bar{\beta}} \right).$$

The main part of the proof is to show inductively that the following bounds hold for all $k = 0, 1, 2, \dots$:

$$(4.9a) \quad \bar{\eta}(z^{k+1}, \gamma^{k+1}) \leq (1/2) \bar{\eta}(z^k, \gamma^k),$$

$$(4.9b) \quad \text{dist}((z^{k+1}, \gamma^{k+1}), \bar{\mathcal{V}}) \leq (1 - 1/2^{k+2}) \delta_3,$$

$$(4.9c) \quad \|(\Delta z^k, \Delta \gamma^k)\|_1 \leq \min(\delta_3, \epsilon) / 2^{k+2},$$

$$(4.9d) \quad \text{dist}(\gamma^{k+1}, \hat{\mathcal{V}}_\gamma) \leq (1 - 1/2^{k+1}) \epsilon.$$

We first consider $k = 0$. From Theorem 3.4 and (4.8), we have

$$\chi_2 \bar{\eta}(z^0, \gamma^0) \leq \chi_2 \bar{\beta} \text{dist}((z^0, \gamma^0), \bar{\mathcal{S}}) \leq \chi_2 \bar{\beta} \delta_5 \leq 1/4.$$

Hence, we have from (4.5) that

$$\bar{\eta}(z^1, \gamma^1) \leq (\chi_2 \bar{\eta}(z^0, \gamma^0)) \bar{\eta}(z^0, \gamma^0) \leq (1/4) \bar{\eta}(z^0, \gamma^0),$$

so that (4.9a) holds for $k = 0$. From (4.4) and (4.8), we have

$$\|(\Delta z^0, \Delta \gamma^0)\|_1 \leq \chi_1 \bar{\eta}(z^0, \gamma^0) \leq \chi_1 \bar{\beta} \text{dist}((z^0, \gamma^0), \bar{\mathcal{S}}) \leq \chi_1 \bar{\beta} \delta_5 \leq (1/4) \min(\delta_3, \epsilon),$$

so (4.9c) holds for $k = 0$. For (4.9b), we have

$$\begin{aligned} \text{dist}((z^1, \gamma^1), \bar{\mathcal{V}}) &\leq \text{dist}((z^0, \gamma^0), \bar{\mathcal{V}}) + \|(\Delta z^0, \Delta \gamma^0)\|_1 \\ &\leq \text{dist}((z^0, \gamma^0), \hat{\mathcal{V}}) + \|(\Delta z^0, \Delta \gamma^0)\|_1 \leq (1/4)\delta_3 + (1/4)\delta_3 < (1 - 1/4)\delta_3. \end{aligned}$$

For (4.9d), we have

$$\text{dist}(\gamma^1, \hat{\mathcal{V}}_\gamma) \leq \text{dist}(\gamma^0, \hat{\mathcal{V}}_\gamma) + \|\Delta \gamma^0\|_1 \leq (1/4 + 1/4) \min(\delta_3, \epsilon) \leq (1/2)\epsilon,$$

proving that this bound also holds at $k = 0$.

We assume now that (4.9) hold for $0, 1, 2, \dots, k$, and prove that these estimates continue to hold when k is replaced by $k + 1$. For (4.9c), we have

$$\begin{aligned} &\|(\Delta z^{k+1}, \Delta \gamma^{k+1})\|_1 \\ &\leq \chi_1 \bar{\eta}(z^{k+1}, \gamma^{k+1}) && \text{by (4.9b) and Lemma 4.1} \\ &\leq (1/2^{k+1}) \chi_1 \bar{\eta}(z^0, \gamma^0) && \text{by (4.9a)} \\ &\leq (1/2^{k+1}) \chi_1 \bar{\beta} \text{dist}((z^0, \gamma^0), \bar{\mathcal{S}}) && \text{by Theorem 3.4} \\ &\leq (1/2^{k+1}) \chi_1 \bar{\beta} \delta_5 && \text{by (4.7) and } \hat{\mathcal{V}} \subset \bar{\mathcal{S}} \\ &\leq (1/2^{k+1}) (1/4) \min(\delta_3, \epsilon) && \text{by (4.8)} \\ &= (1/2^{k+3}) \min(\delta_3, \epsilon), \end{aligned}$$

verifying that this bound continues to hold for $k + 1$. For (4.9b), we have from (4.9b) at k and (4.9c) at $k + 1$ that

$$\begin{aligned} \text{dist}((z^{k+2}, \gamma^{k+2}), \bar{\mathcal{V}}) &\leq \text{dist}((z^{k+1}, \gamma^{k+1}), \bar{\mathcal{V}}) + \|(\Delta z^{k+1}, \Delta \gamma^{k+1})\|_1 \\ &\leq (1 - 1/2^{k+2})\delta_3 + (1/2^{k+3})\delta_3 \\ &= (1 - 1/2^{k+3})\delta_3, \end{aligned}$$

proving this result for $k + 1$. For (4.9a), we have

$$\bar{\eta}(z^{k+1}, \gamma^{k+1}) \leq (1/2^{k+1}) \bar{\eta}(z^0, \gamma^0),$$

so that

$$\begin{aligned} &\chi_2 \bar{\eta}(z^{k+1}, \gamma^{k+1}) \\ &\leq (1/2^{k+1}) \chi_2 \bar{\eta}(z^0, \gamma^0) && \text{by (4.9a)} \\ &\leq (1/2^{k+1}) \chi_2 \bar{\beta} \text{dist}((z^0, \gamma^0), \bar{\mathcal{S}}) && \text{by Theorem 3.4} \\ &\leq (1/2^{k+1}) \chi_2 \bar{\beta} \text{dist}((z^0, \gamma^0), \hat{\mathcal{V}}) && \text{since } \hat{\mathcal{V}} \subset \bar{\mathcal{S}} \\ &\leq (1/2^{k+1}) (1/4) = 1/2^{k+3} && \text{by (4.7) and (4.8)}. \end{aligned}$$

Because we have shown that the bound (4.9b) holds when k is replaced by $k + 1$, we have from Lemma 4.1 that

$$\begin{aligned}\bar{\eta}(z^{k+2}, \gamma^{k+2}) &\leq (\chi_2 \bar{\eta}(z^{k+1}, \gamma^{k+1})) \bar{\eta}(z^{k+1}, \gamma^{k+1}) \\ &\leq (1/2^{k+3}) \bar{\eta}(z^{k+1}, \gamma^{k+1}),\end{aligned}$$

proving that (4.9a) continues to hold at $k + 1$. To prove (4.9d) at $k + 1$, we have from (4.9c) at $k + 1$ and $\delta_3 \leq 1$ that

$$\begin{aligned}\text{dist}(\gamma^{k+2}, \hat{\mathcal{V}}_\gamma) &\leq \text{dist}(\gamma^{k+1}, \hat{\mathcal{V}}_\gamma) + \|\Delta\gamma^{k+1}\|_1 \\ &\leq (1 - 1/2^{k+1}) \min(\delta_3, \epsilon) + (1/2^{k+3}) \min(\delta_3, \epsilon) \leq (1 - 1/2^{k+2}) \min(\delta_3, \epsilon),\end{aligned}$$

as required.

It follows from (4.9c) that the sequence $\{(z^k, \gamma^k)\}$ is Cauchy and hence convergent. Because of (4.9a), the limit point $((z^*, \gamma^*)$, say) must lie in $\bar{\mathcal{S}}$. Finally, Q -quadratic convergence follows from the following chain of inequalities:

$$\begin{aligned}&\|(z^{k+1}, \gamma^{k+1}) - (z^*, \gamma^*)\|_2 \\ &\leq \sum_{i=k+1}^{\infty} \|(\Delta z^i, \Delta \gamma^i)\|_1 \\ &\leq \sum_{i=k+1}^{\infty} \chi_1 \bar{\eta}(z^i, \gamma^i) && \text{by (4.4)} \\ &\leq 2\chi_1 \bar{\eta}(z^{k+1}, \gamma^{k+1}) && \text{by (4.9a)} \\ &\leq 2\chi_1 \chi_2 \bar{\eta}(z^k, \gamma^k)^2 && \text{by (4.5)} \\ &\leq 2\chi_1 \chi_2 \bar{\beta}^2 \text{dist}((z^k, \gamma^k), \bar{\mathcal{S}})^2 && \text{by Theorem 3.4} \\ &\leq 2\chi_1 \chi_2 \bar{\beta}^2 \|(z^k, \gamma^k) - (z^*, \gamma^*)\|^2 && \text{since } (z^*, \gamma^*) \in \bar{\mathcal{S}}. \quad \square\end{aligned}$$

4.2. Discussion. The step (4.2) used by Algorithm SLN is similar to steps that arise in at least two well-known algorithms for the equality constrained problem (2.6), neither of which was designed with the issue of degeneracy in mind. The first example is the quadratic penalty function, which can be defined as

$$P(z; \mu) = \phi(z) + \frac{1}{2\mu} \|h(z)\|_2^2.$$

Algorithms based on this function typically find an approximate minimizer of $P(\cdot; \mu)$, then decrease μ and repeat the process. Under certain assumptions, it is known that the sequence of minimizers approaches a solution z^* as $\mu \downarrow 0$; see [12, Chapter 17].

It is not difficult to show that the Newton step Δz for $P(\cdot; \mu)$ satisfies the following system (in conjunction with another vector $\Delta \gamma$):

$$(4.10) \quad \begin{bmatrix} \nabla_{zz}^2 \bar{\mathcal{L}}(z, h(z)/\mu) & \nabla h(z) \\ -\nabla h(z)^T & \mu I \end{bmatrix} \begin{bmatrix} \Delta z \\ \Delta \gamma \end{bmatrix} = \begin{bmatrix} -\nabla_z \bar{\mathcal{L}}(z, 0) \\ h(z) \end{bmatrix}.$$

The similarities between the systems (4.2) and (4.10) are evident. Note, however, that the upper left block of the coefficient matrix in (4.10) is guaranteed to be similar to that of (4.2) only when $h(z)/\mu \approx \gamma$ —a condition that cannot be expected to hold at many of the iterates visited by the quadratic penalty approach. Hence, the

system does not have the nice property of (4.2) in which Assumption 2 guarantees nonsingularity of the coefficient matrix on later iterations.

The augmented Lagrangian for (2.6) is conventionally defined as follows:

$$\hat{\mathcal{L}}(z, \gamma; \mu) \stackrel{\text{def}}{=} \phi(z) + \gamma^T h(z) + \frac{1}{2\mu} \|h(z)\|_2^2.$$

The basic method of multipliers [10, 13, 2] proceeds by finding an approximate minimizer of $\hat{\mathcal{L}}(\cdot, \gamma; \mu)$ for fixed γ and μ , then updating γ by the formula $\gamma \leftarrow \gamma + h(z)/\mu$, then possibly decreasing μ and repeating the process. When Newton's method is used to minimize $\hat{\mathcal{L}}(\cdot, \gamma; \mu)$, the step Δz satisfies the following system:

$$(4.11) \quad \begin{bmatrix} \nabla_{zz}^2 \bar{\mathcal{L}}(z, \gamma + h(z)/\mu) & \nabla h(z) \\ -\nabla h(z)^T & \mu I \end{bmatrix} \begin{bmatrix} \Delta z \\ \Delta \gamma \end{bmatrix} = \begin{bmatrix} -\nabla_z \bar{\mathcal{L}}(z, \gamma) \\ h(z) \end{bmatrix}.$$

The system (4.11) is indeed very similar to (4.2) when $\|h(z)\| \ll \mu$. The method of multipliers differs from Algorithm SLN, however, in that the $\Delta \gamma$ component of the step is not used (since a different formula is used to update γ) and μ is chosen heuristically rather than via the formula (4.1). Indeed, a main motivation of the method of multipliers is to obtain convergence to z^* without driving μ to zero.

We note that steps for the two methods above have traditionally been obtained using unconstrained minimization algorithms in z alone, rather than the systems (4.10) and (4.11) above. Hence, small values of μ lead to ill conditioned subproblems—a problem that still occurs in the formulations (4.10) and (4.11) above in the degenerate case. Nevertheless, our results in this section, and our observations about numerical implementations in section 6, suggest that local convergence of both approaches could be improved by crossing over to Algorithm SLN on later iterations.

5. Extension to equality constrained problems. We now describe an algorithmic framework for the inequality constrained problem (1.1) that is based on the active constraint identification technique of section 3 and Algorithm SLN of section 4, along with an “outer strategy” that could be derived from any globally convergent algorithm. This framework, which we call Framework INEQ, enters an “EQ phase” when there is reason to believe that it has identified the active constraints correctly. In the EQ phase, it applies Algorithm SLN to the problem obtained by fixing these constraints as equalities and ignoring the others. Checks are applied at each iteration of the EQ phase to ensure that the active constraint identification was in fact correct, and that the algorithm is not converging to a point that does not solve (or, at least, is not a KKT point for) the original problem (1.1). When false convergence of this type is detected, or when the convergence is not sufficiently rapid, Framework INEQ exits the EQ phase and restores the values of (z, λ) that were in force on entry to this phase. (Restoration is necessary because components of λ are allowed to become negative during the EQ phase, making them potentially unsuitable for use as multiplier estimates for the original problem.) Outside of the EQ phase, the iterates (z^k, λ^k) are generated by the outer strategy.

We denote by $\text{EQ}(\mathcal{A})$ the equality-constrained problem obtained by enforcing a subset $\mathcal{A}$ of the components of g in (1.1) as equalities, that is,

$$(5.1) \quad \text{EQ}(\mathcal{A}): \quad \min_z \phi(z) \quad \text{subject to } g_i(z) = 0 \text{ for all } i \in \mathcal{A}.$$

The following result, also noted by Izmailov and Solodov [11, Proposition 3.2], shows that the solution of the inequality constrained problem satisfying the second-order

sufficient condition is a solution of the problem $\text{EQ}(\mathcal{A}^*)$ for the optimal active set $\mathcal{A}^*$ and also satisfies the second-order sufficient condition for this reduced problem.

THEOREM 5.1. *Suppose that Assumption 1 holds at z^* , and that for some multiplier $\lambda^* \in \mathcal{S}_\lambda$, the conditions (2.4) are satisfied. Then the equality constrained problem (5.1) with $\mathcal{A} = \mathcal{A}^*$ has local solution z^* satisfying Assumption 2, and the KKT conditions (2.8) and the second-order condition (2.9) are satisfied by the multiplier $\lambda_{\mathcal{A}^*}^*$.*

The next result is a kind of converse of Theorem 5.1. We omit its proof, which is also simple.

THEOREM 5.2. *Suppose for some $\mathcal{A} \subset \{1, 2, \dots, m\}$ that z^* is a point satisfying Assumption 2 for the problem $\text{EQ}(\mathcal{A})$. Suppose too that there is a multiplier vector $\lambda_{\mathcal{A}}^*$ satisfying (2.8) such that $\lambda_{\mathcal{A}}^* \geq 0$, and in addition that $g_i(z^*) \leq 0$ for all $i \notin \mathcal{A}$. Then z^* satisfies the KKT conditions (1.3) for the problem (1.1).*

An important role in detecting incorrect identification of the active constraints is played by the following test:

$$\begin{aligned} \text{kkt-test}(\bar{z}, \bar{\lambda}_{\mathcal{A}}, \mathcal{A}, B) : & \text{ There exists } \lambda_{\mathcal{A}} \text{ with } \lambda_{\mathcal{A}} \geq 0 \text{ and } \|\lambda_{\mathcal{A}}\|_1 \leq B \text{ such that} \\ (5.2) \quad & \left\| \nabla \phi(\bar{z}) + \sum_{i \in \mathcal{A}} \lambda_i \nabla g_i(\bar{z}) \right\|_1 \leq \bar{\eta}(\bar{z}, \bar{\lambda}_{\mathcal{A}})^\sigma, \end{aligned}$$

where $\sigma \in (.5, 1)$ is a specified constant. Note that the test can be applied by solving the linear program

$$\min_{\lambda_{\mathcal{A}} \geq 0, \|\lambda_{\mathcal{A}}\|_1 \leq B} \left\| \nabla \phi(\bar{z}) + \sum_{i \in \mathcal{A}} \lambda_i \nabla g_i(\bar{z}) \right\|_1,$$

though it may not be necessary to iterate all the way to optimality.

We are now ready to specify the framework.

Framework INEQ.

- O:** Given (z^0, λ^0) with $\lambda^0 \geq 0$, $\tau_{\text{EQ}} \in (0, .5]$, and $\sigma \in (.5, 1)$; Set $k \leftarrow 0$;
- 1:** (* Estimate active constraints *)
 Evaluate $\eta(z^k, \lambda^k)$; Set $\mathcal{A} \leftarrow \mathcal{A}(z^k, \lambda^k)$;
 if $\eta(z^k, \lambda^k) = 0$
 STOP;
- 2:** (* Test for entry into EQ Phase *)
 if $\bar{\eta}(z^k, \lambda_{\mathcal{A}}^k) \leq \tau_{\text{EQ}}$
 $z^{\text{store}} \leftarrow z^k$; $\lambda^{\text{store}} \leftarrow \lambda^k$;
 go to **3**;
 else
 go to **6**;
- 3:** (* EQ Phase *)
 Given $(z^k, \lambda_{\mathcal{A}}^k)$, compute $(\Delta z^k, \Delta \lambda_{\mathcal{A}}^k)$ for problem $\text{EQ}(\mathcal{A})$ from (4.1), (4.2);
- 4:** (* Test validity of EQ Phase step *)
 if $\|(\Delta z^k, \Delta \lambda_{\mathcal{A}}^k)\|_1 > \bar{\eta}(z^k, \lambda_{\mathcal{A}}^k)^\sigma$
 go to **5**;

```

if  $\bar{\eta}(z^k + \Delta z^k, \lambda_{\mathcal{A}}^k + \Delta \lambda_{\mathcal{A}}^k) > \bar{\eta}(z^k, \lambda_{\mathcal{A}}^k)^{1+\sigma}$ 
    go to 5;
if  $g_i(z^k + \Delta z^k) > 0$  for any  $i \notin \mathcal{A}$ 
    go to 5;
if  $\lambda_i^k + \Delta \lambda_i^k < 0$  for some  $i \in \mathcal{A}$  and
     $\text{kkt-test}(z^k + \Delta z^k, \lambda_{\mathcal{A}}^k + \Delta \lambda_{\mathcal{A}}^k, \mathcal{A}, \|\lambda_{\mathcal{A}}^k\|_1 + 2)$  is false
    go to 5;
 $(z^{k+1}, \lambda_{\mathcal{A}}^{k+1}) \leftarrow (z^k + \Delta z^k, \lambda_{\mathcal{A}}^k + \Delta \lambda_{\mathcal{A}}^k)$ ;
 $k \leftarrow k + 1$ ;
if  $\bar{\eta}(z^k, \lambda_{\mathcal{A}}^k) = 0$ 
    STOP;
go to 3;

5:  (* Drop out of EQ phase *)
     $\tau_{\text{EQ}} \leftarrow \tau_{\text{EQ}}/2$ ;  $z^k \leftarrow z^{\text{store}}$ ;  $\lambda^k \leftarrow \lambda^{\text{store}}$ ;

6:  (* Apply outer strategy *)
    Compute  $(z^{k+1}, \lambda^{k+1})$  using the “outer” strategy;
     $k \leftarrow k + 1$ ; go to 1;

```

The main computational costs in each iteration of the EQ phase are (i) the solution of the single linear system (4.2), which has similar structure and sparsity to systems that appear in standard methods for nonlinear programming, and (ii) the solution of the feasibility problem in (5.2). In fact, the latter calculation may not be needed at every step of the EQ phase, since it is performed only if the new iteration survives the earlier tests in step 4 and if some components of the Lagrange multiplier estimates have become negative. In addition, Framework INEQ can be modified easily to perform this test at only a subset of the EQ-phase iterations, reducing the computational requirements further.

Our first convergence result shows that if the iterates enter the EQ phase and remain there, any limit point must be a KKT point for (1.1).

THEOREM 5.3. *Suppose that Framework INEQ does not terminate finitely and that the iterates (z^k, λ^k) remain in the EQ phase for all $k \geq K$, where K is some index. Then the sequence $\{z^k\}$ converges R -superlinearly to a limit $\hat{z}$ which is a KKT point for (1.1).*

Proof. Because of the second test in step 4, $\bar{\eta}(z^k, \lambda_{\mathcal{A}}^k)$ approaches zero geometrically. Using the first test in step 4, we have $\|(\Delta z^k, \Delta \lambda_{\mathcal{A}}^k)\|_1 \leq \bar{\eta}(z^k, \lambda_{\mathcal{A}}^k)^\sigma$, so the sequence $\{(z^k, \lambda_{\mathcal{A}}^k)\}$ is Cauchy and hence convergent, say to $(\hat{z}, \hat{\lambda}_{\mathcal{A}})$. Since, from the third test in step 4, we have $g_i(z^k) \leq 0$ for all $i \notin \mathcal{A}$, then $g_i(\hat{z}) \leq 0$ for all $i \notin \mathcal{A}$. Since $|g_i(z^k)| \leq \bar{\eta}(z^k, \lambda_{\mathcal{A}}^k) \downarrow 0$ for all $i \in \mathcal{A}$, we have $g_i(\hat{z}) = 0$ for $i \in \mathcal{A}$. Therefore, $\hat{z}$ is feasible. Further, by taking limits, we have that $\nabla \phi(\hat{z}) + \sum_{i \in \mathcal{A}} \hat{\lambda}_i \nabla g_i(\hat{z}) = 0$.

Because of the fourth test in step 4, we have either that $\lambda_{\mathcal{A}}^{k+1} \geq 0$ for infinitely many $k \geq K$ or that $\text{kkt-test}(z^{k+1}, \lambda_{\mathcal{A}}^{k+1}, \mathcal{A}, \|\lambda_{\mathcal{A}}^k\|_1 + 2)$ is true for infinitely many $k \geq K$ (or possibly both). In the former case, we have that $\hat{\lambda}_{\mathcal{A}} \geq 0$, so by extending $\hat{\lambda}_{\mathcal{A}}$ to a full multiplier vector $\hat{\lambda}$ by adding zero components at $i \notin \mathcal{A}$, we conclude that $(\hat{z}, \hat{\lambda})$ satisfies the KKT conditions (1.3). In the latter case, let $\tilde{\lambda}_{\mathcal{A}}^{k+1}$ be the vector of multipliers that makes $\text{kkt-test}(z^{k+1}, \lambda_{\mathcal{A}}^{k+1}, \mathcal{A}, \|\lambda_{\mathcal{A}}^k\|_1 + 2)$ true. Since the sequence of such vectors is bounded, there is a subsequential limit $\tilde{\lambda}_{\mathcal{A}}$ such that $\tilde{\lambda}_{\mathcal{A}} \geq 0$ and $\nabla \phi(\hat{z}) + \sum_{i \in \mathcal{A}} \tilde{\lambda}_i \nabla g_i(\hat{z}) = 0$. By extending $\tilde{\lambda}_{\mathcal{A}}$ to a full multiplier vector $\tilde{\lambda}$ by adding

zero components at $i \notin \mathcal{A}$, we conclude that $(\hat{z}, \hat{\lambda})$ satisfies the KKT conditions (1.3). In either case, we conclude that $\hat{z}$ is a KKT point for (1.1).

To verify the rate of convergence, note that for all k sufficiently large, we have from the first and second tests in step 4 that

$$\begin{aligned}
\|(z^k, \lambda_{\mathcal{A}}^k) - (z^*, \hat{\lambda}_{\mathcal{A}})\|_1 &\leq \sum_{j=k}^{\infty} \|(\Delta z^j, \Delta \lambda_{\mathcal{A}}^j)\|_1 \\
&\leq \sum_{j=k}^{\infty} \bar{\eta}(z^j, \lambda_{\mathcal{A}}^j)^{\sigma} \\
&\leq \sum_{j=k}^{\infty} (\bar{\eta}(z^k, \lambda_{\mathcal{A}}^k)^{\sigma})^{(1+\sigma)^{j-k}} \\
&\leq \sum_{j=k}^{\infty} (\bar{\eta}(z^k, \lambda_{\mathcal{A}}^k)^{\sigma})^{1+(j-k)\sigma} \\
&= \bar{\eta}(z^k, \lambda_{\mathcal{A}}^k)^{\sigma} \sum_{j=k}^{\infty} \bar{\eta}(z^k, \lambda_{\mathcal{A}}^k)^{(j-k)\sigma^2} \\
&\leq \bar{\eta}(z^k, \lambda_{\mathcal{A}}^k)^{\sigma} / (1 - \bar{\eta}(z^k, \lambda_{\mathcal{A}}^k)^{\sigma^2}) \\
&\leq 10\bar{\eta}(z^k, \lambda_{\mathcal{A}}^k)^{\sigma},
\end{aligned}$$

where the final inequality follows from

$$\bar{\eta}(z^k, \lambda_{\mathcal{A}}^k)^{\sigma^2} \leq \tau_{\text{EQ}}^{\sigma^2} \leq (1/2)^{1/4} < .9.$$

Since $\{\bar{\eta}(z^k, \lambda_{\mathcal{A}}^k)^{\sigma}\}$ converges Q -superlinearly to zero with Q -order $(1 + \sigma)$, we have that the sequence $\{z^k\}$ converges R -superlinearly to $\hat{z}$ with R -order $1 + \sigma$, as claimed. $\square$

Our next result shows that if the outer strategy is steering the iterates toward the solution set $\mathcal{S}$ for (1.1), and under appropriate assumptions on the boundedness of the multiplier estimates and the second-order condition, the algorithm will eventually enter and remain inside the EQ phase, and will converge superlinearly to a KKT point for (1.1). Note that we cannot be certain (without additional assumptions on the problem) that the stationary limit point lies in $\mathcal{S}$; the limit of $\{z^k\}$ may be different from z^* . Hence, this result combines elements of both local and global convergence analysis.

THEOREM 5.4. *Suppose that Assumption 1 holds for (1.1) at z^* , that all iterates (z^k, λ^k) generated by the outer strategy have $\|\lambda^k\|_1 \leq c$ for some $c > 0$, and that the second-order condition (2.4) is satisfied for all $\lambda^* \in \mathcal{S}_{\lambda}$ with $\|\lambda^*\|_1 \leq c$. Suppose too that the subsequence of iterates (z^k, λ^k) generated by the outer strategy is either finite or else satisfies $\text{dist}((z^k, \lambda^k), \mathcal{S}) \rightarrow 0$. Then Framework INEQ either terminates finitely, with the final point z^k satisfying (1.3) for some λ^* , or else it eventually enters the EQ phase and remains there, and the sequence $\{z^k\}$ converges R -superlinearly to a KKT point $\bar{z}$ satisfying (1.3) with R -order $1 + \sigma$.*

Proof. It is not difficult to show that if Framework INEQ terminates finitely (in either step 1 or step 4), it does so at a point z^k that satisfies the KKT conditions (1.3). (If the termination occurs in step 4, however, the final multiplier $\lambda_{\mathcal{A}}^k$ may not be appropriate for (1.3) as some of its components may be negative. However,

the kkt-test (5.2) ensures the existence of an appropriate multiplier vector.) For the remainder of the proof, we assume that finite termination does not occur.

For the main part of the proof, we assume for contradiction that Framework INEQ never stays permanently in the EQ phase.

We define the set $\mathcal{V}_\lambda$ in Theorem 3.2 as

$$\mathcal{V}_\lambda = \{\lambda^* \in \mathcal{S}_\lambda \mid \|\lambda^*\|_1 \leq c + 1\},$$

and $\hat{\mathcal{V}}_\gamma$ in Theorem 4.2, for the problem $\text{EQ}(\mathcal{A}^*)$, as

$$(5.3) \quad \hat{\mathcal{V}}_\gamma = \{\lambda_{\mathcal{A}^*}^* \in \bar{\mathcal{S}}_\gamma \mid \|\lambda_{\mathcal{A}^*}^*\|_1 \leq c + 1\}.$$

Accordingly, we define $\mathcal{V} = \{z^*\} \times \mathcal{V}_\lambda$ and $\hat{\mathcal{V}} = \{z^*\} \times \hat{\mathcal{V}}_\gamma$. Since $\text{dist}((z^k, \lambda^k), \mathcal{S}) \rightarrow 0$ on the outer iterations, and since $\|\lambda^k\|_1 \leq c$, the projection of (z^k, λ^k) onto $\mathcal{S}$ will eventually lie in $\mathcal{V}$, so by Theorem 3.2, there is $\beta > 0$ such that

$$\eta(z^k, \lambda^k) \in [1/\beta, \beta] \text{dist}((z^k, \lambda^k), \mathcal{S})$$

for all outer strategy iterations k sufficiently large. (For simplicity, as in the proof of Theorem 4.2, we drop the explicit dependence of constants such as β , $\bar{\beta}$, $\bar{\delta}_1$, δ_5 , etc. on the sets $\mathcal{V}_\lambda$, $\hat{\mathcal{V}}_\gamma$, $\bar{\mathcal{V}}_\gamma$, and the scalar ϵ .) It follows from Theorem 3.3 that step 1 sets $\mathcal{A}(z^k, \lambda^k) = \mathcal{A}^*$ at all outer iterations k sufficiently large. Since step 1 is executed infinitely often, we can therefore assume, as we do in the remainder of the proof, that $\mathcal{A} = \mathcal{A}^*$ for *all* k sufficiently large (both for outer-strategy iterations and iterations in the EQ phase).

Consider now the EQ problem $\text{EQ}(\mathcal{A}^*)$. Because of Assumption 1, there is $v > 0$ such that the second-order condition (2.9) is satisfied for problem $\text{EQ}(\mathcal{A}^*)$ for all $\lambda_{\mathcal{A}^*}^* \in \hat{\mathcal{V}}_\gamma$. By choosing ϵ in Theorem 4.2 sufficiently small, we have that (2.9) is in fact satisfied for all $\lambda_{\mathcal{A}^*}^* \in \bar{\mathcal{V}}_\gamma$, where $\bar{\mathcal{V}}_\gamma$ is defined in (4.6), possibly with a reduced (but positive) value of v . Defining $\bar{\mathcal{V}} = \{z^*\} \times \bar{\mathcal{V}}_\gamma$ as before, we have from $\|\lambda_{\mathcal{A}^*}^k\|_1 \leq \|\lambda^k\|_1 \leq c$ that for all outer-strategy iterations k sufficiently large, the following relations hold:

$$\begin{aligned} & \text{dist}((z^k, \lambda_{\mathcal{A}^*}^k), \bar{\mathcal{V}}) \\ & \leq \text{dist}((z^k, \lambda_{\mathcal{A}^*}^k), \hat{\mathcal{V}}) && \text{since } \hat{\mathcal{V}} \subset \bar{\mathcal{V}} \\ & \leq \left[\text{dist}((z^k, \lambda_{\mathcal{A}^*}^k), \hat{\mathcal{V}})^2 + \|\lambda_{(\mathcal{A}^*)^c}^k\|_2^2 \right]^{1/2} \\ & \leq \text{dist}((z^k, \lambda^k), \mathcal{V}) && \text{since } \lambda^* \in \mathcal{V}_\lambda \Rightarrow \lambda_{\mathcal{A}^*}^* \in \hat{\mathcal{V}}_\gamma \\ (5.4) \quad & = \text{dist}((z^k, \lambda^k), \mathcal{S}) \rightarrow 0, \end{aligned}$$

where $\lambda_{(\mathcal{A}^*)^c}^k$ is the vector whose components are λ_i^k , $i \notin \mathcal{A}^*$, and the last equality follows from the fact that the projection (z^*, λ^{k*}) of (z^k, λ^k) onto the set $\mathcal{S}$ must eventually have $\|\lambda^{k*}\|_1 \leq c + 1$, and must therefore also lie in the set $\mathcal{V}$. Hence, the assumptions of Theorem 3.4, Lemma 4.1, and Theorem 4.2 are satisfied by $\text{EQ}(\mathcal{A}^*)$ for all outer-strategy iterations k sufficiently large. In particular, we have from Theorem 3.4 that there is a constant $\bar{\beta} > 0$ such that

$$\bar{\eta}(z^k, \lambda_{\mathcal{A}^*}^k) \in [1/\bar{\beta}, \bar{\beta}] \text{dist}((z^k, \lambda_{\mathcal{A}^*}^k), \bar{\mathcal{S}}),$$

for all outer-strategy iterations k sufficiently large.

Given that Framework INEQ never remains permanently in the EQ phase, we claim now that it enters this phase infinitely often. Otherwise, after the final departure from the EQ phase, τ_{EQ} will be at a fixed nonzero value for the remainder of the iterations. However, in this case, some subsequent iteration k will have

$$\bar{\eta}(z^k, \lambda_{\mathcal{A}^*}^k) \leq \bar{\beta} \text{dist}((z^k, \lambda_{\mathcal{A}^*}^k), \bar{\mathcal{S}}) \leq \bar{\beta} \text{dist}((z^k, \lambda^k), \mathcal{S}) \rightarrow 0,$$

so that the threshold test in step 2 will be satisfied at some subsequent k , activating the EQ phase once again.

Since the algorithm enters the EQ phase infinitely often, it eventually enters with

$$(5.5) \quad \text{dist}((z^k, \lambda_{\mathcal{A}^*}^k), \hat{\mathcal{V}}) \leq \text{dist}((z^k, \lambda_{\mathcal{A}^*}^k), \bar{\mathcal{S}}) \leq \text{dist}((z^k, \lambda^k), \mathcal{S}) \leq \hat{\delta}$$

for any choice of small positive $\hat{\delta}$. We now derive conditions on $\hat{\delta}$ that ensure that every EQ phase iteration $j \geq k$ survives the four tests in step 4, so that all subsequent iterates remain in the EQ phase.

First, we ensure that the conditions of Theorems 3.4 and 4.2 are satisfied by setting

$$(5.6) \quad \hat{\delta} \leq \delta_5.$$

This condition ensures that the estimates (4.4) and (4.5) hold at all subsequent iterations of the EQ phase—and indeed that (4.9) are satisfied as well. In particular, we have from (4.9b), the definition of δ_3 in Lemma 4.1, and Theorem 3.4 that the estimate (3.4) holds for all iterations $j \geq k$ in the EQ phase. Note too from (4.9a) that $\bar{\eta}(z^j, \lambda_{\mathcal{A}^*}^j)$ decreases monotonically at all subsequent iterations $j \geq k$ in the EQ phase.

To handle the first test in step 4, we require that

$$(5.7) \quad \chi_1 \bar{\beta}^{1-\sigma} \hat{\delta}^{1-\sigma} \leq 1.$$

Then for all iterates $j \geq k$ in the EQ phase, we have from (3.4) and (4.9a) that

$$\begin{aligned} & \|(\Delta z^j, \Delta \lambda_{\mathcal{A}^*}^j)\|_1 \\ & \leq \chi_1 \bar{\eta}(z^j, \lambda_{\mathcal{A}^*}^j) && \text{from (4.4) in Lemma 4.1} \\ & = \left[\chi_1 \bar{\eta}(z^j, \lambda_{\mathcal{A}^*}^j)^{1-\sigma} \right] \bar{\eta}(z^j, \lambda_{\mathcal{A}^*}^j)^\sigma \\ & \leq \left[\chi_1 \bar{\eta}(z^k, \lambda_{\mathcal{A}^*}^k)^{1-\sigma} \right] \bar{\eta}(z^j, \lambda_{\mathcal{A}^*}^j)^\sigma && \text{by (4.9a)} \\ & \leq \left[\chi_1 \bar{\beta}^{1-\sigma} \text{dist}((z^k, \lambda_{\mathcal{A}^*}^k), \bar{\mathcal{S}})^{1-\sigma} \right] \bar{\eta}(z^j, \lambda_{\mathcal{A}^*}^j)^\sigma && \text{by Theorem 3.4} \\ & \leq \left[\chi_1 \bar{\beta}^{1-\sigma} \hat{\delta}^{1-\sigma} \right] \bar{\eta}(z^j, \lambda_{\mathcal{A}^*}^j)^\sigma && \text{by (5.5)} \\ & \leq \bar{\eta}(z^j, \lambda_{\mathcal{A}^*}^j)^\sigma && \text{by (5.7).} \end{aligned}$$

Hence, all iterates $j \geq k$ in the EQ phase survive the first test in step 4 when (5.7) holds.

To handle the second conditional statement in step 4, we require that

$$(5.8) \quad \hat{\delta} \leq \chi_2^{-1/(1-\sigma)} / \bar{\beta}.$$

We then have from Theorem 3.4 and (5.5) that

$$\bar{\eta}(z^k, \lambda_{\mathcal{A}^*}^k) \leq \bar{\beta} \text{dist}((z^k, \lambda_{\mathcal{A}^*}^k), \bar{\mathcal{S}}) \leq \bar{\beta} \hat{\delta} \leq \chi_2^{-1/(1-\sigma)}.$$

Hence, at all iterates $j \geq k$ in the EQ phase, we have from the bound above together with Lemma 4.1 and (4.9a) that

$$\begin{aligned}\bar{\eta}(z^{j+1}, \lambda_{\mathcal{A}^*}^{j+1}) &\leq \chi_2 \bar{\eta}(z^j, \lambda_{\mathcal{A}^*}^j)^2 \\ &= \left[\chi_2 \bar{\eta}(z^j, \lambda_{\mathcal{A}^*}^j)^{1-\sigma} \right] \bar{\eta}(z^j, \lambda_{\mathcal{A}^*}^j)^{1+\sigma} \\ &\leq \left[\chi_2 \bar{\eta}(z^k, \lambda_{\mathcal{A}^*}^k)^{1-\sigma} \right] \bar{\eta}(z^j, \lambda_{\mathcal{A}^*}^j)^{1+\sigma} \\ &\leq \bar{\eta}(z^j, \lambda_{\mathcal{A}^*}^j)^{1+\sigma},\end{aligned}$$

ensuring that all iterates $j \geq k$ in the EQ phase survive the second test.

For the third condition in step 4, we first define $\delta_6 > 0$ small enough that $g_i(z) < 0$ for all $i \notin \mathcal{A}^*$ and all z with $\|z - z^*\| \leq \delta_6$. By requiring

$$(5.9) \quad \hat{\delta} \leq \delta_6 / \bar{\beta}^2,$$

we have from Theorem 3.4 and (4.9a) that for all iterations $j \geq k$ in the EQ phase,

$$\begin{aligned}\|z^j - z^*\| &\leq \text{dist}((z^j, \lambda_{\mathcal{A}^*}^j), \bar{\mathcal{S}}) \\ &\leq \bar{\beta} \bar{\eta}(z^j, \lambda_{\mathcal{A}^*}^j) \leq \bar{\beta} \bar{\eta}(z^k, \lambda_{\mathcal{A}^*}^k) \leq \bar{\beta}^2 \text{dist}((z^k, \lambda_{\mathcal{A}^*}^k), \bar{\mathcal{S}}) \leq \bar{\beta}^2 \hat{\delta} \leq \delta_6,\end{aligned}$$

ensuring that each subsequent iteration in the EQ phase also survives the third test in step 4.

We turn now to the fourth test in step 4. We claim that the condition in this test will hold if $\hat{\delta}$ satisfies

$$(5.10) \quad n^{1/2} \bar{\beta}^{3-\sigma} \hat{\delta}^{1-\sigma} < 1.$$

As a preliminary, we show that there exists a vector $\lambda_{\mathcal{A}^*}^* \in \hat{\mathcal{V}}_\gamma$ with $e^T \lambda_{\mathcal{A}^*}^* \leq \|\lambda_{\mathcal{A}^*}^j\|_1 + 2$, for all $j \geq k$ in the EQ phase, and which therefore satisfies the first two conditions in (5.2) for $B = \|\lambda_{\mathcal{A}^*}^j\|_1 + 2$, as defined in step 4. Defining λ^* by projecting (z^k, λ^k) onto $\mathcal{S}$, we have

$$\begin{aligned}\|\lambda_{\mathcal{A}^*}^*\|_1 &\leq \|\lambda_{\mathcal{A}^*}^k\|_1 + n^{1/2} \|\lambda_{\mathcal{A}^*}^k - \lambda_{\mathcal{A}^*}^*\|_2 \\ &\leq \|\lambda_{\mathcal{A}^*}^k\|_1 + n^{1/2} \text{dist}((z^k, \lambda^k), \mathcal{S}) \\ &\leq c + n^{1/2} \hat{\delta} \leq c + 1,\end{aligned}$$

where the final inequality follows from (5.10), since $n^{1/2} \hat{\delta} \leq (n^{1/2} \hat{\delta})(\bar{\beta}^{3-\sigma} \hat{\delta}^{-\sigma}) = n^{1/2} \bar{\beta}^{3-\sigma} \hat{\delta}^{1-\sigma} < 1$. Hence, $\lambda_{\mathcal{A}^*}^* \in \hat{\mathcal{V}}_\gamma \subset \bar{\mathcal{V}}_\gamma$, where $\hat{\mathcal{V}}_\gamma$ is defined in (5.3) and $\bar{\mathcal{V}}_\gamma$ is defined in (4.6). We have from (4.9b) and the definition of δ_3 in Lemma 4.1 that

$$(5.11) \quad \|z^j - z^*\| \leq \text{dist}((z^j, \lambda_{\mathcal{A}^*}^*), \bar{\mathcal{V}}) = \|z^j - z^*\| \leq \delta_3 \leq \bar{\delta}_1,$$

so the estimate (3.4) holds for $(z^j, \lambda_{\mathcal{A}^*}^*)$ for all iterations $j \geq k$ in the EQ phase.

By applying (4.9c), (5.10), and the fact that $\delta_3 \leq 1$ (see Lemma 4.1), we have that

$$\begin{aligned}\|\lambda_{\mathcal{A}^*}^*\|_1 &\leq \|\lambda_{\mathcal{A}^*}^j\|_1 + \|\lambda_{\mathcal{A}^*}^k - \lambda_{\mathcal{A}^*}^*\|_1 + \sum_{i=k}^{j-1} \|\Delta \lambda_{\mathcal{A}^*}^i\|_1 \\ &\leq \|\lambda_{\mathcal{A}^*}^j\|_1 + n^{1/2} \hat{\delta} + \min(\delta_3, \epsilon) \\ &\leq \|\lambda_{\mathcal{A}^*}^j\|_1 + 1 + 1,\end{aligned}$$

showing that the inequality $\|\lambda_{\mathcal{A}^*}^*\|_1 \leq B = \|\lambda_{\mathcal{A}^*}^j\|_1 + 2$ is satisfied for all iterations $j \geq k$ in the EQ phase.

Turning now to kkt-test($z^j + \Delta z^j, \lambda_{\mathcal{A}^*}^j + \Delta \lambda_{\mathcal{A}^*}^j, \mathcal{A}, B$) with $B = \|\lambda_{\mathcal{A}^*}^j\|_1 + 2$, we have for all iterates $j \geq k$ in the EQ phase that

$$\begin{aligned}
& \min_{\lambda_{\mathcal{A}^*} \geq 0, \|\lambda_{\mathcal{A}^*}\|_1 \leq B} \left\| \nabla \phi(z^j) + \sum_{i \in \mathcal{A}^*} \lambda_i \nabla g_i(z^j) \right\|_1 \\
& \leq \min_{\lambda_{\mathcal{A}^*} \geq 0, \|\lambda_{\mathcal{A}^*}\|_1 \leq B} n^{1/2} \left\| \nabla \phi(z^j) + \sum_{i \in \mathcal{A}^*} \lambda_i \nabla g_i(z^j) \right\|_2 \\
& \leq \min_{\lambda_{\mathcal{A}^*} \geq 0, \|\lambda_{\mathcal{A}^*}\|_1 \leq B} n^{1/2} \bar{\eta}(z^j, \lambda_{\mathcal{A}^*}) \\
& \leq n^{1/2} \bar{\eta}(z^j, \lambda_{\mathcal{A}^*}^*) && \text{for } \lambda_{\mathcal{A}^*}^* \text{ defined above} \\
& \leq n^{1/2} \bar{\beta} \text{dist}((z^j, \lambda_{\mathcal{A}^*}^*), \bar{\mathcal{S}}) && \text{by (5.11) and Theorem 3.4} \\
& \leq n^{1/2} \bar{\beta} \|z^j - z^*\| && \text{since } \lambda_{\mathcal{A}^*}^* \in \bar{\mathcal{S}} \\
& \leq n^{1/2} \bar{\beta} \text{dist}((z^j, \lambda_{\mathcal{A}^*}^j), \bar{\mathcal{S}}), \\
& \leq n^{1/2} \bar{\beta}^2 \bar{\eta}(z^j, \lambda_{\mathcal{A}^*}^j) && \text{by Theorem 3.4} \\
& \leq \left[n^{1/2} \bar{\beta}^2 \bar{\eta}(z^j, \lambda_{\mathcal{A}^*}^j)^{1-\sigma} \right] \bar{\eta}(z^j, \lambda_{\mathcal{A}^*}^j)^\sigma \\
& \leq \left[n^{1/2} \bar{\beta}^2 \bar{\eta}(z^k, \lambda_{\mathcal{A}^*}^k)^{1-\sigma} \right] \bar{\eta}(z^j, \lambda_{\mathcal{A}^*}^j)^\sigma && \text{by (4.9a)} \\
& \leq \left[n^{1/2} \bar{\beta}^{3-\sigma} \text{dist}((z^k, \lambda_{\mathcal{A}^*}^k), \bar{\mathcal{S}})^{1-\sigma} \right] \bar{\eta}(z^j, \lambda_{\mathcal{A}^*}^j)^\sigma && \text{by Theorem 3.4} \\
& \leq \left[n^{1/2} \bar{\beta}^{3-\sigma} \hat{\delta}^{1-\sigma} \right] \bar{\eta}(z^j, \lambda_{\mathcal{A}^*}^j)^\sigma && \text{by (5.5)} \\
& < \bar{\eta}(z^j, \lambda_{\mathcal{A}^*}^j)^\sigma && \text{by (5.10).}
\end{aligned}$$

Therefore, we have that

$$\min_{\lambda_{\mathcal{A}^*} \geq 0, \|\lambda_{\mathcal{A}^*}\|_1 \leq B} \left\| \nabla \phi(z^j) + \sum_{i \in \mathcal{A}^*} \lambda_i \nabla g_i(z^j) \right\|_1 \leq \bar{\eta}(z^j, \lambda_{\mathcal{A}^*}^j)^\sigma$$

for all $j \geq k$ in the EQ phase, so iterate j survives the fourth test in step 4.

We conclude that when the EQ phase is entered at some iteration k such that $\mathcal{A}(z^k, \lambda^k) = \mathcal{A}^*$ and (5.5) holds, where $\hat{\delta}$ satisfies the conditions (5.6), (5.7), (5.8), (5.9), (5.10), that all subsequent iterates remain in the EQ phase, contradicting our assertion that the iterates never eventually remain inside the EQ phase.

Therefore, unless Framework INEQ terminates finitely, the iterates eventually enter and remain permanently in the EQ phase for some fixed set $\mathcal{A}$. The final claims of the proof are an immediate consequence of Theorem 5.3. $\square$

Our final result is a more typical local result that assumes a starting point (z^0, λ^0) close to a solution $(z^*, \lambda^*) \in \mathcal{S}$ at which (2.4) is satisfied.

THEOREM 5.5. *Suppose that Assumption 1 holds for (1.1) at some $\lambda^* \in \mathcal{S}_\lambda$ and that the second-order condition (2.4) holds at λ^* . Then if the starting point (z^0, λ^0) is sufficiently close to (z^*, λ^*) , Framework INEQ will enter the EQ phase on the first iteration with $\mathcal{A} = \mathcal{A}^*$, and subsequently $\text{dist}((z^k, \lambda_{\mathcal{A}^*}^k), \bar{\mathcal{S}})$ converges Q -quadratically to zero and $\{z^k\}$ converges R -quadratically to z^* .*

Proof. We start by setting $\mathcal{V}_\lambda = \{\lambda^*\}$ and $\hat{\mathcal{V}}_\gamma = \{\lambda_{\mathcal{A}^*}^*\}$ for purposes of applying the results of sections 3 and 4. By choosing (z^0, λ^0) sufficiently close to (z^*, λ^*) , we

can ensure both that the active set $\mathcal{A}^*$ is identified correctly in step 1 and that the EQ phase is entered in step 2.

It remains to show that all subsequent iterates remain in the EQ phase, that is, that all iterations pass the tests in step 4. The logic here is essentially identical to the proof of Theorem 5.4, so we do not repeat it here. However, by forcing (z^0, λ^0) to be sufficiently close to (z^*, λ^*) we can prove a stronger result concerning the limit and the rate of convergence. We have from Theorem 3.4 that $\text{dist}((z^k, \lambda_{\mathcal{A}^*}^k), \bar{\mathcal{S}}) \leq \bar{\beta} \bar{\eta}(z^k, \lambda_{\mathcal{A}^*}^k)$ for all k , so that in fact $\text{dist}((z^k, \lambda_{\mathcal{A}^*}^k), \bar{\mathcal{S}}) \rightarrow 0$. Hence, we can apply Lemma 4.1 to obtain quadratic convergence of $\bar{\eta}(z^k, \lambda_{\mathcal{A}^*}^k)$ to zero, and by applying Theorem 3.4 again we have that $\text{dist}((z^k, \lambda_{\mathcal{A}^*}^k), \bar{\mathcal{S}})$ also converges Q -quadratically to zero. The R -quadratic convergence claim follows from $\|z^k - z^*\| \leq \text{dist}((z^k, \lambda_{\mathcal{A}^*}^k), \bar{\mathcal{S}})$. $\square$

In this result, the second-order condition and boundedness of the multiplier estimates for the original problem (1.1) are needed only to ensure that the identification of the active constraint set $\mathcal{A}^*$ in step 1 is immediately correct. For fast convergence of the EQ phase, we need only the (weaker) second-order condition for the problem $\text{EQ}(\mathcal{A}^*)$.

Theorem 5.5 is similar to the local convergence results in Hager [8], Fischer [6, Theorem 8], and Wright [17, Theorem 7] the main difference being that the second-order condition assumed here is less restrictive. It uses similar assumptions to the result in Izmailov and Solodov [11], which dealt with a different algorithm.

6. Discussion. Several interesting types of nonlinear optimization problems satisfy conditions such as those listed above. A simple case is the one in which the equality constrained problem

$$(6.1) \quad \min_z \phi(z) \quad \text{subject to } h(z) = 0$$

is reformulated by splitting the equality into two inequalities as follows:

$$(6.2) \quad \min_z \phi(z) \quad \text{subject to } h(z) \geq 0, \quad h(z) \leq 0.$$

Even when z^* is a solution of the original problem at which $\nabla h(z^*)$ has full rank (yielding a unique multiplier γ^*), the reformulation (6.2) satisfies no constraint qualifications. The KKT conditions for (6.2) are that there exist γ_+^* and γ_-^* such that

$$h(z^*) = 0, \quad \nabla \phi(z^*) - \nabla h(z^*)\gamma_+^* + \nabla h(z^*)\gamma_-^*, \quad \gamma_+^* \geq 0, \quad \gamma_-^* \geq 0.$$

Clearly, the multipliers for this formulation are neither unique nor bounded; any vectors γ_+^*, γ_-^* such that

$$(6.3) \quad \gamma^* = \gamma_-^* - \gamma_+^*, \quad \gamma_+^* \geq 0, \quad \gamma_-^* \geq 0,$$

will suffice. The second-order sufficient condition (2.9) for (6.1) is that

$$(6.4) \quad \begin{aligned} & w^T (\nabla^2 \phi(z^*) + \sum_{i=1}^m \gamma_i^* \nabla^2 h_i(z^*)) w \geq v \|w\|^2 \\ & \text{for all } w \text{ such that } \nabla h_i(z^*)^T w = 0, \quad i = 1, 2, \dots, p. \end{aligned}$$

Since $\mathcal{A}_0^* = \emptyset$, we have following (2.4) that the second-order condition for (6.2) is that

$$(6.5) \quad \begin{aligned} & w^T (\nabla^2 \phi(z^*) - \sum_{i=1}^m (\gamma_+^*)_i \nabla^2 h_i(z^*) + \sum_{i=1}^m (\gamma_-^*)_i \nabla^2 h_i(z^*)) w \geq v \|w\|^2 \\ & \text{for all } w \text{ such that } \nabla h_i(z^*)^T w = 0, \quad i = 1, 2, \dots, p. \end{aligned}$$

Obviously, the set of directions in (6.5) is identical to the set in (6.4). It is easy to see that the relation (6.4) holds for *all* γ_+^* , γ_-^* satisfying (6.3). Hence, the second-order assumption in Theorem 5.4 holds for any choice of the multiplier bound c . Therefore, for any “reasonable” outer strategy which ensures boundedness of the multipliers and convergence of $(z^k, \gamma_+^k, \gamma_-^k)$ to the solution set for (6.2), Framework INEQ will converge superlinearly. The stabilization term in the step calculation (4.1), (4.2) overcomes the potentially woeful consequences of the inferior formulation (6.2).

More general examples of a similar type, in which some of the inequality constraints in (1.1) are fixed linear combinations of others, can also be constructed.

Another class of interest is mathematical programs with equilibrium constraints (MPECs) which, when formulated as nonlinear programs in the usual ways, fail to satisfy constraint qualifications at any feasible point. The second-order condition (2.4) corresponds in this case to the second-order condition for the relaxed nonlinear programming formulation (Fletcher et al. [7, eq. (3.2)]). (These are slightly more restrictive than the second-order condition for the MPEC itself.) Izmailov and Solodov [11, section 4] give examples in which these conditions are satisfied locally for some multipliers and show convergence of their algorithm in these cases. Theorem 5.5 implies that Framework INEQ has similar rapid local convergence properties in these cases.

We illustrate the performance of the constraint identification and convergence properties of our approach using the following MPEC in two variables:

$$(6.6) \quad \min z_2 \quad \text{subject to} \quad z_1 \geq 0, \quad z_2 \geq 0, \quad z_1 z_2 \leq 0, \quad z_2^2 \geq 1,$$

whose solution is $z^* = (0, 1)$ with optimal objective 1 and active set $\mathcal{A}^* = \{1, 3, 4\}$ (where we assume that the constraints are ordered as written). The matrix of active constraint gradients is obviously rank deficient and the optimal multiplier set for this problem is

$$\mathcal{S}_\lambda = \{(\alpha, 0, \alpha, 0.5) \mid \alpha \geq 0\},$$

so that $\mathcal{A}_+^* = \{1, 3, 4\}$ and $\mathcal{A}_0^* = \emptyset$. The second-order conditions for this problem are satisfied vacuously, since the direction set in (2.4) is simply the zero vector.

This problem has the property that standard SQP may fail to produce a valid step for some z arbitrarily close to z^* , since the linearized constraints are inconsistent at $z = (\epsilon, 1 - \epsilon)$ for all small positive ϵ [7, section 2.2]. We consider the performance of one step of Framework INEQ from such a point, along with a Lagrange multiplier estimate that is a similar distance from $\mathcal{S}_\lambda$. We set

$$(6.7) \quad z^0 = (\epsilon, 1 - \epsilon), \quad \lambda^0 = (1, 0, 1, .5) + \epsilon v,$$

where $v \in \mathbb{R}^4$ is a vector whose elements are drawn from a uniform distribution over $[-1, 1]$. Obviously, we have that $\text{dist}((z^0, \lambda^0), \mathcal{S}) = O(\epsilon)$ and it is easy to show that $\eta(z^0, \lambda^0) = O(\epsilon)$ as well (in accordance with Theorem 3.2). For all ϵ sufficiently small, we have that $\mathcal{A}(z^0, \lambda^0) = \{1, 3, 4\} = \mathcal{A}^*$ and also that $\bar{\eta}(z^0, \lambda_{\mathcal{A}^*}^0) = O(\epsilon)$. Hence, the test in step 2 of Framework INEQ is satisfied, and the method enters the EQ phase.

Table 6.1 shows the effect of the first step taken in the EQ phase for various values of ϵ . (The results were obtained from a Matlab implementation.) The second column shows that the estimate of $\mathcal{A}$ is precise for all values tried (though it is usually inaccurate for $\epsilon \geq 1/4$). The third and fourth columns show the values of $\bar{\eta}(z^k, \lambda_{\mathcal{A}^*}^k)$ for $k = 0$ and $k = 1$. (We use the notation $3.9(-1)$ to denote 3.9×10^{-1} , and so on.) The final column gives an estimate of the Q -order of convergence of $\bar{\eta}$ to 0 over this step. The Q -quadratic convergence of Theorem 5.5 is apparent.

TABLE 6.1
One step of framework INEQ from starting point (6.7).

ϵ	$\mathcal{A}$	$\bar{\eta}_0$	$\bar{\eta}_1$	$\log \bar{\eta}_1 / \log \bar{\eta}_0$
2^{-3}	$\{1, 3, 4\}$	3.9(-1)	3.9(-2)	3.5
2^{-5}	$\{1, 3, 4\}$	1.1(-1)	1.2(-3)	3.0
2^{-10}	$\{1, 3, 4\}$	3.0(-3)	2.1(-6)	2.2
2^{-15}	$\{1, 3, 4\}$	1.0(-4)	1.4(-9)	2.2
2^{-20}	$\{1, 3, 4\}$	3.1(-6)	1.4(-12)	2.1
2^{-30}	$\{1, 3, 4\}$	2.7(-9)	7.1(-19)	2.1
2^{-40}	$\{1, 3, 4\}$	2.5(-12)	4.2(-25)	2.1

The linear system (4.1), (4.2) that is solved to obtain the step tabulated in Table 6.1 is evidently ill conditioned for small ϵ . (We see this by noting that $\nabla h(z) \in \mathbb{R}^{2 \times 3}$ has column rank 2, while the 3×3 diagonal block below it has $\mu = O(\epsilon)$ and hence is approaching zero.) It is interesting to ask, therefore, if the computed version of the step $(\Delta z, \Delta \gamma)$ will differ from the exact version so markedly as to destroy the rapid convergence properties of the algorithm. This issue has essentially been analyzed in [15]. The analysis takes into account both the errors in evaluating the matrix and right-hand side of (4.2) and the roundoff errors that occur in factorizing and solving this system using a numerically stable scheme. The evaluation errors may be particularly significant in the right-hand side of (4.2), since while all components of this vector are theoretically of size $O(\bar{\eta})$, they are generally computed by manipulating intermediate quantities of size $O(1)$, so may contain errors of absolute size similar to $\mathbf{u}$, where $\mathbf{u}$ is unit roundoff.

The analysis in [15], translated appropriately into the current setting, is based on a singular value decomposition of $\nabla h(z^*)$, namely,

$$(6.8) \quad \nabla h(z^*) = \begin{bmatrix} \hat{U} & \hat{V} \end{bmatrix} \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U^T \\ V^T \end{bmatrix},$$

where S is positive diagonal, and $\hat{U}$, $\hat{V}$, U , and V are orthonormal matrices. While the exact step $(\Delta z, \Delta \gamma)$ is of size $O(\bar{\eta})$ (Lemma 4.1), it is shown in [15] that the step $(\widehat{\Delta z}, \widehat{\Delta \gamma})$ computed under the assumptions above satisfies

$$(\widehat{\Delta z} - \Delta z, U^T(\widehat{\Delta \gamma} - \Delta \gamma)) = O(\bar{\eta}^2) + \delta_{\mathbf{u}},$$

where $\delta_{\mathbf{u}}$ represents a vector whose entries are bounded by a modest multiple of $\mathbf{u}$. The errors introduced into these components are as modest as we could expect, given that the right-hand side contains absolute errors similar to $\mathbf{u}$. In the remaining component of the step, however, the error is more severe; we have

$$V^T(\widehat{\Delta \gamma} - \Delta \gamma) = O(\bar{\eta} + \bar{\eta}^{-1}\delta_{\mathbf{u}}).$$

In this subspace, the ill conditioning of the matrix combines with the evaluation and roundoff errors to produce large differences between true and computed steps. However, as noted in [15], the large error in this space has little effect on the convergence behavior of the algorithm. Roughly speaking, we have

$$\bar{\eta}(z + \widehat{\Delta z}, \gamma + \widehat{\Delta \gamma}) = O(\bar{\eta}(z, \gamma)^2 + \delta_{\mathbf{u}}),$$

TABLE 6.2

One step of framework INEQ from starting point (6.7), with simulated roundoff errors.

ϵ	$\bar{\eta}_0$	$\bar{\eta}_1$	$\log \bar{\eta}_1 / \log \bar{\eta}_0$	$(\widehat{\Delta\lambda}_1, \widehat{\Delta\lambda}_3)$
2^{-5}	9.4(-2)	2.9(-3)	2.5	$10^{-2}(.53, .57)$
2^{-10}	2.8(-3)	3.3(-6)	2.1	$10^{-3}(.12, .23)$
2^{-15}	9.9(-5)	1.5(-9)	2.2	$10^{-4}(-.006, .1)$
2^{-20}	3.0(-6)	2.1(-12)	2.1	$10^{-6}(.06, .84)$
2^{-30}	2.7(-9)	2.1(-12)	1.4	$10^{-4}(-.48, -.48)$
2^{-40}	2.7(-12)	1.0(-12)	1.0	$(-.33, -.33)$

so that we can expect to see finite-precision arithmetic interfering with the quadratic convergence of $\bar{\eta}$ only when $\bar{\eta}$ drops below $\sqrt{\mathbf{u}}$.

Our example (6.6) is too small and simple to exhibit the errors discussed in the previous paragraph, but we can simulate the effects of such errors by introducing arbitrary perturbations into (4.2) and tracking the effects of these perturbations on the computed step. For this “simulated roundoff” experiment, we set $\mathbf{u} = 10^{-12}$ and add perturbations of size $\mathbf{u}(\|A\|_2/2)\rho$ (where A is the coefficient matrix in (4.2) and ρ is a scalar drawn from the uniform distribution on $[-1, 1]$) to all components of the matrix and right-hand side in (4.2). For our example, we have

$$\nabla h(z^*) = \begin{bmatrix} -1 & 1 & 0 \\ 0 & 0 & -2 \end{bmatrix}, \quad V = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix},$$

so we expect to see errors in the computed values of the Lagrange multiplier step of the following form:

$$(6.9) \quad \begin{bmatrix} \Delta\lambda_1 - \widehat{\Delta\lambda}_1 \\ \Delta\lambda_3 - \widehat{\Delta\lambda}_3 \\ \Delta\lambda_4 - \widehat{\Delta\lambda}_4 \end{bmatrix} \approx \beta \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \approx - \begin{bmatrix} \widehat{\Delta\lambda}_1 \\ \widehat{\Delta\lambda}_3 \\ \widehat{\Delta\lambda}_4 \end{bmatrix},$$

where $|\beta| = O(\bar{\eta} + \bar{\eta}^{-1}\mathbf{u})$.

Table 6.2 shows the behavior of the first step of Framework INEQ, computed as in Table 6.1 but with the introduction of simulated roundoff errors as described above. Interference of these errors with the theoretical convergence behavior is evident only in the last two lines of the table, when $\bar{\eta}_0$ drops below about 10^{-6} . The final column verifies that, as predicted in (6.9), the size of the computed multipliers essentially tracks ϵ for $\epsilon \geq \sqrt{\mathbf{u}}$, but grows in the range space of V like $\bar{\eta}^{-1}\mathbf{u}$ for smaller values of $\bar{\eta}$.

Acknowledgments. I am grateful to the editor Sven Leyffer, to the referees, and to Mihai Anitescu, Mikhail Solodov, and Alexey Izmailov for helpful comments on the first version of the paper.

REFERENCES

- [1] M. ANITESCU, *Degenerate nonlinear programming with a quadratic growth condition*, SIAM J. Optim., 10 (2000), pp. 1116–1135.
- [2] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

- [3] J. F. BONNANS AND A. IOFFE, *Second-order sufficiency and quadratic growth for nonisolated minima*, Math. Oper. Res., 20 (1995), pp. 801–819.
- [4] F. FACCHINEI, A. FISCHER, AND C. KANZOW, *On the accurate identification of active constraints*, SIAM J. Optim., 9 (1998), pp. 14–32.
- [5] A. FISCHER, *Modified Wilson method for nonlinear programs with nonunique multipliers*, Math. Oper. Res., 24 (1999), pp. 699–727.
- [6] A. FISCHER, *Local behavior of an iterative framework for generalized equations with nonisolated solutions*, Math. Program. Ser. A, 94 (2002), pp. 91–124.
- [7] R. FLETCHER, S. LEYFFER, D. RALPH, AND S. SCHOLTES, *Local Convergence of SQP Methods for Mathematical Programs with Equilibrium Constraints*, Numerical Analysis Report NA/209, Department of Mathematics, University of Dundee, May 2002.
- [8] W. W. HAGER, *Stabilized sequential quadratic programming*, Comput. Optim. Appl., 12 (1999), pp. 253–273.
- [9] W. W. HAGER AND M. S. GOWDA, *Stability in the presence of degeneracy and error estimation*, Math. Program. Ser. A, 85 (1999), pp. 181–192.
- [10] M. R. HESTENES, *Multiplier and gradient methods*, J. Optim. Theory Appl., 4 (1969), pp. 303–320.
- [11] A. F. IZMAILOV AND M. V. SOLODOV, *Newton-type methods for optimization problems without constraint qualifications*, SIAM J. Optim., 15 (2004), pp. 210–228.
- [12] J. NOCEDAL AND S. J. WRIGHT, *Numerical Optimization*, Springer, New York, 1999.
- [13] M. J. D. POWELL, *A method for nonlinear constraints in minimization problems*, in Optimization, R. Fletcher, ed., Academic Press, New York, 1969, pp. 283–298.
- [14] S. M. ROBINSON, *Generalized equations and their solutions. Part II: Applications to nonlinear programming*, Math. Program. Stud., 19 (1982), pp. 200–221.
- [15] S. J. WRIGHT, *Superlinear convergence of a stabilized SQP method to a degenerate solution*, Comput. Optim. Appl., 11 (1998), pp. 253–275.
- [16] S. J. WRIGHT, *Modifying SQP for degenerate problems*, SIAM J. Optim., 13 (2002), pp. 470–497.
- [17] S. J. WRIGHT, *Constraint identification and algorithm stabilization for degenerate nonlinear programs*, Math. Program. Ser. B, 95 (2003), pp. 137–160.

GENERALIZED LAGRANGIAN DUALS AND SUMS OF SQUARES RELAXATIONS OF SPARSE POLYNOMIAL OPTIMIZATION PROBLEMS*

SUNYOUNG KIM[†], MASAKAZU KOJIMA[‡], AND HAYATO WAKI[†]

Abstract. Sequences of generalized Lagrangian duals and their sums of squares (SOS) of polynomials relaxations for a polynomial optimization problem (POP) are introduced. The sparsity of polynomials in the POP is used to reduce the sizes of the Lagrangian duals and their SOS relaxations. It is proved that the optimal values of the Lagrangian duals in the sequence converge to the optimal value of the POP using a method from the penalty function approach. The sequence of SOS relaxations is transformed into a sequence of semidefinite programming (SDP) relaxations of the POP, which correspond to duals of modification and generalization of SDP relaxations given by Lasserre for the POP.

Key words. polynomial optimization problem, sparsity, global optimization, Lagrangian relaxation, Lagrangian dual, sums of squares optimization, semidefinite program, semidefinite program relaxation

AMS subject classifications. 90C22, 90C25, 49M29

DOI. 10.1137/030601260

1. Introduction. Polynomial optimization problems (POPs) or optimization problems with polynomial objective and constraints represent a broad range of applications in science and engineering. Convex relaxation methods that relax nonconvex constraints and objective function of POPs with convex constraints have been widely used for POPs. Various convex relaxation methods exist for POPs, which include quadratic, linear, and 0-1 integer programming as special cases. After a lift-and-project linear programming (LP) procedure for 0-1 integer linear programs by Balas, Ceria, and Cornuéjols [1], the reformulation-linearization technique (RLT) by Sherali and Adams [18] and a semidefinite programming (SDP) relaxation method by Lovász and Schrijver [11] were introduced. Many convex relaxation methods such as the RLT [20, 19] for 0-1 mixed integer polynomial programs, the successive convex relaxation method (SCRM) [8, 9] for quadratic optimization problems (QOPs), second order cone programming (SOCP) relaxations [4, 5] for QOPs, and sums of squares (SOS) relaxations for POPs [13, 14] have been proposed.

Recently, an important theoretical development has been made by Lasserre [10] toward achieving optimal values of POPs. His method for obtaining a sequence of SDP relaxations can be considered as a primal approach, which is divided into two phases, according to the paper [6]. The first phase is to transform a given POP into an equivalent optimization problem by adding valid polynomial constraints over

*Received by the editors October 19, 2003; accepted for publication (in revised form) July 15, 2004; published electronically April 8, 2005.

<http://www.siam.org/journals/siopt/15-3/60126.html>

[†]Department of Mathematics, Ewha Women's University, 11-1 Dahyun-dong, Sudaemoon-gu, Seoul 120-750, Korea (skim@ewha.ac.kr). A considerable part of this work was conducted while this author was visiting Tokyo Institute of Technology. Research was supported by Kosef R004-000-2001-00200.

[‡]Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, 2-12-1 Oh-okayama, Meguro-ku, Tokyo 152-8552, Japan (kojima@is.titech.ac.jp, hayato.waki@is.titech.ac.jp). The research of the second author was supported by Grant-in-Aid for Scientific Research on Priority Areas 16016234.

semidefinite cones to the POP. The resulting optimization problem forms a polynomial semidefinite programming problem (PSDP). In the second phase, the PSDP is linearized to a semidefinite program. In theory, we can add any size and/or number of valid polynomial constraints over positive semidefinite cones to the POP in the first phase. Lasserre [10] proposed a systematic way of adding valid polynomial constraints over semidefinite cones to the original POP in the first phase to generate a sequence of SDP relaxations in the second phase. It was proved that when the feasible region of the POP is compact, its optimal value can be approximated within any accuracy by the sequence of SDP relaxations. However, the size of an SDP relaxation to be solved in the sequence increases very rapidly as higher accuracy for an approximation to the optimal value of the POP is required.

It is known that practical solvability of semidefinite programs depends on their sizes. A great deal of effort continues to be made to resolve the difficulty of solving large-scale semidefinite programs. See the survey paper [12] and references therein. From a computational point of view, a close approximation to the optimal value of a POP is hard to achieve without efficient methods for handling large-scale semidefinite programs. The rapid increase in the size of SDP relaxations in the sequence in [10] as higher accuracy is desired makes it difficult to achieve their optimal values in practice. As a result, only small-size POPs can be solved by the method in [10].

The purpose of this paper is to present generalized Lagrangian duals and their SOS relaxations [13] for sparse POPs. This approach may be regarded as a dual of Lasserre's SDP relaxations [10] mentioned above. For the standard Lagrangian function of a POP, it is common for nonnegative numbers to be assigned to Lagrangian multipliers. Instead of selecting nonnegative numbers for Lagrangian multipliers, we choose Lagrangian multipliers to be SOS polynomials satisfying sparsity which is similar to the sparsity of the associated constraint polynomials. Then, we define a generalized Lagrangian dual for a POP over such SOS polynomial multipliers and provide a theoretical foundation for the generalized Lagrangian dual. After a sequence of sets of SOS polynomials is constructed, e.g., SOS polynomials of increasing degree, for Lagrangian multipliers, a sequence of Lagrangian duals is obtained. We derive a sufficient condition for the sequence of Lagrangian duals to attain the optimal value of the POP, based on the idea of the penalty function method. For practical purposes, each Lagrangian dual in the sequence is relaxed to an SOS optimization problem, which is further converted into an equivalent semidefinite program. Thus we have sequences of SOS relaxations and SDP relaxations of the POP. The resulting sequence of SDP relaxations corresponds to the dual of the sequence of SDP relaxations obtained from the primal approach.

An advantage of the dual approach in this paper is that the sparsity of objective and constraint polynomials in a POP can be exploited to reduce the size of the SDP relaxations. The size of the SDP relaxations depends on the supports of the polynomials in the dual approach, whereas the size of the SDP relaxations from the primal approach depends on the degree of the polynomials.

This paper is organized as follows. In section 2, we introduce a POP with a compact feasible region and describe the representation of the sparsity of the POP. Two types of POPs are shown for different characterizations of the compact feasible region. Section 3 includes the derivation of generalized Lagrangian duals of the two POPs over Lagrangian multipliers from a set of SOS polynomials satisfying sparsity which is similar to the sparsity of the associated constraint polynomials. We show a sufficient condition for the Lagrangian duals to attain the optimal value of the original POP. In

section 4, we present numerical methods to approximate optimal values of the generalized Lagrangian duals numerically using SOS relaxations. The numerical methods take advantage of the sparsity of the original POP to produce SDP relaxations with smaller size. The relationship between the resulting SDP relaxations and the SDP relaxations in [10] is also discussed. Section 5 contains the proof of a lemma which plays an important role in section 3. In section 6, we report preliminary numerical results for a sparse and structured POP and show a computational advantage of the proposed SOS and SDP relaxations. Section 7 is devoted to concluding discussions.

Throughout the paper, we use the following notation: Let $\mathbb{R}^n$ and $\mathbb{Z}_+^n \subset \mathbb{R}^n$ denote the n -dimensional Euclidean space and the set of n -dimensional nonnegative integer vectors, respectively. Let $f_j : \mathbb{R}^n \rightarrow \mathbb{R}$ be a real valued polynomial in $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ ($j = 0, 1, 2, \dots, m$). We denote each polynomial $f_j(\mathbf{x})$ as $f_j(\mathbf{x}) = \sum_{\mathbf{a} \in \mathcal{F}_j} c_j(\mathbf{a}) \mathbf{x}^{\mathbf{a}}$, where a nonempty finite subset $\mathcal{F}_j$ of $\mathbb{Z}_+^n$ denotes a support of the polynomial $f_j(\mathbf{x})$, $c_j(\mathbf{a}) \in \mathbb{R}$, and $\mathbf{x}^{\mathbf{a}} = x_1^{a_1} x_2^{a_2} \dots x_n^{a_n}$ for every $\mathbf{a} = (a_1, a_2, \dots, a_n) \in \mathcal{F}_j$ ($j = 0, 1, 2, \dots, m$) and $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$. Let r_j denote the degree of each polynomial $f_j(\mathbf{x})$ ($j = 0, 1, 2, \dots, m$); $r_j = \max\{\sum_{i=1}^n a_i : \mathbf{a} \in \mathcal{F}_j\}$.

2. POPs and sparsity. We consider the POP

$$(1) \quad \text{minimize } f_0(\mathbf{x}) \quad \text{subject to } f_j(\mathbf{x}) \geq 0 \quad (j = 1, 2, \dots, m).$$

Let us focus on the support $\mathcal{F}_j$ of the polynomial $f_j(\mathbf{x})$ ($j = 0, 1, 2, \dots, m$) to describe the sparsity of the POP (1). A polynomial $f(\mathbf{x})$ of degree r or its support $\mathcal{F}$ is called sparse if the number of elements in the support $\mathcal{F}$ is much smaller than the number of elements in the support $\mathcal{G}(\xi) \equiv \{\mathbf{a} \in \mathbb{Z}_+^n : \sum_{i=1}^n a_i \leq \xi\}$ of a general fully dense polynomial of degree ξ . In particular, if the number of indices in $I_+(\mathcal{F}) \equiv \{i : a_i > 0 \text{ for some } \mathbf{a} \in \mathcal{F}\}$ is much smaller than n , then $f(\mathbf{x})$ is sparse. We present two examples below.

Example 2.1 (A box constraint POP). Let $m = n$ and $f_j(\mathbf{x}) = 1 - x_j^2$ ($j = 1, 2, \dots, n$). In this case, we have $\mathcal{F}_j = \{\mathbf{0}, 2\mathbf{e}^j\}$ ($j = 1, 2, \dots, n$). Each $\mathcal{F}_j$ has two elements and $I_+(\mathcal{F}_j) = \{j\}$. Here $\mathbf{e}^j$ denotes the j th unit coordinate vector of $\mathbb{R}^n$ with 1 in the j th component and 0 elsewhere.

Example 2.2. Let $m = n$, $f_1(\mathbf{x}) = \alpha_1 - x_1^2$, and $f_j(\mathbf{x}) = \alpha_j - (\beta_{j-1}x_{j-1} - x_j)^2$ ($j = 2, 3, \dots, n$), where α_j ($j = 1, 2, \dots, n$) and β_j ($j = 1, 2, \dots, n-1$) denote positive numbers. Then, $\mathcal{F}_1 = \{\mathbf{0}, 2\mathbf{e}^1\}$, $\mathcal{F}_j = \{\mathbf{0}, 2\mathbf{e}^{j-1}, \mathbf{e}^{j-1} + \mathbf{e}^j, 2\mathbf{e}^j\}$ ($j = 2, 3, \dots, n$). Each $\mathcal{F}_j$ has at most four elements and $I_+(\mathcal{F}_j)$ contains at most two indices.

Another example is given in section 6 with preliminary numerical results.

Let F denote the feasible region of the POP (1):

$$F = \{\mathbf{x} \in \mathbb{R}^n : f_j(\mathbf{x}) \geq 0 \quad (j = 1, 2, \dots, m)\}.$$

Throughout the paper, we assume that F is nonempty and bounded. Then, POP (1) has a finite optimal value ζ^* at an optimal solution $\mathbf{x}^* \in F$:

$$\zeta^* = \min\{f_0(\mathbf{x}) : \mathbf{x} \in F\} = f_0(\mathbf{x}^*) \text{ and } \mathbf{x}^* \in F.$$

In what follows, we further need a bound $\rho > 0$ to be known explicitly for the feasible region. We are concerned with the following cases:

- $F \subset C_\rho \equiv \{\mathbf{x} \in \mathbb{R}^n : \rho^2 - x_i^2 \geq 0 \quad (i = 1, 2, \dots, n)\}$;
- $F \subset B_\rho \equiv \{\mathbf{x} \in \mathbb{R}^n : \rho^2 - \mathbf{x}^T \mathbf{x} \geq 0\}$.

If $F \subset C_\rho$ holds, then POP (1) is equivalent to

$$(2) \quad \text{minimize } f_0(\mathbf{x}) \quad \text{subject to } f_j(\mathbf{x}) \geq 0 \ (j = 1, 2, \dots, m) \text{ and } \mathbf{x} \in C_\rho.$$

Example 2.1 is a special case of the POP (2), where we take $m = 0$ and $\rho = 1$. If $F \subset B_\rho$ is satisfied, then we have $F \subset C_\rho$; hence the two POPs (1) and (2) are equivalent to

$$(3) \quad \text{minimize } f_0(\mathbf{x}) \quad \text{subject to } f_j(\mathbf{x}) \geq 0 \ (j = 1, 2, \dots, m) \text{ and } \mathbf{x} \in B_\rho.$$

In the next section, we present a (generalized) Lagrangian function, a Lagrangian relaxation, and a Lagrangian dual and its SOS relaxation for each of the POPs (2) and (3). For both POPs, the Lagrangian dual converges to the optimal value ζ^* of the original POP (1) under a moderate assumption. But only the SOS relaxation derived from the POP (3) is guaranteed to converge to ζ^* , while the SOS relaxation of the POP (2) inherits more sparsity from the original POP (1) than the SOS relaxation of the POP (3).

3. Generalized Lagrangian duals.

3.1. Lagrangian functions. Let $\bar{\Sigma}$ denote the set of sums of squares of polynomials in $\mathbf{x} \in \mathbb{R}^n$,

$$\bar{\Sigma} = \left\{ \sum_{i=1}^k \chi_i(\mathbf{x})^2 : \begin{array}{l} \chi_i \text{ is a polynomial in } \mathbf{x} \in \mathbb{R}^n \ (i = 1, 2, \dots, k) \\ \text{and } k \text{ is any finite positive integer.} \end{array} \right\}$$

We define two types of (generalized) Lagrangian functions $L_B : B_\rho \times \bar{\Sigma}^m \rightarrow \mathbb{R}$ for POP (3) and $L_C : \mathbb{R}^n \times \bar{\Sigma}^{m+n} \rightarrow \mathbb{R}$ for POP (2):

$$\begin{aligned} L_B(\mathbf{x}, \boldsymbol{\varphi}) &= f_0(\mathbf{x}) - \sum_{j=1}^m \varphi_j(\mathbf{x}) f_j(\mathbf{x}), \\ L_C(\mathbf{x}, \boldsymbol{\varphi}, \boldsymbol{\psi}) &= L_B(\mathbf{x}, \boldsymbol{\varphi}) - \sum_{i=1}^n \psi_i(\mathbf{x}) (\rho^2 - x_i^2) \\ &= f_0(\mathbf{x}) - \sum_{j=1}^m \varphi_j(\mathbf{x}) f_j(\mathbf{x}) - \sum_{i=1}^n \psi_i(\mathbf{x}) (\rho^2 - x_i^2). \end{aligned}$$

Here $\bar{\Sigma}^\ell$ denotes the Cartesian product of ℓ -tuples of $\bar{\Sigma}$,

$$\bar{\Sigma}^\ell = \{(\varphi_1, \varphi_2, \dots, \varphi_\ell) : \varphi_j \in \bar{\Sigma} \ (j = 1, 2, \dots, \ell)\} \quad (\ell = m \text{ or } m+n).$$

Let $(\boldsymbol{\varphi}, \boldsymbol{\psi}) \in \bar{\Sigma}^{m+n}$. Then

$$(4) \quad \left. \begin{aligned} L_C(\mathbf{x}, \boldsymbol{\varphi}, \boldsymbol{\psi}) &\leq f_0(\mathbf{x}) \quad \text{if } \mathbf{x} \in F \cap C_\rho, \\ L_B(\mathbf{x}, \boldsymbol{\varphi}) &\leq f_0(\mathbf{x}) \quad \text{if } \mathbf{x} \in F \cap B_\rho, \\ L_C(\mathbf{x}, \boldsymbol{\varphi}, \boldsymbol{\psi}) &\leq L_B(\mathbf{x}, \boldsymbol{\varphi}) \quad \text{if } \mathbf{x} \in B_\rho. \end{aligned} \right\}$$

3.2. Lagrangian relaxations and duals. We introduce a (generalized) Lagrangian relaxation of the POP (2):

$$L_C^*(\boldsymbol{\varphi}, \boldsymbol{\psi}) = \inf \{L_C(\mathbf{x}, \boldsymbol{\varphi}, \boldsymbol{\psi}) : \mathbf{x} \in \mathbb{R}^n\}$$

for each fixed $(\varphi, \psi) \in \overline{\Sigma}^{m+n}$, and a (generalized) Lagrangian relaxation of the POP (3):

$$L_B^*(\varphi) = \inf \{L_B(\mathbf{x}, \varphi) : \mathbf{x} \in B_\rho\}$$

for each fixed $\varphi \in \overline{\Sigma}^m$. Let $(\varphi, \psi) \in \overline{\Sigma}^{m+n}$. By (4), we see that

$$(5) \quad \left. \begin{aligned} L_C^*(\varphi, \psi) &\leq \zeta^* = \min\{f_0(\mathbf{x}) : \mathbf{x} \in F\} \text{ if } F \subset C_\rho, \\ L_C^*(\varphi, \psi) &\leq L_B^*(\varphi) \leq \zeta^* \text{ if } F \subset B_\rho. \end{aligned} \right\}$$

For every $(\Sigma, \Xi) \subset \overline{\Sigma}^{m+n}$, we define a (generalized) Lagrangian dual of the POP (2):

$$(6) \quad \text{maximize } L_C^*(\varphi, \psi) \quad \text{subject to } (\varphi, \psi) \in \Sigma \times \Xi.$$

For every $\Sigma \subset \overline{\Sigma}^m$, we define a (generalized) Lagrangian dual of the POP (3):

$$(7) \quad \text{maximize } L_B^*(\varphi) \quad \text{subject to } \varphi \in \Sigma.$$

Let $L_C^*(\Sigma \times \Xi)$ and $L_B^*(\Sigma)$ denote the optimal values of (6) and (7), respectively:

$$L_C^*(\Sigma \times \Xi) = \sup_{(\varphi, \psi) \in \Sigma \times \Xi} L_C^*(\varphi, \psi) \quad \text{and} \quad L_B^*(\Sigma) = \sup_{\varphi \in \Sigma} L_B^*(\varphi).$$

It follows from (5) that

$$(8) \quad \left. \begin{aligned} L_C^*(\Sigma \times \Xi) &\leq \zeta^* \text{ if } F \subset C_\rho, \\ L_C^*(\Sigma \times \Xi) &\leq L_B^*(\Sigma) \leq \zeta^* \text{ if } F \subset B_\rho \end{aligned} \right\}$$

holds for every $(\Sigma, \Xi) \subset \overline{\Sigma}^{m+n}$.

Assume that $F \subset C_\rho$. Then, the two POPs (1) and (2) are equivalent. If we restrict (φ, ψ) to the nonnegative orthant $\mathbb{R}_+^{m+n}$ of $\mathbb{R}^{m+n}$, $L_C(\mathbf{x}, \varphi, \psi)$ becomes the standard Lagrangian function for the POP (2). It is well known that a positive duality gap exists in general between the standard Lagrangian dual $L_C^*(\mathbb{R}_+^{m+n})$ and the optimal value ζ^* of the POP (2). In fact, consider a simple example of $f_0(x) = x^3$ ($\forall x \in \mathbb{R}$), where $n = 1$, $m = 0$. Then

$$\begin{aligned} \zeta^* &= \min \{x^3 : x^2 \leq 1\} = -1, \\ L_C(x, \psi) &= x^3 - \psi(1 - x^2) \quad (\forall x \in \mathbb{R}, \forall \psi \in \mathbb{R}_+), \\ L_C^*(\psi) &= -\infty \quad (\forall \psi \in \mathbb{R}_+); \quad \text{hence } L_C^*(\mathbb{R}_+) = -\infty. \end{aligned}$$

As we take a larger set $\Sigma \times \Xi \subset \overline{\Sigma}^{m+n}$, the duality gap between $L_C^*(\Sigma \times \Xi)$ and ζ^* is expected to decrease. In the next section, we derive a sufficient condition on $\Sigma \times \Xi \subset \overline{\Sigma}^{m+n}$ for $L_C^*(\Sigma \times \Xi)$ to attain the optimal value ζ^* of the POP (2). We regard $(\varphi, \psi) \in \overline{\Sigma}^{m+n}$ as “a penalty parameter” in the derivation, and

$$(9) \quad \Phi_C(\mathbf{x}, \varphi, \psi) = -\sum_{j=1}^m \varphi_j(\mathbf{x}) f_j(\mathbf{x}) - \sum_{i=1}^n \psi_i(\mathbf{x}) (\rho^2 - x_i^2)$$

(the terms added to the objective function $f_0(\mathbf{x})$)

in the construction of the Lagrangian function $L_C(\mathbf{x}, \varphi, \psi)$)

as “a penalty function” with a choice of penalty parameters $(\varphi, \psi) = (\varphi^p, \psi^p) \in \bar{\Sigma}^{m+n}$ ($p \in \mathbb{Z}_+$) such that

$$\begin{aligned} & \text{if } \mathbf{x} \in F, \text{ then } \Phi(\mathbf{x}, \varphi^p, \psi^p) \rightarrow 0 \text{ as } p \rightarrow \infty, \\ & \text{if } \mathbf{x} \notin F, \text{ then } \Phi(\mathbf{x}, \varphi^p, \psi^p) \rightarrow \infty \text{ as } p \rightarrow \infty. \end{aligned}$$

Additional properties of the penalty function $\Phi_C(\mathbf{x}, \varphi^p, \psi^p)$ are described in Lemma 3.2. It is shown that the Lagrangian function $L_C(\mathbf{x}, \varphi, \psi) = f_0(\mathbf{x}) + \Phi_C(\mathbf{x}, \varphi, \psi)$ with the penalty parameter $(\varphi, \psi) = (\varphi^p, \psi^p) \in \bar{\Sigma}^{m+n}$ has a global minimizer $\mathbf{x}^p$ over $\mathbb{R}^n$ with the objective value $f_0(\mathbf{x}^p) \rightarrow \zeta^*$ as $p \rightarrow \infty$ and that $\{\mathbf{x}^p\}$ has an accumulation point in the optimal solution set of the POP (2).

3.3. Main theorem. We need some additional notation and symbols. Take a real number $\gamma \geq 1$ such that

$$(10) \quad \left. \begin{aligned} |f_j(\mathbf{x})/\gamma| &\leq 1 \text{ if } \|\mathbf{x}\|_\infty \leq \sqrt{2}\rho, \\ |f_j(\mathbf{x})/\gamma| &\leq \|\mathbf{x}/\rho\|_\infty^r \text{ if } \|\mathbf{x}\|_\infty \geq \sqrt{2}\rho \end{aligned} \right\}$$

($j = 0, 1, 2, \dots, m$). Define

$$(11) \quad \left. \begin{aligned} \varphi_j^p(\mathbf{x}) &= (1 - f_j(\mathbf{x})/\gamma)^{2p} \quad (j = 1, 2, \dots, m, \quad p \in \mathbb{Z}_+), \\ \varphi^p(\mathbf{x}) &= (\varphi_1^p(\mathbf{x}), \varphi_2^p(\mathbf{x}), \dots, \varphi_m^p(\mathbf{x})) \quad (p \in \mathbb{Z}_+), \\ \psi_i^p(\mathbf{x}) &= ((m+2)\gamma/\rho^2) (x_i/\rho)^{2r(p+1)} \quad (i = 1, 2, \dots, n, \quad p \in \mathbb{Z}_+), \\ \psi^p(\mathbf{x}) &= (\psi_1^p(\mathbf{x}), \psi_2^p(\mathbf{x}), \dots, \psi_n^p(\mathbf{x})) \quad (p \in \mathbb{Z}_+). \end{aligned} \right\}$$

Here $r = \max\{r_j : j = 0, 1, 2, \dots, m\}$ denotes the maximum of the degree r_j of the polynomial $f_j(\mathbf{x})$ ($j = 0, 1, 2, \dots, m$). Obviously $(\varphi^p, \psi^p) \in \bar{\Sigma}^{m+n}$ ($p \in \mathbb{Z}_+$).

THEOREM 3.1. *Assume that $\Xi \times \Sigma \subset \bar{\Sigma}^{m+n}$ contains an infinite subsequence of $\{(\varphi^p, \psi^p) \mid p \in \mathbb{Z}_+\}$. Then $L_C^*(\Sigma \times \Xi) = \zeta^*$ if $F \subset C_\rho$, and $L_C^*(\Sigma \times \Xi) = L_B^*(\Sigma) = \zeta^*$ if $F \subset B_\rho$.*

To prove the theorem, we use the following lemma, whose proof is given in section 5.

LEMMA 3.2. *Suppose that $F \subset C_\rho$. Let p^* be the smallest nonnegative integer such that $(m+2)n - 2^{r(p^*+1)} \leq 0$.*

(a) *If $\sqrt{2}\rho \leq \|\mathbf{x}\|_\infty$, then*

$$L_C(\mathbf{x}, \varphi^p, \psi^p) \equiv f_0(\mathbf{x}) + \Phi_C(\mathbf{x}, \varphi^p, \psi^p) \geq \|\mathbf{x}/\rho\|_\infty^{2rp} \geq \|\mathbf{x}/\rho\|_\infty^2$$

for every $p \geq p^*$.

(b) *If $\tilde{\mathbf{x}} \notin F$ and $\kappa \in \mathbb{R}$, then there exist $\delta > 0$ and $\tilde{p} \geq p^*$ such that*

$$L_C(\mathbf{x}, \varphi^p, \psi^p) \equiv f_0(\mathbf{x}) + \Phi_C(\mathbf{x}, \varphi^p, \psi^p) \geq \kappa$$

for every $p \geq \tilde{p}$ and $\mathbf{x} \in U_\delta(\tilde{\mathbf{x}})$. Here $U_\delta(\tilde{\mathbf{x}}) = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \tilde{\mathbf{x}}\| < \delta\}$.

(c) *If $\hat{\mathbf{x}} \in F$ and $\epsilon > 0$, then there exist $\delta > 0$ and $\tilde{p} \geq p^*$ such that*

$$-\epsilon \leq \Phi_C(\mathbf{x}, \varphi^p, \psi^p)$$

for every $\mathbf{x} \in U_\delta(\hat{\mathbf{x}})$ and $p \geq \tilde{p}$.

(d) $\lim_{p \rightarrow \infty} L_C^*(\varphi^p, \psi^p) = \zeta^*$.

Proof of Theorem 3.1. Note that $B_\rho \subset C_\rho$. Hence, by (8), we need only show that if $F \subset C_\rho$, then $L_C^*(\Sigma \times \Xi) = \zeta^*$. Suppose that $F \subset C_\rho$. Since the sequence $\{(\varphi^p, \psi^p) \mid (p \in \mathbb{Z}_+)\}$ lies in the set $\Sigma \times \Xi$ by the assumption, we see that

$$\limsup_{p \rightarrow \infty} L_C^*(\varphi^p, \psi^p) \leq L_C^*(\Sigma \times \Xi) \leq \zeta^*.$$

By (d) of Lemma 3.2,

$$\limsup_{p \rightarrow \infty} L_C^*(\varphi^p, \psi^p) \geq \lim_{p \rightarrow \infty} L_C^*(\varphi^p, \psi^p) = \zeta^*.$$

Therefore $L_C^*(\Sigma \times \Xi) = \zeta^*$. $\square$

3.4. Construction of $\Sigma \times \Xi$ satisfying the assumption of Theorem 3.1.

For every nonempty subset $\mathcal{A}$ of $\mathbb{Z}_+^n$, we define

$$\Sigma(\mathcal{A}) = \left\{ \sum_{i=1}^k \chi_i(\mathbf{x})^2 : \begin{array}{l} \chi_i \text{ is a polynomial in } \mathbf{x} \in \mathbb{R}^n \text{ with a support in } \mathcal{A} \\ (i = 1, 2, \dots, k) \text{ and } k \text{ is any finite positive integer.} \end{array} \right\}$$

Suppose that $\mathcal{A}_j^q$ ($j = 1, 2, \dots, m, q \in \mathbb{Z}_+$) and $\mathcal{B}_i^q$ ($i = 1, 2, \dots, n, q \in \mathbb{Z}_+$) are nonempty finite subsets of $\mathbb{Z}_+^n$ such that

$$(12) \quad \varphi_j^{\lambda(q)} \in \Sigma(\mathcal{A}_j^q) \text{ and } \psi_i^{\lambda(q)}(\mathbf{x}) \in \Sigma(\mathcal{B}_i^q) \text{ if } q \geq q^*$$

for some $q^* \in \mathbb{Z}_+$ and some mapping λ from $\mathbb{Z}_+$ into itself satisfying

$$\lambda(q) \leq \lambda(q+1) \quad (q \in \mathbb{Z}_+) \quad \text{and} \quad \lim_{q \rightarrow \infty} \lambda(q) = \infty.$$

Let

$$\begin{aligned} \Sigma^q &= \prod_{j=1}^m \Sigma(\mathcal{A}_j^q) \quad (q \in \mathbb{Z}_+), \quad \Xi^q = \prod_{i=1}^n \Sigma(\mathcal{B}_i^q) \quad (q \in \mathbb{Z}_+), \\ \Sigma^\infty &= \bigcup_{q \in \mathbb{Z}_+} \Sigma^q, \quad \text{and} \quad \Xi^\infty = \bigcup_{q \in \mathbb{Z}_+} \Xi^q. \end{aligned}$$

By construction, we see that

$$(\varphi^{\lambda(q)}, \psi^{\lambda(q)}) \in \Sigma^q \times \Xi^q \quad (q \geq q^*).$$

Hence $\Sigma^\infty \times \Xi^\infty$ contains an infinite subsequence $\{(\varphi^{\lambda(q)}, \psi^{\lambda(q)}) \mid (q \geq q^*)\}$. Therefore $\Sigma \times \Xi = \Sigma^\infty \times \Xi^\infty$ satisfies the assumption of Theorem 3.1.

We give some examples of $\mathcal{A}_j^q$ ($j = 1, 2, \dots, m, q \in \mathbb{Z}_+$) and $\mathcal{B}_i^q$ ($i = 1, 2, \dots, n, q \in \mathbb{Z}_+$) satisfying the assumption (12).

Example 3.3. For every $j = 1, 2, \dots, m, i = 1, 2, \dots, n, q \in \mathbb{Z}_+$, let

$$\begin{aligned} \mathcal{A}_j^0 &= \{\mathbf{0}\}, \quad \mathcal{A}_j^1 = \{\mathbf{0}\} \bigcup \mathcal{F}_j, \quad \mathcal{A}_j^{q+1} = \{\mathbf{a} + \mathbf{b} : \mathbf{a} \in \mathcal{A}_j^q, \mathbf{b} \in \mathcal{A}_j^1\} \quad (q \geq 1), \\ \mathcal{B}_i^q &= \{k\mathbf{e}^i : k = 0, 1, 2, \dots, (q+1)r\}. \end{aligned}$$

Example 3.4. For every $j = 1, 2, \dots, m, i = 1, 2, \dots, n, q \in \mathbb{Z}_+$, let

$$\begin{aligned} \mathcal{A}_j^0 &= \{\mathbf{0}\}, \quad \mathcal{A}_j^1 = \{\mathbf{0}\} \bigcup \{\mathbf{e}^k : k \in I_+(\mathcal{F}_j)\}, \\ \mathcal{A}_j^{q+1} &= \{\mathbf{a} + \mathbf{b} : \mathbf{a} \in \mathcal{A}_j^q, \mathbf{b} \in \mathcal{A}_j^1\} \quad (q \geq 1), \\ \mathcal{B}_i^q &= \{k\mathbf{e}^i : k = 0, 1, 2, \dots, q+1\} \quad (q \in \mathbb{Z}_+). \end{aligned}$$

Here $I_+(\mathcal{F}_j) = \{i : a_i > 0 \text{ for some } \mathbf{a} \in \mathcal{F}_j\}$.

Example 3.5. For every $j = 1, 2, \dots, m$, $i = 1, 2, \dots, n$, $q \in \mathbb{Z}_+$, let

$$\begin{aligned} \mathcal{A}_j^0 &= \mathcal{B}_i^0 = \{\mathbf{0}\}, \quad \mathcal{A}_j^1 = \mathcal{B}_i^1 = \{\mathbf{0}\} \bigcup \{\mathbf{e}^k : k = 1, 2, \dots, n\}, \\ \mathcal{A}_j^{q+1} &= \mathcal{B}_i^{q+1} = \{\mathbf{a} + \mathbf{b} : \mathbf{a} \in \mathcal{A}_j^q, \mathbf{b} \in \mathcal{A}_j^1\} \quad (q \geq 1). \end{aligned}$$

In all the examples, both $\mathcal{A}_j^q$ and $\mathcal{B}_i^q$ expand monotonically as q increases, and for any $\bar{p} \in \mathbb{Z}_+$ there exists $\bar{q} \in \mathbb{Z}_+$ such that $\varphi_j^p(\mathbf{x}) \in \Sigma(\mathcal{A}_j^q)$ and $\psi_i^p(\mathbf{x}) \in \Sigma(\mathcal{B}_i^q)$ for all $p \leq \bar{p}$ and $q \geq \bar{q}$. It should be noted that if $\mathcal{F}_j$ is sparse, then $\mathcal{A}_j^q$ remains sparse in Example 3.3. The choice of $\mathcal{A}_j^q$ in Example 3.4 may be also reasonable when the number of the indices $I_+(\mathcal{F}_j)$ is smaller than n . When we take $\mathcal{A}_j^q$ and $\mathcal{B}_i^q$ as in Example 3.5, we have

$$\begin{aligned} \Sigma(\mathcal{A}_j^q) &= \Sigma(\mathcal{B}_i^q) \\ &= \left\{ \sum_{i=1}^k \chi_i(\mathbf{x})^2 : \begin{array}{l} \chi_i \text{ is a polynomial in } \mathbf{x} \in \mathbb{R}^n \text{ with degree } \leq q \\ (i = 1, 2, \dots, k) \text{ and } k \text{ is any finite positive integer.} \end{array} \right\} \end{aligned}$$

Hence the sparsity of the original POP (1) is destroyed in the Lagrangian duals (6) and (7); the Lagrangian duals involve at least all the monomials with degree $\leq q$, many of which are not contained in the original POP (1) when the POP is sparse. We remark here that the choice of $\mathcal{A}_j^q$ as in Example 3.5 for the Lagrangian dual (7) leads to the dual of Lasserre's SDP relaxation [10] applied to the POP (3). This is discussed briefly in section 4.5.

4. Numerical methods for generalized Lagrangian duals. In the following five subsections, we discuss numerical methods for generalized Lagrangian duals. The first three subsections include how $L_C^*(\Sigma^\infty \times \Xi^\infty)$ can be approximated numerically, and the fourth section is for the approximation of $L_B^*(\Sigma^\infty)$. We briefly mention the relationship between the proposed methods and Lasserre's SDP relaxation [10] applied to the POP (2) in the last subsection. Throughout this section, we assume that $F \subset C_\rho$ so that $L_C^*(\Sigma^\infty \times \Xi^\infty) = \zeta^*$.

4.1. Approximation of generalized Lagrangian duals. We introduce a sequence of subproblems of the Lagrangian dual (6):

$$(13) \quad \text{Maximize } L_C^*(\varphi, \psi) \quad \text{subject to } (\varphi, \psi) \in \Sigma^q \times \Xi^q$$

($q \in \mathbb{Z}_+$).

LEMMA 4.1.

- (a) $L_C^*(\Sigma^q \times \Xi^q) \leq L_C^*(\Sigma^\infty \times \Xi^\infty)$ ($q \in \mathbb{Z}_+$).
- (b) For any $\epsilon > 0$, there exists a nonnegative integer $\bar{q}$ such that

$$L_C^*(\Sigma^\infty \times \Xi^\infty) - \epsilon \leq L_C^*(\Sigma^q \times \Xi^q) \quad (q \geq \bar{q}).$$

Proof. By construction, we know that $\Sigma^q \times \Xi^q \subset \Sigma^\infty \times \Xi^\infty$ ($q \in \mathbb{Z}_+$). Thus, (a) follows. Let $\epsilon > 0$. By (d) of Lemma 3.2, there exists $\bar{p}$ such that

$$L_C^*(\Sigma^\infty \times \Xi^\infty) - \epsilon \leq L_C^*(\varphi^p, \psi^p) \quad (p \geq \bar{p}).$$

Take $\bar{q} \in \mathbb{Z}_+$ such that $\lambda(q) \geq \bar{p}$ ($q \geq \bar{q}$). Then

$$(\varphi^{\lambda(q)}, \psi^{\lambda(q)}) \in \Sigma^q \times \Xi^q \text{ and } \lambda(q) \geq \bar{p} \text{ for every } q \geq \bar{q}.$$

Therefore we obtain that

$$L_C^*(\Sigma^\infty \times \Xi^\infty) - \epsilon \leq L_C^*(\varphi^{\lambda(q)}, \psi^{\lambda(q)}) \leq L_C^*(\Sigma^q \times \Xi^q) \text{ for every } q \geq \bar{q}. \quad \square$$

4.2. SOS relaxations. Let $q \in \mathbb{Z}_+$ be fixed throughout this subsection. We can rewrite the problems in (13) as

$$(14) \quad \begin{array}{ll} \text{maximize} & \zeta \\ \text{subject to} & \left. \begin{array}{l} L_C(\mathbf{x}, \boldsymbol{\varphi}, \boldsymbol{\psi}) - \zeta \geq 0 \ (\forall \mathbf{x} \in \mathbb{R}^n), \\ (\boldsymbol{\varphi}, \boldsymbol{\psi}) \in \Sigma^q \times \Xi^q. \end{array} \right\} \end{array}$$

Note that $\mathbf{x} \in \mathbb{R}^n$ is not a vector variable but serves as an index vector for an infinite number of inequality constraints $L_C(\mathbf{x}, \boldsymbol{\varphi}, \boldsymbol{\psi}) - \zeta \geq 0 \ (\forall \mathbf{x} \in \mathbb{R}^n)$. Replacing the inequality constraints $L_C(\mathbf{x}, \boldsymbol{\varphi}, \boldsymbol{\psi}) - \zeta \geq 0 \ (\forall \mathbf{x} \in \mathbb{R}^n)$ by an SOS condition $L_C(\mathbf{x}, \boldsymbol{\varphi}, \boldsymbol{\psi}) - \zeta \in \bar{\Sigma}$ in (14), we obtain an SOS optimization problem (SOSOP)

$$(15) \quad \begin{array}{ll} \text{maximize} & \zeta \\ \text{subject to} & \left. \begin{array}{l} L_C(\mathbf{x}, \boldsymbol{\varphi}, \boldsymbol{\psi}) - \zeta = \varphi_0(\mathbf{x}) \ (\forall \mathbf{x} \in \mathbb{R}^n), \\ (\boldsymbol{\varphi}, \boldsymbol{\psi}) \in \Sigma^q \times \Xi^q, \ \varphi_0(\mathbf{x}) \in \bar{\Sigma}. \end{array} \right\} \end{array}$$

Let ζ_C^q denote the optimal value of the SOSOP (15),

$$\zeta_C^q = \sup \left\{ \zeta : \begin{array}{l} L_C(\mathbf{x}, \boldsymbol{\varphi}, \boldsymbol{\psi}) - \zeta = \varphi_0(\mathbf{x}) \ (\forall \mathbf{x} \in \mathbb{R}^n), \\ (\boldsymbol{\varphi}, \boldsymbol{\psi}) \in \Sigma^q \times \Xi^q, \ \varphi_0(\mathbf{x}) \in \bar{\Sigma}. \end{array} \right\}$$

If $(\zeta, \boldsymbol{\varphi}, \boldsymbol{\psi}, \varphi_0)$ is a feasible solution of the SOSOP (15), then $(\zeta, \boldsymbol{\varphi}, \boldsymbol{\psi})$ is a feasible solution of the problem (14). It follows that $\zeta_C^q \leq L_C^*(\Sigma^q \times \Xi^q)$. Although neither $\zeta_C^q = L_C^*(\Sigma^q \times \Xi^q)$ nor the convergence of ζ_C^q to $L_C^*(\Sigma^\infty \times \Xi^\infty)$ as $q \rightarrow \infty$ is guaranteed, we can solve the SOSOP (15), as we show in the next subsection, while the problem (14) is difficult to solve in general.

4.3. Reduction to semidefinite programs. Let us fix $q \in \mathbb{Z}_+$ throughout this subsection. We show how to solve the SOSOP (15) as a semidefinite program. If we rewrite the constraint $(\boldsymbol{\varphi}, \boldsymbol{\psi}) \in \Sigma^q \times \Xi^q$ for each component, we have

$$\varphi_j(\mathbf{x}) \in \Sigma(\mathcal{A}_j^q) \ (j = 1, 2, \dots, m) \text{ and } \psi_i(\mathbf{x}) \in \Sigma(\mathcal{B}_i^q) \ (i = 1, 2, \dots, n).$$

Notice that finite supports $\mathcal{A}_j^q$ and $\mathcal{B}_i^q$ are given for generating variable polynomials $\varphi_j(\mathbf{x})$ and $\psi_i(\mathbf{x})$ ($j = 1, 2, \dots, m$, $i = 1, 2, \dots, n$). But, no finite support is specified for the variable polynomial $\varphi_0(\mathbf{x})$. The first step for constructing a semidefinite program is to find an appropriate finite set $\mathcal{G} \subset \mathbb{Z}_+^n$ so that $\varphi_0(\mathbf{x})$ can be chosen from $\Sigma(\mathcal{G})$.

To choose such a $\mathcal{G} \subset \mathbb{Z}_+^n$, we focus on the support of the left-hand side polynomial $L_C(\mathbf{x}, \boldsymbol{\varphi}, \boldsymbol{\psi}) - \zeta$ of the equality constraint in the SOSOP (15). From the support $\mathcal{F}_0$ of the objective polynomial function $f_0(\mathbf{x})$, the support of each term $\varphi_j(\mathbf{x})f_j(\mathbf{x})$,

$$\widehat{\mathcal{A}}_j^q \equiv \{\mathbf{a} + \mathbf{b} + \mathbf{c} : \mathbf{a} \in \mathcal{A}_j^q, \ \mathbf{b} \in \mathcal{A}_j^q, \ \mathbf{c} \in \mathcal{F}_j\} \quad (j = 1, 2, \dots, m),$$

and the support of term $\psi_i(\mathbf{x})(\rho^2 - x_i^2)$,

$$\widehat{\mathcal{B}}_i^q \equiv \{\mathbf{a} + \mathbf{b} + \mathbf{c} : \mathbf{a} \in \mathcal{B}_i^q, \ \mathbf{b} \in \mathcal{B}_i^q, \ \mathbf{c} \in \{0, 2\mathbf{e}^i\}\} \quad (i = 1, 2, \dots, n),$$

we know that the support of $L_C(\mathbf{x}, \boldsymbol{\varphi}, \boldsymbol{\psi}) - \zeta$ becomes

$$\mathcal{F}_L = \mathcal{F}_0 \cup \{\mathbf{0}\} \cup \left(\bigcup_{j=1}^m \widehat{\mathcal{A}}_j^q \right) \cup \left(\bigcup_{i=1}^n \widehat{\mathcal{B}}_i^q \right).$$

Here $\{\mathbf{0}\}$ stands for the support of the term ζ .

By Theorem 1 of [17], we can use

$$\mathcal{G}^0 = \left(\text{the convex hull of } \left\{ \mathbf{a}/2 : \begin{array}{l} \mathbf{a} \in \mathcal{F}_L, \\ \text{every } a_i \text{ is even } (i = 1, 2, \dots, n) \end{array} \right\} \right) \cap \mathbb{Z}_+^n$$

for such a support $\mathcal{G}$ that $\varphi_0(\mathbf{x})$ can be chosen from $\Sigma(\mathcal{G})$. We can further apply a method proposed recently in [7] for reducing the size of $\mathcal{G}^0$ to obtain the smallest support $\mathcal{G}^*$ in a class of supports including $\mathcal{G}^0$. See the paper [7] for more details.

Remark 4.2. Even when all $\mathcal{F}_j$ ($j = 0, 1, 2, \dots, m$) are sparse, $\mathcal{G}^*$ becomes dense. This is because the method proposed in [7] does not eliminate any integer point in the simplex with vertices $\bar{k}\mathbf{e}^i$ ($i = 1, 2, \dots, n$) and $\mathbf{0}$ contained in

$$\text{the convex hull of } \{ \mathbf{a}/2 : \mathbf{a} \in \mathcal{F}_L, \text{ every } a_i \text{ is even } (i = 1, 2, \dots, n), \}$$

which has induced $\mathcal{G}^0$ above. Here $\bar{k} = \min_i \max\{k : k\mathbf{e}^i \in \mathcal{B}_i^q\}$. Hence the support $\mathcal{G}^*$ does not benefit much from the sparsity of $\mathcal{F}_j$ ($j = 0, 1, 2, \dots, m$).

To transform the SOSOP (15) into a semidefinite program, we need some notation and symbols. Let $\mathcal{F} \in \mathbb{Z}_+^n$ be a nonempty finite set. Let $|\mathcal{F}|$ denote the cardinality of $\mathcal{F}$ and $\mathbb{R}(\mathcal{F})$ the $|\mathcal{F}|$ -dimensional Euclidean space whose coordinates are indexed by $\mathbf{a} \in \mathcal{F}$. Although the order of the coordinates is not relevant in the succeeding discussions, we may assume that the coordinates are arranged according to the lexicographical order. Each element of $\mathbb{R}(\mathcal{F})$ is denoted as $\mathbf{v} = (v_{\mathbf{a}} : \mathbf{a} \in \mathcal{F})$. We use the symbol $\mathcal{S}(\mathcal{F})_+$ for the set of $|\mathcal{F}| \times |\mathcal{F}|$ symmetric positive semidefinite matrices with coordinates $\mathbf{a} \in \mathcal{F}$; each $\mathbf{V} \in \mathcal{S}(\mathcal{F})_+$ has elements $V_{\mathbf{ab}}$ ($\mathbf{a} \in \mathcal{F}, \mathbf{b} \in \mathcal{F}$) such that $V_{\mathbf{ab}} = V_{\mathbf{ba}}$ and that $\mathbf{w}^T \mathbf{V} \mathbf{w} = \sum_{\mathbf{a} \in \mathcal{F}} \sum_{\mathbf{b} \in \mathcal{F}} V_{\mathbf{ab}} w_{\mathbf{a}} w_{\mathbf{b}} \geq 0$ for every $\mathbf{w} = (w_{\mathbf{a}} : \mathbf{a} \in \mathcal{F}) \in \mathbb{R}(\mathcal{F})$. For every $\mathbf{x} \in \mathbb{R}^n$, let $\mathbf{u}(\mathbf{x}, \mathcal{F}) = (\mathbf{x}^{\mathbf{a}} : \mathbf{a} \in \mathcal{F})$ be a column vector consisting of elements $\mathbf{x}^{\mathbf{a}}$ ($\mathbf{a} \in \mathcal{F}$).

The lemma below is well known [2, 7, 13, 15].

LEMMA 4.3. *Let $\mathcal{F}$ be a nonempty finite subset of $\mathbb{Z}_+^n$. A polynomial $\varphi(\mathbf{x})$ is contained in $\Sigma(\mathcal{F})$ if and only if there exists a $\mathbf{V} \in \mathcal{S}(\mathcal{F})_+$ such that*

$$(16) \quad \varphi(\mathbf{x}) = \mathbf{u}(\mathbf{x}, \mathcal{F})^T \mathbf{V} \mathbf{u}(\mathbf{x}, \mathcal{F}) = \sum_{\mathbf{a} \in \mathcal{F}} \sum_{\mathbf{b} \in \mathcal{F}} V_{\mathbf{ab}} \mathbf{x}^{\mathbf{a}+\mathbf{b}}.$$

Applying Lemma 4.3 to the polynomials $\varphi_j(\mathbf{x}) \in \Sigma(\mathcal{A}_j^{(\lambda_j)})$ ($j = 1, 2, \dots, m$), $\psi_i(\mathbf{x}) \in \Sigma(\mathcal{B}_i^{(\mu_i)})$ ($i = 1, 2, \dots, n$), and $\varphi_0(\mathbf{x}) \in \Sigma(\mathcal{G}^*)$, we represent these polynomials as follows:

$$\begin{aligned} \varphi_j(\mathbf{x}) &= \mathbf{u}(\mathbf{x}, \mathcal{A}_j^q)^T \mathbf{V}^j \mathbf{u}(\mathbf{x}, \mathcal{A}_j^q), \quad \mathbf{V}^j \in \mathcal{S}(\mathcal{A}_j^q)_+, \\ \psi_i(\mathbf{x}) &= \mathbf{u}(\mathbf{x}, \mathcal{B}_i^q)^T \mathbf{V}^{m+i} \mathbf{u}(\mathbf{x}, \mathcal{B}_i^q), \quad \mathbf{V}^{m+i} \in \mathcal{S}(\mathcal{B}_i^q)_+, \\ \varphi_0(\mathbf{x}) &= \mathbf{u}(\mathbf{x}, \mathcal{G}^*)^T \mathbf{V}^0 \mathbf{u}(\mathbf{x}, \mathcal{G}^*), \quad \mathbf{V}^0 \in \mathcal{S}(\mathcal{G}^*)_+. \end{aligned}$$

Substituting these functions in the SOSOP (15) leads to

$$\left. \begin{array}{ll} \text{maximize} & \zeta \\ \text{subject to} & \begin{aligned} & f_0(\mathbf{x}) - \sum_{j=1}^m f_j(\mathbf{x}) \mathbf{u}(\mathbf{x}, \mathcal{A}_j^q)^T \mathbf{V}^j \mathbf{u}(\mathbf{x}, \mathcal{A}_j^q) \\ & - \sum_{i=1}^n (\rho - x_i^2) \mathbf{u}(\mathbf{x}, \mathcal{B}_i^q)^T \mathbf{V}^{m+i} \mathbf{u}(\mathbf{x}, \mathcal{B}_i^q) \\ & - \mathbf{u}(\mathbf{x}, \mathcal{G}^*)^T \mathbf{V}^0 \mathbf{u}(\mathbf{x}, \mathcal{G}^*) - \zeta = 0 \quad (\forall \mathbf{x} \in \mathbb{R}^n), \\ & \mathbf{V}^j \in \mathcal{S}(\mathcal{A}_j^q)_+ \quad (j = 1, 2, \dots, m), \\ & \mathbf{V}^{m+i} \in \mathcal{S}(\mathcal{B}_i^q)_+ \quad (i = 1, 2, \dots, n), \\ & \mathbf{V}^0 \in \mathcal{S}(\mathcal{G}^*)_+. \end{aligned} \end{array} \right\}$$

Since the left-hand side of the equality constraint in the problem above is a polynomial with the support

$$\mathcal{F}_C = \mathcal{F}_0 \cup \{0\} \cup \left(\bigcup_{j=1}^m \hat{\mathcal{A}}_j^q \right) \cup \left(\bigcup_{i=1}^n \hat{\mathcal{B}}_i^q \right) \cup \{\mathbf{a} + \mathbf{b} : \mathbf{a} \in \mathcal{G}^*, \mathbf{b} \in \mathcal{G}^*\},$$

and the coefficients are linear functions of the matrix variables

$$\mathbf{V}^j \quad (j = 1, 2, \dots, m), \quad \mathbf{V}^{m+i} \quad (i = 1, 2, \dots, n), \quad \mathbf{V}^0$$

and the real variable ζ , we can rewrite the equality constraint of the problem above as

$$\sum_{\mathbf{a} \in \mathcal{F}_C} d(\mathbf{a}, \mathbf{V}, \zeta) \mathbf{x}^{\mathbf{a}} = 0,$$

where $d(\mathbf{a}, \mathbf{V}, \zeta)$ is a linear function in the matrix variables $\mathbf{V}^j$ ($j = 0, 1, 2, \dots, m+n$) and a real variable ζ for each $\mathbf{a} \in \mathcal{F}_C$. This identity needs to be satisfied for all $\mathbf{x} \in \mathbb{R}^n$ in the problem, and the equality constraint is equivalent to a system of linear equations

$$d(\mathbf{a}, \mathbf{V}, \zeta) = 0 \quad (\mathbf{a} \in \mathcal{F}_C).$$

Consequently, we obtain the following semidefinite program, which is equivalent to the SOSOP (15):

$$(17) \quad \left. \begin{array}{ll} \text{Maximize} & \zeta \\ \text{subject to} & \begin{aligned} & d(\mathbf{a}, \mathbf{V}, \zeta) = 0 \quad (\mathbf{a} \in \mathcal{F}_C), \\ & \mathbf{V}^j \in \mathcal{S}(\mathcal{A}_j^q)_+ \quad (j = 1, 2, \dots, m), \\ & \mathbf{V}^{m+i} \in \mathcal{S}(\mathcal{B}_i^q)_+ \quad (i = 1, 2, \dots, n), \\ & \mathbf{V}^0 \in \mathcal{S}(\mathcal{G}^*)_+. \end{aligned} \end{array} \right\}$$

The numerical efficiency of solving a semidefinite program depends largely on its size. In the semidefinite program (17) above, the number of equality constraints and the sizes of matrix variables are determined by the supports $\mathcal{F}_j$ of the polynomial functions $f_j(\mathbf{x})$ ($j = 0, 1, \dots, m$) in the original POP (1) to be solved and $q \in \mathbb{Z}_+$. When the supports are sparse, the size of the resulting semidefinite program becomes small. As we take a larger $q \in \mathbb{Z}_+$, we can expect to have a more accurate lower bound for the unknown optimal value ζ^* of the POP (1), but the number of equality constraints and the size of the matrix variables increase.

4.4. SOS relaxation of $L_B^*(\Sigma^\infty)$. In this section, we derive a sequence of SOSOPs from the Lagrangian dual (7). The SOSOPs obtained here are less sparse than (15) in the previous section, but their optimal values converge to $L_B^*(\Sigma^\infty)$. Thus, this compensates the theoretical weakness of the SOSOP (15), whose optimal value is not guaranteed to converge to $L_C^*(\Sigma^\infty \times \Xi^\infty)$. Throughout this section, we assume that $F \subset B_\rho$; hence $L_C^*(\Sigma^\infty \times \Xi^\infty) = L_B^*(\Sigma^\infty) = \zeta^*$ is satisfied by Theorem 3.1.

As we have obtained the sequence of SOSOPs (15) from the Lagrangian dual (6), we can similarly derive the following sequence of SOSOPs from the Lagrangian dual (7):

$$(18) \quad \begin{array}{ll} \text{Maximize} & \zeta \\ \text{subject to} & \left. \begin{array}{l} L_B(\mathbf{x}, \boldsymbol{\varphi}) - \zeta - \varphi_{m+1}(\mathbf{x})(\rho^2 - \mathbf{x}^T \mathbf{x}) = \varphi_0(\mathbf{x}) \ (\forall \mathbf{x} \in \mathbb{R}^n), \\ \boldsymbol{\varphi} \in \Sigma^q, \ \varphi_{m+1} \in \Sigma(\mathcal{G}(\tau(q) - 1)), \ \varphi_0 \in \bar{\Sigma}, \ (q \in \mathbb{Z}_+). \end{array} \right\} \end{array}$$

Here

$$\tau(q) = \lceil (\text{the degree of } L_B(\mathbf{x}, \boldsymbol{\varphi}) \text{ with } \boldsymbol{\varphi} \in \Sigma^q) / 2 \rceil,$$

$$\mathcal{G}(\xi) = \left\{ \mathbf{a} \in \mathbb{Z}_+^n : \sum_{i=1}^n a_i \leq \xi \right\} \quad (\xi = 0, 1, 2, \dots).$$

Let ζ_B^q denote the optimal value of the SOSOP (18),

$$\zeta_B^q = \sup \left\{ \zeta : \begin{array}{l} L_B(\mathbf{x}, \boldsymbol{\varphi}) - \zeta - \varphi_{m+1}(\mathbf{x})(\rho^2 - \mathbf{x}^T \mathbf{x}) = \varphi_0(\mathbf{x}), \\ \boldsymbol{\varphi} \in \Sigma^q, \ \varphi_{m+1} \in \Sigma(\mathcal{G}(\tau(q) - 1)), \ \varphi_0 \in \bar{\Sigma}. \end{array} \right\}$$

THEOREM 4.4.

(a) $\zeta_B^q \leq L_B^*(\Sigma^\infty)$ ($q \in \mathbb{Z}_+$).

(b) In addition to (12), assume that $\mathcal{A}_j^q \subset \mathcal{A}_j^{q+1}$ ($j = 1, 2, \dots, m, \ q \in \mathbb{Z}_+$). For any $\epsilon > 0$, there exists a $\hat{q} \in \mathbb{Z}_+$ such that $L_B^*(\Sigma^\infty) - \epsilon \leq \zeta_B^q$ ($q \geq \hat{q}$).

Proof.

(a) Let $q \in \mathbb{Z}_+$. Suppose that $(\zeta, \boldsymbol{\varphi}, \varphi_{m+1}, \varphi_0) = (\bar{\zeta}, \bar{\boldsymbol{\varphi}}, \bar{\varphi}_{m+1}, \bar{\varphi}_0)$ is a feasible solution of (18). Then $0 \leq L_B(\mathbf{x}, \boldsymbol{\varphi}) - \zeta$ for every $\mathbf{x} \in B_\rho$; hence $\bar{\zeta} \leq L_B^*(\boldsymbol{\varphi}) \leq L_B^*(\Sigma^\infty)$. This ensures that $\zeta_B^q \leq L_B^*(\Sigma^\infty)$.

(b) It follows from the assumption $\mathcal{A}_j^q \subset \mathcal{A}_j^{q+1}$ ($j = 1, 2, \dots, m, \ q \in \mathbb{Z}_+$) that $\Sigma^q \subset \Sigma^{q+1}$ ($q \in \mathbb{Z}$). Let $\epsilon > 0$. By the construction of $L_B^*(\Sigma^\infty)$, there exist $\hat{q} \in \mathbb{Z}_+$ and $\hat{\boldsymbol{\varphi}} \in \Sigma^{\hat{q}}$ such that

$$L_B^*(\Sigma^\infty) - \epsilon/2 \leq L_B^*(\hat{\boldsymbol{\varphi}}).$$

Then

$$L_B(\mathbf{x}, \hat{\boldsymbol{\varphi}}) - L_B^*(\Sigma^\infty) + \epsilon$$

is a polynomial that is positive in the ball B_ρ . By Lemma 4.1 of [16],

$$L_B(\mathbf{x}, \hat{\boldsymbol{\varphi}}) - L_B^*(\Sigma^\infty) + \epsilon - \hat{\varphi}_{m+1}(\mathbf{x})(\rho^2 - \mathbf{x}^T \mathbf{x}) \in \bar{\Sigma}$$

for some $\hat{\omega} \geq \tau(\hat{q})$ and some $\hat{\varphi}_{m+1} \in \Sigma(\mathcal{G}(\hat{\omega} - 1))$. Choose a $\tilde{q} \in \mathbb{Z}_+$ such that $\hat{q} \leq \tilde{q}$ and $\tau(\tilde{q}) \geq \hat{\omega}$. Let $\tilde{\omega} = \tau(\tilde{q})$. Then

$$\begin{aligned} \zeta_B^{\tilde{q}} &= \sup \left\{ \zeta : \begin{array}{l} L_B(\mathbf{x}, \boldsymbol{\varphi}) - \zeta - \varphi_{m+1}(\mathbf{x})(\rho^2 - \mathbf{x}^T \mathbf{x}) = \varphi_0(\mathbf{x}), \\ \boldsymbol{\varphi} \in \Sigma^{\tilde{q}}, \ \varphi_{m+1} \in \Sigma(\mathcal{G}(\tau(\tilde{q}) - 1)), \ \varphi_0 \in \bar{\Sigma} \end{array} \right\} \\ &\geq \sup \left\{ \zeta : L_B(\mathbf{x}, \hat{\boldsymbol{\varphi}}) - \zeta - \hat{\varphi}_{m+1}(\mathbf{x})(\rho^2 - \mathbf{x}^T \mathbf{x}) = \varphi_0(\mathbf{x}), \ \varphi_0 \in \bar{\Sigma} \right\} \\ &\quad (\text{since } \hat{\boldsymbol{\varphi}} \in \Sigma^{\hat{q}} \subset \Sigma^{\tilde{q}} \text{ and } \hat{\varphi}_{m+1} \in \Sigma(\mathcal{G}(\tau(\hat{q}) - 1)) \subset \Sigma(\mathcal{G}(\tau(\tilde{q}) - 1))) \\ &\geq L_B^*(\Sigma^\infty) - \epsilon. \quad \square \end{aligned}$$

Comparing the SOSOP (15) with the SOSOP (18) shows that the latter problem involves the term $\varphi_{m+1}(\mathbf{x})(\rho^2 - \mathbf{x}^T \mathbf{x})$ with $\varphi_{m+1} \in \Sigma(\mathcal{G}(\tau(q) - 1))$, which is a fully dense polynomial with degree $2\tau(q)$; hence we need to prepare a fully dense polynomial variable $\varphi_0 \in \Sigma(\mathcal{G}(\tau(q)))$ in general. If we convert the SOSOP (18) into a semidefinite program, as we converted the SOSOP (15) into the semidefinite program (17), the polynomial variables $\varphi_{m+1} \in \Sigma(\mathcal{G}(\tau(q) - 1))$ and $\varphi_0 \in \Sigma(\mathcal{G}(\tau(q)))$ induce matrix variables

$$\mathbf{V}^{m+1} \in \mathcal{S}(\mathcal{G}(\tau(q) - 1))_+ \text{ and } \mathbf{V}^0 \in \mathcal{S}(\mathcal{G}(\tau(q)))_+,$$

respectively. Let $\mathcal{F}_B$ denote the support of the polynomial

$$L_B(\mathbf{x}, \boldsymbol{\varphi}) - \zeta - \varphi_{m+1}(\mathbf{x})(\rho^2 - \mathbf{x}^T \mathbf{x}) - \varphi_0(\mathbf{x}) \text{ with } \\ \boldsymbol{\varphi} \in \overline{\Sigma}^q, \varphi_{m+1} \in \Sigma(\mathcal{G}(\tau(q) - 1)), \text{ and } \varphi_0 \in \Sigma(\mathcal{G}(\tau(q))).$$

Then the resulting SDP formulation of the SOSOP (18) turns out to be

$$(19) \quad \left. \begin{array}{ll} \text{maximize} & \zeta \\ \text{subject to} & d'(\mathbf{a}, \mathbf{V}, \zeta) = 0 \ (\mathbf{a} \in \mathcal{F}_B), \\ & \mathbf{V}^j \in \mathcal{S}(\mathcal{A}_j^q)_+ \ (j = 1, 2, \dots, m), \\ & \mathbf{V}^{m+1} \in \mathcal{S}(\mathcal{G}(\tau(q) - 1))_+, \\ & \mathbf{V}^0 \in \mathcal{S}(\mathcal{G}(\tau(q)))_+. \end{array} \right\}$$

Here $d'(\mathbf{a}, \mathbf{V}, \zeta)$ denotes a linear function in the matrix variables $\mathbf{V}^j$ ($j = 0, 1, 2, \dots, m+1$) and a real variable ζ for each $\mathbf{a} \in \mathcal{F}_B$. If the polynomials $f_j(\mathbf{x})$ ($j = 0, 1, 2, \dots, m$) are sparse, the size of (19) is larger than the size of (17); the number $|\mathcal{F}_B|$ of linear equations in (19) is larger than the number $|\mathcal{F}_C|$ of linear equations in (17), and the sizes of two matrix variables $\mathbf{V}^{m+1}$ and $\mathbf{V}^0$ in (19) are larger than the sizes of the matrix variables in (17).

4.5. Comparison to Lasserre's SDP relaxation. In this subsection, we briefly mention a modification of Lasserre's SDP relaxation [10] that takes into account the supports $\mathcal{F}_j$ of the polynomials $f_j(\mathbf{x})$ ($j = 0, 1, \dots, m$) in the POP (3). The modification leads to the dual semidefinite program of (19).

First, we describe the original Lasserre's SDP relaxation [10] applied to the POP (3) according to the interpretation given in the paper [6] on the relaxation. Let $\lceil r/2 \rceil \leq N \in \mathbb{Z}_+$ and $\nu_j = N - \lceil r_j/2 \rceil$ ($j = 1, 2, \dots, m$). Consider the following polynomial optimization problem on positive semidefinite cones, which is equivalent to the POP (1):

$$(20) \quad \left. \begin{array}{ll} \text{Minimize} & f_0(\mathbf{x}) \\ \text{subject to} & f_j(\mathbf{x})\mathbf{u}(\mathbf{x}, \mathcal{G}(\nu_j))\mathbf{u}(\mathbf{x}, \mathcal{G}(\nu_j))^T \in \mathcal{S}(\mathcal{G}(\nu_j))_+ \ (j = 1, 2, \dots, m), \\ & (\rho^2 - \mathbf{x}^T \mathbf{x})\mathbf{u}(\mathbf{x}, \mathcal{G}(N-1))\mathbf{u}(\mathbf{x}, \mathcal{G}(N-1))^T \in \mathcal{S}(\mathcal{G}(N-1))_+, \\ & \mathbf{u}(\mathbf{x}, \mathcal{G}(N))\mathbf{u}(\mathbf{x}, \mathcal{G}(N))^T \in \mathcal{S}(\mathcal{G}(N))_+. \end{array} \right\}$$

We then apply a linearization to this problem to obtain an SDP relaxation of the POP (1). See the paper [6].

When the modification for the primal approach corresponding to (15) is implemented, the POP (3) is converted into a polynomial optimization problem on positive semidefinite cones:

$$(21) \quad \left. \begin{array}{ll} \text{Minimize} & f_0(\mathbf{x}) \\ \text{subject to} & f_j(\mathbf{x})\mathbf{u}(\mathbf{x}, \mathcal{A}_j^q)\mathbf{u}(\mathbf{x}, \mathcal{A}_j^q)^T \in \mathcal{S}(\mathcal{A}_j^q)_+ \ (j = 1, 2, \dots, m), \\ & (\rho^2 - \mathbf{x}^T \mathbf{x})\mathbf{u}(\mathbf{x}, \mathcal{G}(\tau(q) - 1))\mathbf{u}(\mathbf{x}, \mathcal{G}(\tau(q) - 1))^T \in \mathcal{S}(\mathcal{G}(\tau(q) - 1))_+, \\ & \mathbf{u}(\mathbf{x}, \mathcal{G}(\tau(q)))\mathbf{u}(\mathbf{x}, \mathcal{G}(\tau(q)))^T \in \mathcal{S}(\mathcal{G}(\tau(q)))_+. \end{array} \right\}$$

We obtain a semidefinite program by applying the linearization to the problem (21). Then the resulting semidefinite program is the dual of (19) derived from the Lagrangian dual (7) of the POP (3). See section 6 of [6] for more details on the linearization technique and the proof of the fact above.

Notice that N in the problem (20) corresponds to $\tau(q)$ in the problem (21). Using Lemma 4.1 of [16], which we have used to prove the convergence of ζ_B^q to ζ^* , Lasserre proved the convergence of the optimal value of the semidefinite program derived from the problem (20) to the optimal value ζ^* of the POP (1) as the parameter N tends to ∞ in the paper [10]. An essential difference between the original and the modified Lasserre's SDP relaxations is that the modified relaxation takes into account the supports $\mathcal{F}_j$ of the polynomials $f_j(\mathbf{x})$ ($j = 1, \dots, m$) in the POP (3). As the supports $\mathcal{F}_j$ ($j = 0, 1, 2, \dots, m$) are more sparse, the modified relaxation results in a smaller SDP relaxation. In other words, (17) derived from the Lagrangian dual (6) utilizes the sparsity of the POP (1) more effectively and therefore, its size is smaller than (19) derived from the Lagrangian dual (7) of the POP (3).

5. Proof of Lemma 3.2.

(a) Suppose that $\sqrt{2}\rho \leq \|\mathbf{x}\|_\infty$. Let $p \geq p^*$. If $f_j(\mathbf{x}) < 0$, then

$$f_j(\mathbf{x})\varphi_j^p(\mathbf{x}) = f_j(\mathbf{x})(1 - f_j(\mathbf{x})/\gamma)^{2p} < 0.$$

Otherwise, we obtain that

$$\begin{aligned} 0 &\leq f_j(\mathbf{x})\varphi_j^p(\mathbf{x}) \\ &= f_j(\mathbf{x})(1 - f_j(\mathbf{x})/\gamma)^{2p} \\ &\leq \gamma \|\mathbf{x}/\rho\|_\infty^r \max \left\{ 1, (f_j(\mathbf{x})/\gamma)^{2p} \right\} \\ &\quad (\text{by (10) and } |1 - f_j(\mathbf{x})/\gamma| \leq \max\{1, |f_j(\mathbf{x})/\gamma|\}) \\ &\leq \gamma \|\mathbf{x}/\rho\|_\infty^r (\|\mathbf{x}/\rho\|_\infty^r)^{2p} \quad (\text{by (10)}) \\ &= \gamma \|\mathbf{x}/\rho\|_\infty^{r(2p+1)}. \end{aligned}$$

Hence

$$\sum_{j=1}^m f_j(\mathbf{x})\varphi_j^p(\mathbf{x}) \leq m\gamma \|\mathbf{x}/\rho\|_\infty^{r(2p+1)}.$$

On the other hand, we see that

$$\begin{aligned}
& \sum_{i=1}^n (\rho^2 - x_i^2) \psi_i^p(\mathbf{x}) \\
&= ((m+2)\gamma/\rho^2) \sum_{i=1}^n (\rho^2 - x_i^2) (x_i/\rho)^{2r(p+1)} \\
&= ((m+2)\gamma/\rho^2) \left(\sum_{x_i^2 \leq \rho^2} (\rho^2 - x_i^2) (x_i/\rho)^{2r(p+1)} + \sum_{\rho^2 < x_i^2} (\rho^2 - x_i^2) (x_i/\rho)^{2r(p+1)} \right) \\
&\leq (m+2)\gamma \left(n + \left(1 - \|\mathbf{x}/\rho\|_\infty^2\right) \|\mathbf{x}/\rho\|_\infty^{2r(p+1)} \right) \quad (\text{since } \{i \mid \rho^2 < x_i^2\} \neq \emptyset) \\
&\leq \gamma \left((m+2)n - (m+2) \|\mathbf{x}/\rho\|_\infty^{2r(p+1)} \right) \quad (\text{since } \sqrt{2} \leq \|\mathbf{x}/\rho\|_\infty) \\
&\leq \gamma \left((m+2)n - \|\mathbf{x}/\rho\|_\infty^{2r(p^*+1)} - (m+1) \|\mathbf{x}/\rho\|_\infty^{2r(p+1)} \right) \\
&\quad (\text{since } p^* \leq p \text{ and } \sqrt{2} \leq \|\mathbf{x}/\rho\|_\infty) \\
&\leq -(m+1)\gamma \|\mathbf{x}/\rho\|_\infty^{2r(p+1)} \\
&\quad (\text{by } \sqrt{2} \leq \|\mathbf{x}/\rho\|_\infty \text{ and } (m+2)n - 2^{r(p^*+1)} \leq 0).
\end{aligned}$$

Therefore

$$\begin{aligned}
& L_C(\mathbf{x}, \boldsymbol{\varphi}^p, \boldsymbol{\psi}^p) \\
&= f_0(\mathbf{x}) - \sum_{j=1}^m f_j(\mathbf{x}) \varphi_j^p(\mathbf{x}) - \sum_{i=1}^n (\rho^2 - x_i^2) \psi_i^p(x_i) \\
&\geq f_0(\mathbf{x}) - m\gamma \|\mathbf{x}/\rho\|_\infty^{r(2p+1)} + (m+1)\gamma \|\mathbf{x}/\rho\|_\infty^{2r(p+1)} \\
&\geq f_0(\mathbf{x}) + \gamma \|\mathbf{x}/\rho\|_\infty^{2r(p+1)} \quad (\text{since } -m \|\mathbf{x}/\rho\|_\infty^{r(2p+1)} + m \|\mathbf{x}/\rho\|_\infty^{2r(p+1)} \geq 0) \\
&\geq -\gamma \|\mathbf{x}/\rho\|_\infty^r + \gamma \|\mathbf{x}/\rho\|_\infty^{2r(p+1)} \quad (\text{by (10)}) \\
&= \gamma \|\mathbf{x}/\rho\|_\infty^{2r(p+1)} \left(1 - \|\mathbf{x}/\rho\|_\infty^{-r(2p+1)} \right) \\
&\geq \frac{1}{2} \|\mathbf{x}/\rho\|_\infty^{2r} \|\mathbf{x}/\rho\|_\infty^{2rp} \quad (\text{since } \gamma \geq 1, r \geq 1, p \geq 1, \text{ and } \|\mathbf{x}/\rho\|_\infty \geq \sqrt{2}) \\
&\geq \|\mathbf{x}/\rho\|_\infty^{2rp} \geq \|\mathbf{x}/\rho\|_\infty^2 \quad (\text{since } r \geq 1, p \geq 1, \text{ and } \|\mathbf{x}/\rho\|_\infty \geq \sqrt{2}).
\end{aligned}$$

Thus we have shown (a).

(b) Suppose that $\tilde{\mathbf{x}} \notin F$ and $\kappa > 0$. If $\|\tilde{\mathbf{x}}\|_\infty > \sqrt{2}\rho$, take $\delta > 0$ such that $\|\mathbf{x}\|_\infty > \sqrt{2}\rho$ for every $\mathbf{x} \in U_\delta(\tilde{\mathbf{x}})$. Then the desired result follows from (a). Now assume that $\|\tilde{\mathbf{x}}\|_\infty \leq \sqrt{2}\rho$. By (10), we see that

$$|f_j(\tilde{\mathbf{x}})/\gamma| \leq 1 \quad (j = 0, 1, 2, \dots, m).$$

Since $\tilde{\mathbf{x}} \notin F$, $f_k(\tilde{\mathbf{x}}) < 0$ also holds for some $k \in \{1, 2, \dots, m\}$. Hence we can take $\epsilon > 0$ and $\delta > 0$ such that

$$f_k(\mathbf{x})/\gamma < -\epsilon \quad \text{and} \quad |f_j(\mathbf{x})/\gamma| \leq 2 \quad (j = 0, 1, 2, \dots, m) \quad \text{for every } \mathbf{x} \in U_\delta(\tilde{\mathbf{x}}).$$

Let $\tilde{p} \geq p^*$ be a positive integer such that

$$\epsilon(1 + \epsilon)^{2\tilde{p}} - 2 - 2m - (m+2)n \geq \kappa/\gamma.$$

Then, for every $\mathbf{x} \in U_\delta(\tilde{\mathbf{x}})$ and $p \geq \tilde{p}$,

$$\begin{aligned}
L_C(\mathbf{x}, \boldsymbol{\varphi}^p, \boldsymbol{\psi}^p) &= f_0(\mathbf{x}) - \sum_{j=1}^m f_j(\mathbf{x}) \varphi_j^p(\mathbf{x}) - \sum_{i=1}^n (\rho^2 - x_i^2) \psi_i^p(\mathbf{x}) \\
&\geq -2\gamma - f_k(\mathbf{x}) \varphi_k^p(\mathbf{x}) - \sum_{f_j(\mathbf{x}) \geq 0} f_j(\mathbf{x}) \varphi_j^p(\mathbf{x}) - \sum_{\rho^2 - x_i^2 \geq 0} (\rho^2 - x_i^2) \psi_i^p(\mathbf{x}) \\
&= -2\gamma - f_k(\mathbf{x}) (1 - f_k(\mathbf{x})/\gamma)^{2p} - \sum_{f_j(\mathbf{x}) \geq 0} f_j(\mathbf{x}) (1 - f_j(\mathbf{x})/\gamma)^{2p} \\
&\quad - ((m+2)\gamma/\rho^2) \sum_{\rho^2 - x_i^2 \geq 0} (\rho^2 - x_i^2) (x_i/\rho)^{2r(p+1)} \\
&\geq -2\gamma + \gamma\epsilon(1+\epsilon)^{2p} - 2m\gamma - (m+2)n\gamma \\
&\geq \gamma(\epsilon(1+\epsilon)^{2p} - 2 - 2m - (m+2)n) \\
&\geq \kappa.
\end{aligned}$$

(c) Suppose that $\hat{\mathbf{x}} \in F$ and $\epsilon > 0$. Then

$$\|\hat{\mathbf{x}}\|_\infty \leq \rho \text{ and } 0 \leq f_j(\hat{\mathbf{x}})/\gamma \leq 1 \text{ (} j = 1, 2, \dots, m \text{)}.$$

We can take a $\delta > 0$ such that

$$f_j(\hat{\mathbf{x}})/\gamma \leq 1.1 \text{ (} j = 1, 2, \dots, m \text{) for every } \mathbf{x} \in U_\delta(\hat{\mathbf{x}}),$$

and $\tilde{p} \geq p^*$ such that

$$\begin{aligned}
0 &\leq \eta(1-\eta)^{2\tilde{p}} \leq \epsilon/(2m\gamma) \text{ if } 0 \leq \eta \leq 1.1, \\
0 &\leq (1-\xi)\xi^{r(\tilde{p}+1)} \leq \epsilon/((2m+4)n\gamma) \text{ if } 0 \leq \xi \leq 1.
\end{aligned}$$

Then, for every $\mathbf{x} \in U_\delta(\hat{\mathbf{x}})$ and $p \geq \tilde{p}$,

$$\begin{aligned}
\Phi_C(\mathbf{x}, \boldsymbol{\varphi}^p, \boldsymbol{\psi}^p) &= - \sum_{j=1}^m f_j(\mathbf{x}) \varphi_j^p(\mathbf{x}) - \sum_{i=1}^n (\rho^2 - x_i^2) \psi_i^p(\mathbf{x}) \\
&\geq - \sum_{f_j(\mathbf{x}) \geq 0} f_j(\mathbf{x}) (1 - f_j(\mathbf{x})/\gamma)^{2p} \\
&\quad - ((m+2)\gamma/\rho^2) \sum_{\rho^2 - x_i^2 \geq 0} (\rho^2 - x_i^2) (x_i/\rho)^{2r(p+1)} \\
&\geq - \sum_{f_j(\mathbf{x}) \geq 0} \gamma\epsilon/(2m\gamma) - (m+2)\gamma \sum_{\rho^2 - x_i^2 \geq 0} \epsilon/((2m+4)n\gamma) \\
&= -\epsilon \sum_{f_j(\mathbf{x}) \geq 0} 1/(2m) - \sum_{\rho^2 - x_i^2 \geq 0} \epsilon/(2n) \\
&\geq -\epsilon.
\end{aligned}$$

Thus we have shown (c).

(d) Define a compact subset A of $\mathbb{R}^n$ by

$$A = \left\{ \mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_\infty \leq \sqrt{2}\rho \right\} \cup \left\{ \mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}/\rho\|_\infty^2 \leq \zeta^* \right\}.$$

In view of property (a), we know that

$$\{\mathbf{x} \in \mathbb{R}^n : L_C(\mathbf{x}, \boldsymbol{\varphi}^p, \boldsymbol{\psi}^p) \leq \zeta^*\} \subset A \quad (p \geq p^*).$$

By property (4) with $(\boldsymbol{\varphi}, \boldsymbol{\psi}) = (\boldsymbol{\varphi}^p, \boldsymbol{\psi}^p)$, $\mathbf{x} = \mathbf{x}^* \in F$, and $\zeta^* = f_0(\mathbf{x}^*)$, we also see that

$$L_C(\mathbf{x}^*, \boldsymbol{\varphi}^p, \boldsymbol{\psi}^p) \leq f_0(\mathbf{x}^*) = \zeta^* \quad (p \geq p^*) \text{ and } \mathbf{x}^* \in F \subset A.$$

Hence

$$L_C^*(\boldsymbol{\varphi}^p, \boldsymbol{\psi}^p) = \inf \{L_C(\mathbf{x}, \boldsymbol{\varphi}^p, \boldsymbol{\psi}^p) : \mathbf{x} \in \mathbb{R}^n\} = \inf \{L_C(\mathbf{x}, \boldsymbol{\varphi}^p, \boldsymbol{\psi}^p) : \mathbf{x} \in A\} \quad (p \geq p^*).$$

Since A is compact, $L_C(\mathbf{x}, \boldsymbol{\varphi}^p, \boldsymbol{\psi}^p)$ has a minimizer over $\mathbb{R}^n$ at an $\mathbf{x} = \mathbf{x}^p \in A$:

$$L_C^*(\boldsymbol{\varphi}^p, \boldsymbol{\psi}^p) = L_C(\mathbf{x}^p, \boldsymbol{\varphi}^p, \boldsymbol{\psi}^p) \text{ and } \mathbf{x}^p \in A \quad (p \geq p^*).$$

Now assume on the contrary that the sequence $\{L_C(\mathbf{x}^p, \boldsymbol{\varphi}^p, \boldsymbol{\psi}^p) \mid p \geq p^*\}$ does not converge to ζ^* . Then we can take an $\epsilon > 0$ and a subsequence $\{L_C(\mathbf{x}^p, \boldsymbol{\varphi}^p, \boldsymbol{\psi}^p) \mid p \in J\}$ such that

$$(22) \quad L_C(\mathbf{x}^p, \boldsymbol{\varphi}^p, \boldsymbol{\psi}^p) < \zeta^* - \epsilon \quad (p \in J).$$

Since the sequence $\{\mathbf{x}^p \mid p \in J\}$ lies in the compact set A , we can take a subsequence $\{\mathbf{x}^p \mid p \in K\}$ with $K \subset J$ which converges to some $\bar{\mathbf{x}} \in A$. Since $L_C(\mathbf{x}^p, \boldsymbol{\varphi}^p, \boldsymbol{\psi}^p) \leq \zeta^* \quad (p \in K)$, we know by property (b) that $\bar{\mathbf{x}} \in F$. By property (c), there exists a $\tilde{p} \geq p^*$ such that

$$f_0(\mathbf{x}^p) - \epsilon/2 \leq L_C(\mathbf{x}^p, \boldsymbol{\varphi}^p, \boldsymbol{\psi}^p) \leq \zeta^* \text{ if } p \in K \text{ and } p \geq \tilde{p}.$$

Taking the limit along the subsequence $\{\mathbf{x}^p \mid p \in K\}$, the left-hand side $f_0(\mathbf{x}^p) - \epsilon/2$ of the inequality above converges to $f_0(\bar{\mathbf{x}}) - \epsilon/2 \geq \zeta^* - \epsilon/2$. Hence

$$\zeta^* - \epsilon \leq L_C(\mathbf{x}^p, \boldsymbol{\varphi}^p, \boldsymbol{\psi}^p) \leq \zeta^* \text{ for every sufficiently large } p \in K.$$

This contradicts assumption (22). Therefore, (d) follows.

6. Preliminary numerical results. We provide an illustrative example of structured and sparse POPs and show how the choice of SOS polynomials in SOS relaxations can enhance the efficiency of the proposed relaxations greatly while preserving the effectiveness.

As mentioned in Remark 4.2, the support set $\mathcal{G}^*$ in the proposed SOS relaxation of (2) becomes dense even for sparse $\mathcal{F}_j$ ($j = 0, 1, 2, \dots, m$) of the POP (2) because a polynomial $\varphi_0(\mathbf{x})$ is determined from the support of $L_C(\mathbf{x}, \boldsymbol{\varphi}) - \xi$. The convergence result shown in section 4 is based on this choice of φ_0 . In practical implementation of the proposed SOS relaxations, however, it may be more important to obtain a good lower bound with relatively small size SDP relaxations. We show that the formulation of SOS relaxation presented in this paper can be easily adapted in practice with the following example. The aim of this example is not to propose a practical method for general structured and sparse POPs, but to show how the SOS relaxation with convergent property can be modified for a specific problem in practice.

We consider an example:

$$(23) \quad \left. \begin{array}{ll} \text{Minimize} & f_0(\mathbf{x}) \equiv \sum_{i=1}^{n-1} f_{0i}(x_i, x_{i+1}) \\ \text{subject to} & f_{ij}(x_i, x_{i+1}) \geq 0 \quad (i = 1, 2, \dots, n-1, \quad j = 1, 2, \dots, m). \end{array} \right\}$$

Here $m \in \{1, n, n^2\}$, each $f_{0i}(x_i, x_{i+1})$ denotes a (fully dense) polynomial with degree 6 in two variables x_i and x_{i+1} whose coefficients are chosen randomly from the interval $(-1, 1)$ ($i = 1, 2, \dots, n-1$), and each $f_{ij}(x_i, x_{i+1})$ denotes a polynomial in two variables x_i and x_{i+1} of the form

$$1 - (x_i^\ell, x_{i+1}) \left(\frac{1}{\lambda_1^2} \begin{pmatrix} -a_2 \\ a_1 \end{pmatrix} (-a_2, a_1) + \frac{1}{\lambda_2^2} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} (a_1, a_2) \right) \begin{pmatrix} x_i^\ell \\ x_{i+1} \end{pmatrix}$$

for some $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$ chosen from the unit circle, λ_1, λ_2 chosen randomly from the interval $(0.5, 2)$, and $\ell \in \{1, 3\}$ ($i = 1, 2, \dots, n-1, j = 1, 2, \dots, m$). When $\ell = 1$, each constraint $f_{ij}(x_i, x_{i+1}) \geq 0$ forms an ellipsoid in the (x_i, x_{i+1}) space with the center at the origin; if $\lambda_1 > \lambda_2$, $\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$ corresponds to the major axis and $\begin{pmatrix} -a_2 \\ a_1 \end{pmatrix}$ to the minor axis.

Let us derive three relaxations of (23): the dual of Lasserre's SDP relaxation, the SOS relaxation presented in section 4.2, and a practical version of the SOS relaxation. If we want to have the POP (23) in the form of (2) and to follow literally the theory described so far, the redundant inequalities

$$1 - x_i^2 \geq 0 \quad (i = 1, 2, \dots, n)$$

need to be added to the POP (23). However, for simplicity of discussion, we consider the problem without these inequalities. Notice that if these inequalities are added, the three relaxations get stronger. As far as the size of the relaxations is concerned, adding the inequalities increases the size of all three relaxations. The biggest increase in the size occurs in the case of the dual of Lasserre's SDP relaxation given in (24). We also note that all the SOS relaxations presented below remain effective without the inequalities.

Define the Lagrangian function

$$L(\mathbf{x}, \boldsymbol{\varphi}) = f_0(\mathbf{x}) - \sum_{i=1}^{n-1} \sum_{j=1}^m \varphi_{ij}(\mathbf{x}) f_{ij}(x_i, x_{i+1}) \text{ for every } \mathbf{x} \in \mathbb{R}^n$$

$$\text{and every } \boldsymbol{\varphi} = (\varphi_{ij} \ (i = 1, 2, \dots, n-1, j = 1, 2, \dots, m)) \in \overline{\Sigma}^{m(n-1)}.$$

For every $i = 1, 2, \dots, n-1$ and $q = 0, 1, \dots$, let

$$\begin{aligned} \mathcal{A}_i^q &= \{ \mu \mathbf{e}^i + \nu \mathbf{e}^{i+1} : \mu \in \mathbb{Z}_+, \nu \in \mathbb{Z}_+, \mu + \nu \leq q \}, \\ \mathcal{A}_0^q &= \left\{ \mathbf{a} \in \mathbb{Z}_+^n : \sum_{i=1}^n a_i \leq q \right\}. \end{aligned}$$

Then we have two types of SOS relaxations. One is

$$(24) \quad \left. \begin{array}{ll} \text{maximize} & \zeta \\ \text{subject to} & L(\mathbf{x}, \boldsymbol{\varphi}) - \zeta = \varphi_0(\mathbf{x}) \ (\forall \mathbf{x} \in \mathbb{R}^n), \\ & \varphi_{ij} \in \Sigma(\mathcal{A}_0^q) \ (i = 1, 2, \dots, n-1, j = 1, 2, \dots, m), \\ & \varphi_0 \in \Sigma(\mathcal{A}_0^{q+\ell}), \end{array} \right\}$$

which corresponds to the dual of Lasserre's SDP relaxation applied to the POP (23).

The other is

$$(25) \quad \left. \begin{array}{ll} \text{maximize} & \zeta \\ \text{subject to} & L(\mathbf{x}, \boldsymbol{\varphi}) - \zeta = \varphi_0(\mathbf{x}) \ (\forall \mathbf{x} \in \mathbb{R}^n), \\ & \varphi_{ij} \in \Sigma(\mathcal{A}_i^q) \ (i = 1, 2, \dots, n-1, j = 1, 2, \dots, m), \\ & \varphi_0 \in \Sigma(\mathcal{A}_0^{q+\ell}), \end{array} \right\}$$

TABLE 6.1

Numerical results on the POP (23) with $m = 1$, $\ell = 0$, and $q = 2$.

POP (23)	cpu time in seconds		
n	relaxation (24)	relaxation (25)	relaxation (26)
4	0.6	0.4	0.1
5	4.9	2.2	0.1
6	22.3	21.5	0.1
7	153.8	98.6	0.2

TABLE 6.2

Numerical results on the POP (23) with $m = n$, $\ell = 0$, and $q = 2$.

POP (23)	cpu time in seconds		
n	relaxation (24)	relaxation (25)	relaxation (26)
4	1.6	0.5	0.1
5	11.2	2.6	0.2
6	83.3	13.2	0.4
7	607.1	64.4	0.5

TABLE 6.3

Numerical results on the POP (23) with $m = n^2$, $\ell = 0$, and $q = 2$.

POP (23)	cpu time in seconds		
n	relaxation (24)	relaxation (25)	relaxation (26)
4	6.6	1.0	0.5
5	83.7	5.1	1.1
6	717.1	23.2	2.7
7	7402.5	135.6	4.1

which exploits the sparsity of the constraint inequalities of (23). In both relaxations, we take nonnegative integers q and ℓ such that $q + \ell \geq 3$; hence $q = 2, 3, \dots$ if $\ell = 1$, and $q = 0, 1, 2, \dots$ if $\ell = 3$.

The SDP relaxation (24) uses

$$m(n-1) \text{ copies of support sets } \mathcal{A}_0^q \text{ of size } \#\mathcal{A}_0^q = \binom{n+q}{n},$$

$$\text{a support set } \mathcal{A}_0^{q+\ell} \text{ of size } \#\mathcal{A}_0^{q+\ell} = \binom{n+q+\ell}{n},$$

while the SDP relaxation (25) uses

$$m \text{ copies of support sets } \mathcal{A}_i^q \text{ of size } \#\mathcal{A}_i^q = \binom{2+q}{2} \text{ } (i = 1, 2, \dots, n-1),$$

$$\text{a support set } \mathcal{A}_0^{q+\ell} \text{ of size } \#\mathcal{A}_0^{q+\ell} = \binom{n+q+\ell}{n}.$$

The two SOS relaxations (24) and (25) share the support set $\mathcal{A}_0^{q+\ell}$ of size $\binom{n+q+\ell}{n}$. The difference between them lies in the support sets $\mathcal{A}_0^q$ and $\mathcal{A}_i^q$. We can see that the size of the SOS relaxation (25) is smaller than the size of the SOS relaxation (24). When q is fixed, the advantage of the SOS relaxation (25) in the size of the problem over the SOS relaxation (24) becomes larger as m increases. This will be shown in Tables 6.1, 6.2, and 6.3. In the case of m fixed, the common support set $\mathcal{A}_0^{q+\ell}$ dominates all other support sets in both SOS relaxations in terms of size. As

a result, the advantage of the size when increasing q is not as much as the case of q fixed.

Exploiting the structure of polynomials may improve the performance of the SOS relaxation (25) of the POP (23). We focus on “tridiagonal structure” of the support of the left-hand side polynomial $L(\mathbf{x}, \boldsymbol{\varphi}) - \zeta$ of the equality constraint of the SOS relaxation (25), where φ_{ij} is assumed to be chosen from $\Sigma(\mathcal{A}_i^q)$. Specifically, the support of the polynomial $L(\mathbf{x}, \boldsymbol{\varphi}) - \zeta$ is covered by $\bigcup_{i=1}^{n-1} (\mathcal{A}_i^{q+\ell} + \mathcal{A}_i^{q+\ell})$. Here we assume that $q + \ell \geq 3$. From this, we can expect that the polynomial $L(\mathbf{x}, \boldsymbol{\varphi}) - \zeta$ is represented as SOS of polynomials, each of which has a single support from $\mathcal{A}_i^{q+\ell}$ ($i = 1, 2, \dots, n-1$). We replace $\varphi_0(\mathbf{x})$ and $\varphi_0 \in \Sigma(\mathcal{A}_0^{q+\ell})$ by $\sum_{i=1}^{n-1} \psi_i(\mathbf{x})$ and $\psi_i \in \Sigma(\mathcal{A}_i^{q+\ell})$ ($i = 1, 2, \dots, n-1$) in the SOS relaxation (25), respectively, to obtain a new SOS relaxation of the POP (23):

$$(26) \quad \left. \begin{array}{ll} \text{Maximize} & \zeta \\ \text{subject to} & L(\mathbf{x}, \boldsymbol{\varphi}) - \zeta = \sum_{i=1}^{n-1} \psi_i(\mathbf{x}) \quad (\forall \mathbf{x} \in \mathbb{R}^n), \\ & \varphi_{ij} \in \Sigma(\mathcal{A}_i^q) \quad (i = 1, 2, \dots, n-1, j = 1, 2, \dots, m), \\ & \psi_i \in \Sigma(\mathcal{A}_i^{q+\ell}) \quad (i = 1, 2, \dots, n-1). \end{array} \right\}$$

It should be noted that the size of every support set in the SOS relaxation (26) is independent of the dimension n of the POP (23). When m and q are fixed, the total size of support sets in the SOS relaxation (26) grows linearly with the dimension n while the growth rate of the total size of support sets in the SOS relaxation (24) as well as that in the SOS relaxation (25) are of $O(n^{q+\ell})$. This shows that the SOS relaxation (26) has a considerable computational advantage in solving the POP (23) of large dimension n .

The numerical experiment was done using semidefinite programming algorithm (SDPA) 6.0 [21] on a 2.4 GHz Pentium IV (XEON) with 6GB memory, and the optimal values of the POP (23) with $m \in \{1, n, n^2\}$, $\ell \in \{1, 3\}$, and $n \in \{4, 5, 6, 7\}$ were computed by GloptiPoly [3]. Tables 6.1, 6.2, and 6.3 show numerical results from the three SOS relaxations (24), (25), and (26) of the POP (23) with $m \in \{1, n, n^2\}$, $\ell = 1$, and dimension $n \in \{4, 5, 6, 7\}$. We observe that

- all the SOS relaxations (24), (25), and (26) attain optimal values of the POP (23) with the lowest order $q = 2$.
- the SOS relaxation (25) requires less cpu time than the SOS relaxation (24), and the difference in cpu time becomes larger as m increases.

Table 6.4 shows numerical results from the three SOS relaxations (24), (25), and (26) of the POP (23) with $m = n$, $\ell = 3$, and dimension $n \in \{3, 4, 5, 6\}$. In this case,

- the SOS relaxations (25) and (26) obtain optimal values of the POP (23) or their lower bounds of (almost) the same quality as the SOS relaxation (24).
- the SOS relaxation (25) spends less cpu time than the SOS relaxation (24), but the difference is small.
- when $n = 6$, the SOS relaxation (24) with the order $q = 2$ attains the optimal value, but the other two SOS relaxations with the same order $q = 2$ provide only lower bounds for the optimal value. It should also be noted that the optimal value is obtained by the SOS relaxation (26) for the order $q = 3$.

In all cases reported in Tables 6.1, 6.2, 6.3 and 6.4, the SOS relaxation (26) has a clear advantage over the other SOS relaxations.

TABLE 6.4
Numerical results on the POP (23) with $m = n$ and $\ell = 3$.

POP (23)		cpu time in seconds (optimal value)		
n (optimal value)	q	relaxation (24)	relaxation (25)	relaxation (26)
3 (−1.782266)	0	0.1 (−148.0654)	0.1 (−148.0654)	0.1 (−148.0654)
	1	0.4 (−1.872454)	0.4 (−1.888884)	0.1 (−1.888890)
	2	1.9 (−1.782266)	1.6 (−1.782266)	0.2 (−1.782266)
4 (−2.244005)	0	0.4 (−129.5713)	0.4 (−129.5713)	0.1 (−129.5713)
	1	11.7 (−2.277639)	5.6 (−2.277639)	0.2 (−2.277844)
	2	46.2 (−2.244005)	36.1 (−2.244005)	0.4 (−2.244005)
5 (−3.848386)	0	2.6 (−120.1503)	2.5 (−120.2150)	0.1 (−120.1503)
	1	65.3 (−3.888779)	61.7 (−3.888779)	0.3 (−3.892605)
	2	787.1 (−3.848386)	644.6 (−3.848386)	0.8 (−3.848386)
6 (−3.531009)	0	13.3 (−120.2150)	13.4 (−120.2150)	0.1 (−120.2168)
	1	500.4 (−3.603920)	469.2 (−3.696910)	0.5 (−3.698462)
	2	11,912.4 (−3.531009)	11,718.1 (−3.535911)	1.4 (−3.537123)
	3	can't solve	can't solve	2.8 (−3.531009)

7. Concluding discussions. Considering the two types of POPs (2) and (3) obtained from different characterizations of the feasible region of the POP (1), we have proposed a sequence of SOS relaxations from generalized Lagrangian duals of each POP. We have also provided a theoretical foundation on the convergence of the sequence of SOS relaxations to the optimal value of (3). The sequence of SOS relaxations has been transformed into a sequence of SDP relaxations to solve the POP (1) computationally. We have shown that the sequence of SDP relaxations derived here has a primal-dual relationship with the one obtained after modifying Lasserre's SDP relaxation [10].

Theoretically, we have proved that the SOS relaxation of the Lagrangian dual of (3) attains the optimal value ζ^* of the POP (3). But there remains a gap between the Lagrangian dual of (2) and its SOS relaxation in section 4.2; the former attains ζ^* but the latter is not guaranteed to attain ζ^* . Thus it is interesting to prove or disprove that the SOS relaxation of the Lagrangian dual of (2) attains ζ^* . This will be a subject of future study.

The size of the SOS relaxation or the SDP relaxation obtained from the Lagrangian dual approach by exploiting sparsity is smaller than the size of Lasserre's SDP relaxation. This is of course a nice feature, but this may not necessarily mean that the former SDP relaxation is as effective as the latter in practice. To attain an approximation to the optimal value ζ^* of the POP (1) with accuracy as high as the approximation from Lasserre's SDP relaxation, we may need higher degree SOS polynomials in our dual approach, which makes the size of the resulting SDP relaxation larger.

One of the advantages of the proposed method is that it gives us much flexibility in implementation of the SOS relaxation and the SDP relaxation of the POP (1); sets of Lagrangian multiplier SOS polynomials satisfying the assumption of Theorem 3.1 can be freely chosen to strengthen the resulting relaxations. We have presented an illustrative example of how the framework of the proposed SOS relaxation can be used to obtain a practical SOS relaxation which exploits a structured sparsity. Numerical results of the example have indicated that it is possible to dramatically improve computational efficiency of SOS relaxations by making proper heuristic choices of supports, depending on the problems.

In addition to the preliminary numerical results in section 6, numerical experiments on the SDP relaxation with heuristically chosen supports were performed for various types of polynomial optimization problems with certain types of sparsity, but were not included here because we believe that the discussion of heuristics is beyond the scope of this paper. The main purpose of this paper has been to propose general methods for sparsity in SDP relaxations for polynomial optimization and introduce Lagrangian dual and penalty function approaches into SDP relaxations for polynomial optimization. Although the numerical results supported the claim that the SOS relaxations could improve the efficiency, to prove this claim it would be necessary to address issues such as (i) a reasonable definition of structured sparsity in polynomial optimization problems, (ii) technical details of heuristic choices of supports, and (iii) extensive numerical experiments on various problems with structured sparsity. These will be the subjects of a paper on practical performance of heuristics, which we hope to present in the near future.

REFERENCES

- [1] E. BALAS, S. CERIA, AND G. CORNUÉJOLS, *A lift-and-project cutting plane algorithm for mixed 0-1 programs*, Math. Program., 58 (1993), pp. 295–323.
- [2] M. D. CHOI, T. Y. LAM, AND B. REZNICK, *Sums of squares of real polynomials*, in *k-theory and Algebra Geometry: Connections with Quadratic Forms and Division Algebras*, Proc. Sympos. Pure Math, 58 Part 2, AMS, Providence RI, 1995, pp. 103–126.
- [3] D. HENRION AND J. B. LASSERRE, *GloptiPoly: Global optimization over polynomials with Matlab and SeDuMi*, ACM Trans. Math. Software, 29 (2003), pp. 165–194.
- [4] S. KIM AND M. KOJIMA, *Second order cone programming relaxation of nonconvex quadratic optimization problems*, Optim. Method Softw., 15 (2001), pp. 201–224.
- [5] S. KIM AND M. KOJIMA, *Exact solutions of some nonconvex quadratic optimization problems via SDP and SOCP relaxations*, Comput. Optim. Appl., 26 (2003), pp. 143–154.
- [6] M. KOJIMA, S. KIM, AND H. WAKI, *A general framework for convex relaxation of polynomial optimization problems over cones*, J. Oper. Res. Soc. Japan, 46 (2003), pp. 125–144.
- [7] M. KOJIMA, S. KIM, AND H. WAKI, *Sparsity in Sums of Squares of Polynomials*, Research Report B-391, Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, Meguro, Tokyo 2003.
- [8] M. KOJIMA AND L. TUNÇEL, *Cones of matrices and successive convex relaxations of nonconvex sets*, SIAM J. Optim., 10 (2000), pp. 750–778.
- [9] M. KOJIMA AND L. TUNÇEL, *Discretization and localization in successive convex relaxation methods for nonconvex quadratic optimization problems*, Math. Program., 89 (2000), pp. 79–111.
- [10] J. B. LASSERRE, *Global optimization with polynomials and the problem of moments*, SIAM J. Optim., 11 (2001), pp. 796–817.
- [11] L. LOVÁSZ, AND A. SCHRIJVER, *Cones of matrices and set functions and 0-1 optimization*, SIAM J. Optim., 1 (1991), pp. 166–190.
- [12] R. D. C. MONTEIRO, *First- and second-order methods for semidefinite programming*, Math. Program., 97 (2003), pp. 209–244.
- [13] P. A. PARRILO, *Semidefinite programming relaxations for semialgebraic problems*, Math. Program., 96 (2003), pp. 293–320.
- [14] S. PRAJNA, A. PAPACHRISTODOULOU, AND P. A. PARRILO, *SOSTOOLS: Sum of Squares Optimization Toolbox for MATLAB – User’s Guide*, Control and Dynamical Systems, California Institute of Technology, Pasadena, CA, 2002. Available online at <http://www.eds.caltech.edu/sostools>
- [15] V. POWERS AND T. WÖRMANN, *An algorithm for sums of squares of real polynomials*, J. Pure Appl. Algebra, 127 (1998), pp. 99–104.
- [16] M. PUTINAR, *Positive polynomials on compact semi-algebraic sets*, Indiana U. Math. J., 42 (1993), pp. 969–984.
- [17] B. REZNICK, *Extremal psd forms with few terms*, Duke Math. J., 45 (1978), pp. 363–374.
- [18] H. D. SHERALI AND W. P. ADAMS, *A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems*, SIAM J. Discrete Math., 3 (1990), pp. 411–430.

- [19] H. D. SHERALI AND C. H. TUNCBILEK, *A global optimization algorithm for polynomial programming problems using a reformulation-linearization technique*, J. Global Optim., 2 (1992), pp. 101–112.
- [20] H. D. SHERALI AND W. P. ADAMS, *A hierarchy of relaxations and convex hull characterizations for mixed-integer zero-one programming problems*, Discrete Appl. Math., 52 (1994), pp. 83–106.
- [21] M. YAMASHITA, K. FUJISAWA, AND M. KOJIMA, *Implementation and evaluation of SDPA 6.0 (semidefinite programming algorithm 6.0)*, Optim. Method Softw., 18 (2003), pp. 491–505.

AN INTERIOR POINT METHOD FOR MATHEMATICAL PROGRAMS WITH COMPLEMENTARITY CONSTRAINTS (MPCCs)*

ARVIND U. RAGHUNATHAN[†] AND LORENZ T. BIEGLER[†]

Abstract. Interior point methods for nonlinear programs (NLPs) are adapted for solution of mathematical programs with complementarity constraints (MPCCs). The constraints of the MPCC are suitably relaxed so as to guarantee a strictly feasible interior for the inequality constraints. The standard primal-dual algorithm has been adapted with a modified step calculation. The algorithm is shown to be superlinearly convergent in the neighborhood of the solution set under assumptions of MPCC-LICQ, strong stationarity and upper level strict complementarity. The modification can be easily accommodated within most nonlinear programming interior point algorithms with identical local behavior. Numerical experience is also presented and holds promise for the proposed method.

Key words. MPEC, equilibrium constraints, complementarity, nonlinear programming, interior point methods

AMS subject classifications. 90C30, 90C33, 90C51, 49M37, 65K10

DOI. 10.1137/S1052623403429081

1. Introduction. The mathematical program with complementarity constraints (MPCC) considered in the paper is

$$(1.1) \quad \begin{array}{ll} \min_{x,w,y} & f(x, w, y) \\ \text{s.t.} & h(x, w, y) = 0, \\ & x \geq 0, \\ & 0 \leq w \perp y \geq 0, \end{array}$$

where $x \in \mathbb{R}^n$, $w, y \in \mathbb{R}^m$, $f : \mathbb{R}^{n+2m} \rightarrow \mathbb{R}$, $h : \mathbb{R}^{n+2m} \rightarrow \mathbb{R}^q$. For ease of presentation, we consider simple bounds on x , as general inequality constraints can be posed in a similar form by the introduction of slacks. The above problem is difficult to solve due to the nonsatisfaction of constraint qualification (CQ), such as the Mangasarian–Fromovitz constraint qualification (MFCQ), which is readily satisfied for nonlinear programs (NLPs). The failure of MPCC (1.1) to satisfy MFCQ renders the set of multipliers, if nonempty, unbounded. This has been put forth as an argument for inapplicability of NLP algorithms in the context of MPCCs. A number of algorithms specific to solving mathematical programs with equilibrium constraints (MPECs) and MPCCs [14, 16] have been proposed. Global [11, 14, 19] and local [5, 15] convergence of the algorithms have also been studied.

In a parallel development researchers have investigated the applicability of nonlinear programming algorithms for solving MPCCs. Animescu [1], in a recent work, has investigated the conditions for the existence of multipliers for MPCCs. Fletcher et al. [7] have shown local convergence of a SQP algorithm when applied to MPCCs.

*Received by the editors May 30, 2003; accepted for publication (in revised form) April 23, 2004; published electronically April 8, 2005.

<http://www.siam.org/journals/siopt/15-3/42908.html>

[†]Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 (aur@andrew.cmu.edu, lb01@andrew.cmu.edu).

Anitescu [1] and Fletcher et al. [7] pose the MPCC as the following NLP:

$$(1.2) \quad \begin{aligned} \min_{x,w,y} \quad & f(x, w, y) \\ \text{s.t.} \quad & h(x, w, y) = 0, \\ & w^T y \leq 0, \\ & x, w, y \geq 0, \end{aligned}$$

where all the complementarity constraints are aggregated to a single constraint. Interior point approaches have also been proposed for the solution of the MPCC. Benson, Shanno, and Vanderbei [2] pose the complementarity constraints as

$$(1.3) \quad \begin{aligned} w_i y_i &\leq 0, \\ w_i, y_i &\geq 0, \end{aligned} \quad i = 1, \dots, m,$$

and solve the resulting problem with an NLP interior point algorithm LOQO [20]. In the above, w_i and y_i denote i th component of the vectors w and y , respectively. Observe that the problem does not possess strictly feasible points, yet Benson, Shanno, and Vanderbei [2] have reported encouraging numerical results on the MacMPEC test suite [12]. Liu and Sun [13] relax the complementarity constraint as

$$(1.4) \quad \begin{aligned} w_i y_i &\leq t, \\ w_i, y_i &\geq 0, \end{aligned} \quad i = 1, \dots, m,$$

for $t > 0$. The parameter t is driven to zero in the limit. The authors describe an interior point approach that achieves global convergence for a given relaxation parameter t under certain assumptions. De Miguel et al. [3] propose solving the MPCC by relaxing the complementarity constraints and the bounds on the complementary variables as

$$\begin{aligned} w_i y_i &\leq t_i, \\ w_i, y_i &\geq -\theta_i, \end{aligned} \quad i = 1, \dots, m,$$

for $t_i, \theta_i > 0$. De Miguel et al. show that in the limit only one of t_i, θ_i needs to be driven to zero. As a result, a problem with a nonempty interior can be guaranteed under standard MPCC regularity assumptions and an NLP interior point algorithm can be readily used. De Miguel et al. prove superlinear convergence of their algorithm. In this paper we employ an approach similar to Liu and Sun [13] to relax the complementarity constraints. We present a modified NLP barrier algorithm and show that it achieves superlinear convergence in the neighborhood of a solution.

MPECs allow specification of parametric variational inequalities as constraints and generalize MPCCs. A survey of the applications involving MPECs and algorithms can be found in the monograph of Luo, Pang, and Ralph [14]. Bilevel programs with convex inner problems can be converted to an MPEC, with the parametric variational inequality representing the first order optimality conditions of the program. The MPEC can be converted to an MPCC using the Karush–Kuhn–Tucker (KKT) conditions of the variational inequality, provided the constraint set is expressed as a finite set of inequalities that satisfy a CQ. The theory that is developed in this work can also be applied to the above class of bilevel problems and MPECs.

The paper is organized as follows. We define the constraint qualification, stationarity, and second order conditions associated with MPCCs in section 2. A

characterization of the solution set of the MPCC when posed as an NLP (2.5) and some related results are also presented. A modified interior point algorithm is presented in section 3. The local convergence analysis in section 4 shows that the generated iterates converge quadratically to a solution of the MPCC. We present numerical experience using our proposed algorithm on a test suite of MPCCs in section 5 and conclude in section 6.

Notation. We will denote the i th component of a vector $u \in \mathbb{R}^n$ by u_i and denote by u^k the k th iterate generated by an algorithm or the k th element of a sequence which will be clear from the context. Given $u \in \mathbb{R}^n$, the uppercase letter U will denote an $n \times n$ matrix with elements of the vector u on the diagonal, i.e., $U = \text{diag}(u_1, \dots, u_n)$. Norms $\|\cdot\|$ will denote a fixed vector norm and its compatible matrix norm unless otherwise noted. Given $x, y \in \mathbb{R}^n$, $\min(x, y)$ denotes the componentwise minimum, i.e., $(\min(x, y))_i = \min(x_i, y_i)$. We will use the order notation in the following standard way. We will denote by $\mathcal{O}(t^k)$ a sequence $\{u^k\}$ satisfying $\|u^k\| \leq \beta_1 t^k$ for some constant $\beta_1 > 0$ independent of k , by $\Theta(t^k)$ a sequence $\{u^k\}$ satisfying $\beta_2 t^k \leq \|u^k\| \leq \beta_3 t^k$ for some $\beta_2, \beta_3 > 0$ independent of k , and by $o(t^k)$ a sequence $\{u^k\}$ satisfying $\lim_{k \rightarrow \infty} (u^k/t^k) = 0$.

2. Preliminaries. We will first state some definitions. For any feasible point $\tilde{z} := (\tilde{x}, \tilde{w}, \tilde{y})$ of the MPCC, we associate a *relaxed* NLP, in the sense of Luo et al. [15] and Scheel and Scholtes [17], as follows:

$$\begin{aligned}
 & \min_{x, w, y} && f(x, w, y) \\
 & \text{s.t.} && h(x, w, y) = 0, \\
 & && x \geq 0, \\
 (2.1) \quad & && w_i = 0, \quad i \in I_w(\tilde{z}) \setminus I_y(\tilde{z}), \\
 & && y_i = 0, \quad i \in I_y(\tilde{z}) \setminus I_w(\tilde{z}), \\
 & && w_i \geq 0, \quad i \in I_y(\tilde{z}), \\
 & && y_i \geq 0, \quad i \in I_w(\tilde{z}),
 \end{aligned}$$

where the index sets are defined as $I_w(\tilde{z}) := \{i \mid \tilde{w}_i = 0\}$ and $I_y(\tilde{z}) := \{i \mid \tilde{y}_i = 0\}$. Due to the complementarity constraint, $I_w(\tilde{z}) \cup I_y(\tilde{z}) = \{1, \dots, m\}$. The set $I_w(\tilde{z}) \cap I_y(\tilde{z})$ consists of indices $i : \tilde{w}_i = \tilde{y}_i = 0$. Let $\mathcal{F}_{\text{MPCC}}$ and $\mathcal{F}_{\text{RNLP}}$, respectively, denote the feasible regions of MPCC (1.1) and relaxed NLP (2.1). Note that the relaxed NLP (2.1) does not enforce complementarity for $i \in I_w(\tilde{z}) \cap I_y(\tilde{z})$. Consequently, there exists a neighborhood $\mathcal{U}(\tilde{z})$ such that $\mathcal{F}_{\text{MPCC}} \cap \mathcal{U}(\tilde{z}) \subseteq \mathcal{F}_{\text{RNLP}} \cap \mathcal{U}(\tilde{z})$. This type of relationship has been exploited by Luo, Pang, and Ralph [15] and Scheel and Scholtes [17] to obtain stationary conditions of the MPCC in terms of stationary conditions of the relaxed NLP for which constraint qualifications such as linear independence of active constraint gradients (LICQ) readily hold.

DEFINITION 2.1 (MPCC-LICQ). *The MPCC (1.1) satisfies an MPCC-LICQ at a feasible point $\tilde{z}$ if the corresponding relaxed NLP (2.1) satisfies LICQ at $\tilde{z}$.*

DEFINITION 2.2 (strong stationarity). *A point $\bar{z}$ feasible to the MPCC (1.1) is called a strong stationary point of the MPCC if there exists a multiplier $\bar{\lambda} := (\bar{\lambda}_h, \bar{\lambda}_x, \bar{\lambda}_w, \bar{\lambda}_y)$ such that $(\bar{z}, \bar{\lambda})$ satisfy NLP stationarity for the relaxed NLP (2.1), where $\bar{\lambda}_h \in \mathbb{R}^q$, $\bar{\lambda}_x \in \mathbb{R}^n$, $\bar{\lambda}_w, \bar{\lambda}_y \in \mathbb{R}^m$.*

Definition 2.2 implies that if $\bar{z}$ is a strongly stationary point of the MPCC, then

$$(2.2) \quad \begin{aligned} & \begin{bmatrix} \nabla_x f(\bar{z}) \\ \nabla_w f(\bar{z}) \\ \nabla_y f(\bar{z}) \end{bmatrix} + \begin{bmatrix} \nabla_x h(\bar{z}) \\ \nabla_w h(\bar{z}) \\ \nabla_y h(\bar{z}) \end{bmatrix} \bar{\lambda} - \begin{bmatrix} \bar{\lambda}_x \\ \bar{\lambda}_w \\ \bar{\lambda}_y \end{bmatrix} = 0, \\ & \bar{\lambda}_{x,i} \geq 0, \quad \bar{x}_i \bar{\lambda}_{x,i} = 0, \quad i = 1, \dots, n, \\ & \bar{\lambda}_{w,i}, \bar{\lambda}_{y,i} \geq 0, \quad i \in I_w(\bar{z}) \cap I_y(\bar{z}), \\ & \bar{w}_i \bar{\lambda}_{w,i} = 0, \quad i = 1, \dots, m, \\ & \bar{y}_i \bar{\lambda}_{y,i} = 0, \quad i = 1, \dots, m, \\ & \bar{z} \in \mathcal{F}_{\text{MPCC}}. \end{aligned}$$

If the MPCC satisfies MPCC-LICQ at $\bar{z}$, then the multiplier $\bar{\lambda}$ is unique.

We now state a second order sufficient condition for MPCCs. The set of critical directions, $\mathcal{T}$, is defined as

$$(2.3) \quad \mathcal{T} = \left\{ d = \begin{bmatrix} dx \\ dw \\ dy \end{bmatrix} \neq 0 \left| \begin{array}{lll} \nabla_z h(\bar{z})^T d = 0, \\ dx_i = 0, & i : \bar{x}_i = 0, & \bar{\lambda}_{x,i} > 0 \\ dx_i \geq 0, & i : \bar{x}_i = 0, & \bar{\lambda}_{x,i} = 0 \\ dw_i = 0, & i : i \in I_w(\bar{z}), & \bar{\lambda}_{w,i} \neq 0 \\ dw_i \geq 0, & i : i \in I_w(\bar{z}), & \bar{\lambda}_{w,i} = 0 \\ dy_i = 0, & i : i \in I_y(\bar{z}), & \bar{\lambda}_{y,i} \neq 0 \\ dy_i \geq 0, & i : i \in I_y(\bar{z}), & \bar{\lambda}_{y,i} = 0 \end{array} \right. \right\}.$$

DEFINITION 2.3 (MPCC-SOSC). A point $\bar{z}$ satisfies an MPCC-SOSC [18] if strong stationarity (2.2) holds and

$$d^T \nabla_{zz} \mathcal{L}^{\text{RNLP}}(\bar{z}, \bar{\lambda}_h, \bar{\lambda}_x, \bar{\lambda}_w, \bar{\lambda}_y) d > 0 \quad \forall d \in \mathcal{T},$$

where $\nabla_{zz} \mathcal{L}^{\text{RNLP}}$ is the Hessian of the Lagrangian of the relaxed NLP, $\mathcal{L}^{\text{RNLP}}$,

$$(2.4) \quad \mathcal{L}^{\text{RNLP}}(z, \lambda_h, \lambda_x, \lambda_w, \lambda_y) := f(z) + \lambda_h^T h(z) - \lambda_x^T x - \lambda_w^T w - \lambda_y^T y,$$

and $\mathcal{T}$ is the set of critical directions defined above.

For the remainder of the paper, we make the following standing assumptions.

Assumptions (A).

(A1) The functions f, h are all twice Lipschitz continuously differentiable.

(A2) $\bar{z} := (\bar{x}, \bar{w}, \bar{y})$ is a strong stationary point of MPCC (1.1) with associated MPCC multiplier, $\bar{\lambda} := (\bar{\lambda}_h, \bar{\lambda}_x, \bar{\lambda}_w, \bar{\lambda}_y)$, and MPCC-LICQ, MPCC-SOSC hold.

(A3) Strict complementarity is required to hold for certain components of the MPCC multiplier, $\bar{\lambda}$, i.e., $\bar{x}_i + \bar{\lambda}_{x,i} > 0 \quad \forall i \in 1, \dots, n$ and $\bar{\lambda}_{w,i}, \bar{\lambda}_{y,i} > 0 \quad \forall i \in I_w(\bar{z}) \cap I_y(\bar{z})$.

The assumption of strict complementarity for $i \in I_w(\bar{z}) \cap I_y(\bar{z})$ is also the upper level strict complementarity assumption introduced by Scheel and Scholtes [17] and employed by Scholtes [18]. Assumption (A3) is weaker than requiring $\bar{w}_i + \bar{\lambda}_{w,i} \neq 0$ for all $i \in I_w(\bar{z})$, $\bar{y}_i + \bar{\lambda}_{y,i} \neq 0$ for all $i \in I_y(\bar{z})$ and $\bar{x}_i + \bar{\lambda}_{x,i} > 0$ for all $i = 1, \dots, n$.

In this paper, we will pose the MPCC (1.1) equivalently as the following NLP:

$$(2.5) \quad \begin{aligned} & \min_{x,w,y} \quad f(x, w, y) \\ & \text{s.t.} \quad h(x, w, y) = 0, \\ & \quad Wy \leq 0, \\ & \quad x, w, y \geq 0, \end{aligned}$$

where $W = \text{diag}(w_1, \dots, w_m)$ is a diagonal matrix with components of w on the diagonal. A point $(\bar{x}, \bar{w}, \bar{y}) \in \mathcal{F}_{\text{MPCC}}$ satisfies stationary conditions of the NLP (2.5) if there exists $\lambda := (\lambda_h, \lambda_{cc}, \lambda_x, \lambda_w, \lambda_y) \in \mathbb{R}^q \times \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$ satisfying

$$\begin{aligned}
 (2.6a) \quad & \begin{bmatrix} \nabla_x f(\bar{z}) \\ \nabla_w f(\bar{z}) \\ \nabla_y f(\bar{z}) \end{bmatrix} + \begin{bmatrix} \nabla_x h(\bar{z}) \\ \nabla_w h(\bar{z}) \\ \nabla_y h(\bar{z}) \end{bmatrix} \lambda_h + \begin{bmatrix} 0 \\ \bar{Y} \\ \bar{W} \end{bmatrix} \lambda_{cc} - \begin{bmatrix} \lambda_x \\ \lambda_w \\ \lambda_y \end{bmatrix} = 0, \\
 (2.6b) \quad & \lambda_{cc,i} \geq 0, \quad \lambda_{cc,i}(\bar{w}_i \bar{y}_i) = 0, \quad i = 1, \dots, m, \\
 (2.6c) \quad & \lambda_{x,i} \geq 0, \quad \lambda_{x,i} \bar{x}_i = 0, \quad i = 1, \dots, n, \\
 (2.6d) \quad & \lambda_{w,i} \geq 0, \quad \lambda_{w,i} \bar{w}_i = 0, \quad i = 1, \dots, m, \\
 (2.6e) \quad & \lambda_{y,i} \geq 0, \quad \lambda_{y,i} \bar{y}_i = 0, \quad i = 1, \dots, m, \\
 (2.6f) \quad & \bar{z} \in \mathcal{F}_{\text{MPCC}}.
 \end{aligned}$$

The following proposition relates stationary conditions of MPCC (2.2) and NLP (2.6). This result is closely related to [1, Theorem 2.2].

PROPOSITION 2.4 (NLP-MPCC stationarity). *$\bar{z}$ satisfies the NLP stationary conditions (2.6) of NLP (2.5) if and only if $\bar{z}$ satisfies the strong stationary conditions (2.2) of MPCC (1.1), where $\bar{z} := (\bar{x}, \bar{w}, \bar{y})$.*

Proof. The NLP formulation (2.5) ensures that $\bar{z}$ is feasible to the MPCC (1.1) if and only if it is feasible to the NLP (2.5). It remains for us to verify the remaining conditions.

Suppose $(\bar{x}, \bar{w}, \bar{y}, \lambda_h, \lambda_{cc}, \lambda_x, \lambda_w, \lambda_y)$ satisfies the NLP stationary conditions (2.6). Then define the MPCC multipliers as follows:

$$\left. \begin{aligned} \bar{\lambda}_h &:= \lambda_h \\ \bar{\lambda}_x &:= \lambda_x \\ \bar{\lambda}_w &:= \lambda_w - \bar{Y} \lambda_{cc} \\ \bar{\lambda}_y &:= \lambda_y - \bar{W} \lambda_{cc} \end{aligned} \right\} \Rightarrow \left. \begin{aligned} \bar{\lambda}_{w,i} &= \lambda_{w,i} \stackrel{(2.6d)}{\geq} 0 \\ \bar{\lambda}_{y,i} &= \lambda_{y,i} \stackrel{(2.6e)}{\geq} 0 \end{aligned} \right\} \quad \forall i \in I_w(\bar{z}) \cap I_y(\bar{z}).$$

Hence, $(\bar{x}, \bar{w}, \bar{y}, \bar{\lambda}_h, \bar{\lambda}_x, \bar{\lambda}_w, \bar{\lambda}_y)$ solves the strong stationarity conditions (2.2).

For the reverse implication, suppose $(\bar{x}, \bar{w}, \bar{y}, \bar{\lambda}_h, \bar{\lambda}_x, \bar{\lambda}_w, \bar{\lambda}_y)$ satisfies the strong stationarity conditions (2.2). In order to satisfy the NLP stationary conditions we need to find $(\lambda_{cc}, \lambda_w, \lambda_y) \geq 0$ such that

$$\begin{aligned}
 -\bar{Y} \lambda_{cc} + \lambda_w &= \bar{\lambda}_w, \\
 -\bar{W} \lambda_{cc} + \lambda_y &= \bar{\lambda}_y
 \end{aligned}$$

holds. Define the NLP multipliers as

$$\begin{aligned}
 (2.7) \quad & \lambda_h := \bar{\lambda}_h, \quad \lambda_x := \bar{\lambda}_x, \\
 & \lambda_{cc,i} := \begin{cases} \max \left\{ -\frac{\bar{\lambda}_{w,i}}{\bar{y}_i}, 0 \right\}, & i \in I_w(\bar{z}) \setminus I_y(\bar{z}), \\ 0, & i \in I_w(\bar{z}) \cap I_y(\bar{z}), \\ \max \left\{ -\frac{\bar{\lambda}_{y,i}}{\bar{w}_i}, 0 \right\}, & i \in I_y(\bar{z}) \setminus I_w(\bar{z}), \end{cases} \\
 & \lambda_{w,i} := \begin{cases} \max \{ 0, \bar{\lambda}_{w,i} \}, & i \in I_w(\bar{z}) \setminus I_y(\bar{z}), \\ \bar{\lambda}_{w,i}, & i \in I_y(\bar{z}), \end{cases} \\
 & \lambda_{y,i} := \begin{cases} \bar{\lambda}_{y,i}, & i \in I_w(\bar{z}), \\ \max \{ 0, \bar{\lambda}_{y,i} \}, & i \in I_y(\bar{z}) \setminus I_w(\bar{z}). \end{cases}
 \end{aligned}$$

Then $(\bar{x}, \bar{w}, \bar{y}, \lambda_h, \lambda_{cc}, \lambda_x, \lambda_w, \lambda_y)$ solves the NLP stationary conditions (2.6). $\square$

The multipliers to the NLP obtained in the proof are analogous to the *basic* multipliers that Fletcher et al. [7] describe. We will denote the multipliers as λ^{basic} .

The following are results concerning the solution set of the NLP (2.5), which poses the complementarity constraint as an inequality. Given a strong stationary point $\bar{z}$, $\mathcal{S}_\lambda$ will denote the set of multipliers satisfying NLP stationary conditions (2.6) at the point $\bar{z}$. The set $\mathcal{S}_\lambda$ is defined as

$$(2.8) \quad \mathcal{S}_\lambda := \left\{ \lambda \left| \begin{array}{l} \lambda_h = \bar{\lambda}_h, \\ \lambda_x = \bar{\lambda}_x, \\ \lambda_{cc}, \lambda_w, \lambda_y \geq 0, \end{array} \quad \begin{array}{l} \bar{Y}\lambda_{cc} - \lambda_w = -\bar{\lambda}_w \\ \bar{W}\lambda_{cc} - \lambda_y = -\bar{\lambda}_y \end{array} \right. \right\}.$$

Since our interest is in the behavior near a particular strongly stationary point, $\bar{z}$, we will suppress the dependence of $\mathcal{S}_\lambda$ on $\bar{z}$. Further, our assumption of MPCC-LICQ (A2) renders $\bar{\lambda}$ unique and there exists no ambiguity in the definition of $\mathcal{S}_\lambda$. Using the definition of $\mathcal{S}_\lambda$ we can prove the following result on λ^{basic} .

PROPOSITION 2.5 (characterization of λ^{basic}). *Let $\bar{z}$ be a strong stationary point of MPCC (1.1) satisfying MPCC-LICQ. Then*

$$(2.9) \quad \lambda^{\text{basic}} = \arg \min_{\lambda \in \mathcal{S}_\lambda} \|\lambda\|_1.$$

Proof. From the assumption of MPCC-LICQ, the multiplier $\bar{\lambda}$ is unique and as a result there is ambiguity only in the choice of λ_{cc} , λ_w , and λ_y . The form of the objective and constraints in (2.9) imply that we have a separable problem. Hence, it suffices to show that for each $i \in \{1, \dots, m\}$, $(\lambda_{cc,i}^{\text{basic}}, \lambda_{w,i}^{\text{basic}}, \lambda_{y,i}^{\text{basic}})$ solves

$$(2.10) \quad \begin{array}{ll} \min_{\lambda_{cc,i}, \lambda_{w,i}, \lambda_{y,i}} & \lambda_{cc,i} + \lambda_{w,i} + \lambda_{y,i} \\ \text{s.t.} & \begin{array}{l} \bar{y}_i \lambda_{cc,i} - \lambda_{w,i} = -\bar{\lambda}_{w,i}, \\ \bar{w}_i \lambda_{cc,i} - \lambda_{y,i} = -\bar{\lambda}_{y,i}, \\ \lambda_{cc,i}, \lambda_{w,i}, \lambda_{y,i} \geq 0, \end{array} \end{array} \quad \equiv \quad \begin{array}{ll} \min_{\lambda_{cc,i}} & (1 + \bar{w}_i + \bar{y}_i) \lambda_{cc,i} \\ \text{s.t.} & \begin{array}{l} \bar{y}_i \lambda_{cc,i} + \bar{\lambda}_{w,i} \geq 0, \\ \bar{w}_i \lambda_{cc,i} + \bar{\lambda}_{y,i} \geq 0, \\ \lambda_{cc,i} \geq 0, \end{array} \end{array}$$

where the reduced problem follows from the elimination of $\lambda_{w,i}$ and $\lambda_{y,i}$. The bounds on $\lambda_{cc,i}$ can be decomposed further based on $I_w(\bar{z})$ and $I_y(\bar{z})$ as

$$\begin{aligned} \lambda_{cc,i} &\geq \max \left(0, -\frac{\bar{\lambda}_{w,i}}{\bar{y}_i} \right) & : & \text{ if } i \in I_w(\bar{z}) \setminus I_y(\bar{z}), \\ \lambda_{cc,i} &\geq 0 & : & \text{ if } i \in I_w(\bar{z}) \cap I_y(\bar{z}), \\ \lambda_{cc,i} &\geq \max \left(0, -\frac{\bar{\lambda}_{y,i}}{\bar{w}_i} \right) & : & \text{ if } i \in I_y(\bar{z}) \setminus I_w(\bar{z}). \end{aligned}$$

From the above it is clear that the solution $(\lambda_{cc,i}^*, \lambda_{w,i}^*, \lambda_{y,i}^*)$ to (2.10) satisfies

$$\lambda_{cc,i}^* = \begin{cases} \max \left(0, -\frac{\bar{\lambda}_{w,i}}{\bar{y}_i} \right) & \text{ if } i \in I_w(\bar{z}) \setminus I_y(\bar{z}), \\ 0 & \text{ if } i \in I_w(\bar{z}) \cap I_y(\bar{z}), \\ \max \left(0, -\frac{\bar{\lambda}_{y,i}}{\bar{w}_i} \right) & \text{ if } i \in I_y(\bar{z}) \setminus I_w(\bar{z}), \end{cases} \quad \begin{aligned} \lambda_{w,i}^* &= \bar{y}_i \lambda_{cc,i}^* + \bar{\lambda}_{w,i}, \\ \lambda_{y,i}^* &= \bar{w}_i \lambda_{cc,i}^* + \bar{\lambda}_{y,i}. \end{aligned}$$

It is easily seen that $(\lambda_{cc,i}^*, \lambda_{w,i}^*, \lambda_{y,i}^*)$ is the same as $(\lambda_{cc,i}^{\text{basic}}, \lambda_{w,i}^{\text{basic}}, \lambda_{y,i}^{\text{basic}})$ (2.7). $\square$

The above result continues to hold for any p -norm with $p < \infty$. The existence of λ^{basic} given $(\bar{z}, \bar{\lambda})$ satisfying the strong stationary conditions of the MPCC yields that $\mathcal{S}_\lambda \neq \emptyset$. Hence, the Euclidean projection of any (z, λ) onto the solution set of NLP, $\{\bar{z}\} \times \mathcal{S}_\lambda$, which we denote as $(\hat{z}, \hat{\lambda})$, is well defined. Furthermore, the satisfaction of MPCC-SOSC at $\bar{z}$ yields the following result regarding the satisfaction of a second order condition for the NLP (2.5) on the subspace $\mathcal{T}_\lambda^{\text{NLP}}$ for $\lambda \in \mathcal{S}_\lambda$:

$$(2.11) \quad \mathcal{T}_\lambda^{\text{NLP}} = \left\{ d = \begin{bmatrix} dx \\ dw \\ dy \end{bmatrix} \neq 0 \left| \begin{array}{l} \nabla_z h(\bar{z})^T d = 0, \\ dx_i = 0, \quad i : \bar{x}_i = 0, \lambda_{x,i} > 0 \\ dx_i \geq 0, \quad i : \bar{x}_i = 0, \lambda_{x,i} = 0 \\ dw_i = 0, \quad i : i \in I_w(\bar{z}), \lambda_{w,i} > 0 \\ dw_i \geq 0, \quad i : i \in I_w(\bar{z}), \lambda_{w,i} = 0 \\ dy_i = 0, \quad i : i \in I_y(\bar{z}), \lambda_{y,i} > 0 \\ dy_i \geq 0, \quad i : i \in I_y(\bar{z}), \lambda_{y,i} = 0 \\ \bar{y}_i dw_i + \bar{w}_i dy_i = 0, \quad i : \lambda_{cc,i} > 0 \\ \bar{y}_i dw_i + \bar{w}_i dy_i \leq 0, \quad i : \lambda_{cc,i} = 0 \end{array} \right. \right\}.$$

LEMMA 2.6 (NLP-SOSC). *Suppose Assumptions (A) hold. Then*

$$d^T \nabla_{zz} \mathcal{L}^{\text{NLP}}(\bar{z}, \lambda) d > 0 \quad \forall d \in \mathcal{T}_\lambda^{\text{NLP}}, \lambda \in \mathcal{S}_\lambda,$$

where $\mathcal{L}^{\text{NLP}}$ is the Hessian of the Lagrangian of the NLP (2.5), defined as

$$(2.12) \quad \mathcal{L}^{\text{NLP}}(z, \lambda) := f(z) + \lambda_h^T h(z) + \lambda_{cc}^T (Wy) - \lambda_x^T x - \lambda_w^T w - \lambda_y^T y.$$

Proof. We first show that $\mathcal{T}_\lambda^{\text{NLP}} \subseteq \mathcal{T} \forall \lambda \in \mathcal{S}_\lambda$. Next, the terms in $\mathcal{L}^{\text{NLP}}$ due to $Wy \leq 0$ are shown not to contribute to the curvature of the Hessian of $\mathcal{L}^{\text{NLP}}$ for all $d \in \mathcal{T}_\lambda^{\text{NLP}}$ and for all $\lambda \in \mathcal{S}_\lambda$. This allows us to prove the claim by noting that the Lagrangian of the NLP (2.12) differs from the Lagrangian for the relaxed NLP (2.4) only in the terms due to the constraint $Wy \leq 0$.

From the definition of $\mathcal{T}_\lambda^{\text{NLP}}$ (2.11) we observe that for all $i \in I_w(\bar{z})$ and $\lambda \in \mathcal{S}_\lambda$,

$$dw_i \geq 0, \quad \bar{y}_i dw_i \leq 0 \quad \forall d \in \mathcal{T}_\lambda^{\text{NLP}}$$

with one of the inequalities possibly required to hold as an equality. If $i \in I_w(\bar{z}) \setminus I_y(\bar{z})$, i.e., $\bar{y}_i > 0$, then $dw_i = 0$. As for $i \in I_w(\bar{z}) \cap I_y(\bar{z})$, Assumption (A3) yields that $dw_i = 0$ since $\lambda_{w,i} \geq \lambda_{w,i}^{\text{basic}} > 0$ for all $\lambda \in \mathcal{S}_\lambda$ (refer Lemma 2.5). A similar result can be shown to hold for dy_i for $i \in I_y(\bar{z})$. Hence,

$$(2.13) \quad \left. \begin{array}{l} dw_i = 0 \quad \forall i \in I_w(\bar{z}) \\ dy_i = 0 \quad \forall i \in I_y(\bar{z}) \end{array} \right\} \quad \forall d \in \mathcal{T}_\lambda^{\text{NLP}}, \lambda \in \mathcal{S}_\lambda.$$

On the other hand, the critical directions in $\mathcal{T}$ satisfy

$$\begin{array}{ll} dw_i = 0 & \text{if } i \in I_w(\bar{z}) : \bar{\lambda}_{w,i} \neq 0, \quad dy_i = 0 \quad \text{if } i \in I_y(\bar{z}) : \bar{\lambda}_{y,i} \neq 0, \\ dw_i \geq 0 & \text{if } i \in I_w(\bar{z}) : \bar{\lambda}_{w,i} = 0, \quad dy_i \geq 0 \quad \text{if } i \in I_y(\bar{z}) : \bar{\lambda}_{y,i} = 0. \end{array}$$

Further, the definition of $\mathcal{S}_\lambda$ (2.8) shows that $\lambda_x = \bar{\lambda}_x \forall \lambda \in \mathcal{S}_\lambda$. As a result, we obtain

$$(2.14) \quad \mathcal{T}_\lambda^{\text{NLP}} \subseteq \mathcal{T} \quad \forall \lambda \in \mathcal{S}_\lambda.$$

From (2.13) observe that

$$(2.15) \quad \left. \begin{aligned} dw_i = 0 \text{ or } dy_i = 0 \quad \forall i = 1, \dots, m \\ \Rightarrow d^T(\nabla_{zz}(w_i y_i)(\bar{z}))d = 2(dw_i)(dy_i) = 0 \end{aligned} \right\} \quad \forall d \in \mathcal{T}_\lambda^{\text{NLP}}, \lambda \in \mathcal{S}_\lambda.$$

In other words, the additional terms in the Lagrangian of NLP (2.5) do not contribute to the curvature of the Hessian of $\mathcal{L}^{\text{NLP}} \forall d \in \mathcal{T}_\lambda^{\text{NLP}}$ and $\lambda \in \mathcal{S}_\lambda$.

Hence,

$$\begin{aligned} d^T \nabla_{zz} \mathcal{L}^{\text{NLP}}(\bar{z}, \lambda) d &\stackrel{(2.12)}{=} d^T \left(\nabla_{zz} f(\bar{z}) + \sum_{j=1}^q \lambda_{h,j} \nabla_{zz} h_j(\bar{z}) \right. \\ &\quad \left. + \sum_{j=1}^m \lambda_{cc,i} \nabla_{zz}(w_j y_j)(\bar{z}) \right) d \\ &\stackrel{(2.15)}{=} d^T \left(\nabla_{zz} f(\bar{z}) + \sum_{j=1}^q \lambda_{h,j} \nabla_{zz} h_j(\bar{z}) \right) d \\ &\stackrel{(2.8)}{=} d^T \left(\nabla_{zz} f(\bar{z}) + \sum_{j=1}^q \bar{\lambda}_{h,j} \nabla_{zz} h_j(\bar{z}) \right) d \\ &\stackrel{(2.4)}{=} d^T \nabla_{zz} \mathcal{L}^{\text{RNLP}}(\bar{z}, \bar{\lambda}) d \\ &\stackrel{(2.14)}{>} 0 \quad \forall d \in \mathcal{T}_\lambda^{\text{NLP}}, \lambda \in \mathcal{S}_\lambda, \end{aligned}$$

proving the claim. $\square$

To summarize, the assumptions of MPCC-LICQ, strong stationarity, MPCC-SOSC (A2), and strict complementarity condition (A3) yield that the NLP (2.5) satisfies the stationary and second order sufficient conditions. This motivates the idea of approaching the solution to the MPCC (1.1) using nonlinear programming algorithms. The algorithm we employ is a *barrier method*. This method solves a sequence of *barrier problems*,

$$(2.16) \quad \begin{aligned} \min_{x, w, y, s_{cc}} \quad & \varphi_\mu(x, w, y, s_{cc}) \\ \text{s.t.} \quad & h(x, w, y) = 0, \\ & Wy + s_{cc} = te_m, \end{aligned}$$

where $s_{cc} \in \mathbb{R}^m$, $e_m \in \mathbb{R}^m$ is a vector of all ones, $\mu > 0$ is the *barrier parameter*, $t > 0$ relaxes the complementarity constraint, and φ_μ is the barrier objective defined as

$$\varphi_\mu(x, w, y, s_{cc}) := f(x, w, y) - \mu \sum_{i=1}^n \ln(x_i) - \mu \sum_{i=1}^m \ln(w_i) - \mu \sum_{i=1}^m \ln(y_i) - \mu \sum_{i=1}^m \ln(s_{cc,i}).$$

As is evident, we have relaxed the constraint $Wy \leq 0$ as $Wy \leq te_m$. This gives the barrier problem a strictly feasible region which the MPCC does not possess but is necessary for barrier methods. We solve the barrier problem (2.16) for a decreasing sequence of parameters $(\mu, t) \rightarrow 0$.

The first order optimality conditions or KKT conditions for the barrier problem in the *primal-dual* form are as follows:

$$(2.17a) \quad \nabla_x f(z) + \nabla_x h(z)\lambda_h - \lambda_x = 0,$$

$$(2.17b) \quad \nabla_w f(z) + \nabla_w h(z)\lambda_h + Y\lambda_{cc} - \lambda_w = 0,$$

$$(2.17c) \quad \nabla_y f(z) + \nabla_y h(z)\lambda_h + W\lambda_{cc} - \lambda_y = 0,$$

$$(2.17d) \quad h(z) = 0,$$

$$(2.17e) \quad Wy + s_{cc} - te_m = 0,$$

$$(2.17f) \quad S_{cc}\Lambda_{cc}e_m - \mu e_m = 0,$$

$$(2.17g) \quad X\Lambda_x e_n - \mu e_n = 0,$$

$$(2.17h) \quad W\Lambda_w e_m - \mu e_m = 0,$$

$$(2.17i) \quad Y\Lambda_y e_m - \mu e_m = 0.$$

The vectors λ_h and λ_{cc} denote the Lagrange multipliers of the appropriate constraints and λ_x , λ_w , λ_y are multipliers for the bounds on x , w , y , respectively. It is clear that if a limit exists as $(\mu, t) \rightarrow 0$, then we obtain the NLP stationary conditions (2.6) for the MPCC. The stationary conditions (2.17) also provide additional information on the limiting multiplier. Under Assumption (A3), the minimal multiplier in $\mathcal{S}_\lambda$, λ^{basic} (2.7) may not satisfy strict complementarity as

$$(2.18) \quad \begin{aligned} \bar{w}_i &= 0 = \lambda_{w,i}^{\text{basic}} & \forall i \in I_w(\bar{z}) : \bar{\lambda}_{w,i} \leq 0, \\ \bar{y}_i &= 0 = \lambda_{y,i}^{\text{basic}} & \forall i \in I_y(\bar{z}) : \bar{\lambda}_{y,i} \leq 0, \\ \bar{s}_{cc,i} &= 0 = \lambda_{cc,i}^{\text{basic}} & \forall i \in 1, \dots, m : \bar{\lambda}_{w,i} \geq 0 \text{ or } \bar{\lambda}_{y,i} \geq 0. \end{aligned}$$

From the proof of Proposition 2.5, for any $\lambda \in \mathcal{S}_\lambda$ and all $i \in \{1, \dots, m\}$ we have that

$$(2.19) \quad \lambda_{w,i} \geq \lambda_{w,i}^{\text{basic}}, \quad \lambda_{y,i} \geq \lambda_{y,i}^{\text{basic}}, \quad \text{and} \quad \lambda_{cc,i} \geq \lambda_{cc,i}^{\text{basic}}.$$

Equations (2.18) and (2.19) indicate that $\mathcal{S}_\lambda$ includes multipliers that do not satisfy strict complementarity. More significantly, (2.19) shows that there exist multipliers satisfying strict complementarity for the indices in (2.18). In the following we show that convergence to multipliers in $\mathcal{S}_\lambda$ violating strict complementarity can be avoided by bounding the multipliers with indices in (2.18) away from zero.

LEMMA 2.7 (lower bounds for $\lambda_{cc}^j, \lambda_w^j, \lambda_y^j$). *Suppose Assumptions (A) hold. Let $\{(z^j, \lambda^j)\}$ solve the barrier problems (2.17) for some $(t^j, \mu^j) > 0$ with $(t^j, \mu^j) \rightarrow 0$ and $z^j \rightarrow \bar{z}$. Then, for all j sufficiently large,*

$$(2.20) \quad \begin{aligned} \lambda_{cc,i}^j &> \frac{\mu^j}{t^j} & \forall i \in 1, \dots, m, \\ \lambda_{w,i}^j &\geq \frac{\mu^j}{2t^j} \bar{y}_i & \forall i \in I_w(\bar{z}) \setminus I_y(\bar{z}), \\ \lambda_{y,i}^j &\geq \frac{\mu^j}{2t^j} \bar{w}_i & \forall i \in I_y(\bar{z}) \setminus I_w(\bar{z}). \end{aligned}$$

Proof. For $\mu^j > 0$, the solution to the barrier problem satisfies the inequality constraints strictly. Consider $\lambda_{cc,i}^j$,

$$\lambda_{cc,i}^j = \frac{\mu^j}{s_{cc,i}^j} = \frac{\mu^j}{t^j - w_i^j y_i^j} > \frac{\mu^j}{t^j}.$$

As for $\lambda_{w,i}^j$ with $i \in I_w(\bar{z}) \setminus I_y(\bar{z})$,

$$w_i^j = \frac{t^j - s_{cc,i}^j}{y_i^j}.$$

The above is well defined for all j sufficiently large since $\bar{y}_i > 0$ for all $i \in I_w(\bar{z}) \setminus I_y(\bar{z})$ by definition of the index partitioning. This implies that

$$\lambda_{w,i}^j = \frac{\mu^j}{w_i^j} = \frac{\mu^j}{t^j - s_{cc,i}^j} y_i^j > \frac{\mu^j}{t^j} y_i^j \geq \frac{\mu^j}{2t^j} \bar{y}_i,$$

and the claim follows. The remaining claim in (2.20) can be proved similarly. $\square$

From Proposition 2.4 we have that any solution to NLP stationary conditions satisfies the MPCC stationary conditions (2.2). When Assumption (A3) holds, Lemma 2.6 ensures that the limit satisfies a second order sufficient condition, and Lemma 2.7 ensures that the multiplier limit satisfies strict complementarity by appropriate choice of μ and t . This motivates the use of a barrier approach for solving MPCCs.

We present an interior point algorithm for the solution of MPCCs in the next section, which shows quadratic convergence in the neighborhood of a solution.

3. Algorithm. To simplify the notation we will assume without loss of generality that the NLP (2.5) with relaxed complementarity constraints can be posed as follows:

$$(3.1) \quad \begin{aligned} \min_z \quad & f(z) \\ \text{s.t.} \quad & c(z; t) = 0, \\ & z \geq 0, \end{aligned}$$

where $z \in \mathbb{R}^n$, $f : \mathbb{R}^n \rightarrow \mathbb{R}$, and $c(\cdot; t) : \mathbb{R}^n \rightarrow \mathbb{R}^q$. We may identify the variables z and constraints c , respectively, with the variables (x, w, y, s_{cc}) and constraints $(h(x, w, y), Wy + s_{cc} - te_m)$ of the NLP (2.16). We denote the Lagrangian of the above NLP (3.1) as

$$(3.2) \quad \mathcal{L}(z, \lambda) := f(z) + \lambda_c^T c(z; t) - \lambda_z^T z,$$

where $\lambda := (\lambda_c, \lambda_z)$. The algorithm attempts to solve the barrier problem (2.16) approximately and then decrease t and μ . In the following sections we will describe an algorithm, without any enhancement to promote global convergence, which achieves quadratic convergence when started sufficiently close to a solution of the MPCC (1.1). The stationary conditions of a barrier problem associated with NLP in (3.1) are

$$(3.3) \quad \begin{aligned} \nabla_z \mathcal{L}(z, \lambda) + \nabla_z c(z; t) \lambda_c - \lambda_z &= 0, \\ c(z; t) &= 0, \\ Z \Lambda_z e_n &= \mu e_n. \end{aligned}$$

We will assume that the MPCC has a solution $(\bar{z}, \bar{\lambda})$ satisfying Assumptions (A). Let (z^*, λ^*) denote any solution of the NLP where $z^* := (\bar{z}, s_{cc}^*)$, $\lambda^* \in \mathcal{S}_\lambda$ with $s_{cc}^* = 0$. The indices of z are partitioned based on those active at the solution z^* as

$$(3.4) \quad B := \{i \mid z_i^* > 0\} \quad \text{and} \quad N := \{i \mid z_i^* = 0\}.$$

The above partitioning is required only for the analysis and not the algorithm.

Given an initial estimate (z^0, λ^0) with $(z^0, \lambda^0) > 0$, the algorithm generates a sequence of improved estimates (z^k, λ^k) of the solution to the MPCC. For this purpose, in each iteration k the standard primal-dual interior point method solves the linearization of the KKT conditions (3.3),

$$(3.5) \quad \begin{bmatrix} \nabla_{zz}\mathcal{L}(z^k, \lambda^k) & \nabla_z c(z^k; t) & -I \\ \nabla_z c(z^k; t)^T & 0 & 0 \\ \Lambda_z^k & 0 & Z^k \end{bmatrix} \begin{pmatrix} dz^k \\ d\lambda_c^k \\ d\lambda_z^k \end{pmatrix} = - \begin{pmatrix} \nabla_z \mathcal{L}(z^k, \lambda^k) \\ c(z^k; t) \\ Z^k \Lambda_z^k e_n - \mu e_n \end{pmatrix},$$

where $I \in \mathbb{R}^{n \times n}$ is the identity matrix. The Jacobian of the constraints $c(\bar{z}; 0)$ is

$$(3.6) \quad \nabla_z c(\bar{z}; 0)^T = \begin{bmatrix} \nabla_x h(\bar{z})^T & \nabla_w h(\bar{z})^T & \nabla_y h(\bar{z})^T & 0 \\ 0 & \bar{Y} & \bar{W} & I_m \end{bmatrix},$$

where $I_m \in \mathbb{R}^{m \times m}$ is the identity matrix. Under the assumed MPCC-LICQ the full rank of $\nabla_z h(\bar{z})^T$ is guaranteed and the full rank of $c(\bar{z}; 0)$ is immediate. As has been noted earlier, the MPCC (1.1) does not satisfy even the MFCQ at any feasible point. Hence, columns of $\nabla_z c(\bar{z}; 0)$ and gradients of the active bounds at $\bar{z}$ are linearly dependent. Further, strict complementarity can be violated for some $i \in N$; $z_i^* = \lambda_{z,i}^* = 0$. The loss of strict complementarity can result in poor local convergence behavior of the algorithm [24]. Lemma 2.7 shows that convergence to nonstrict complementarity multipliers can be avoided by careful choice of μ and t , provided Assumptions (A3) hold. But singularity of the system (3.5) due to non-satisfaction of MFCQ in the limit requires modification of the step calculation as described below. This modification can be shown to allow for rapid local convergence of the algorithm.

The modification we propose is motivated by, but differs from, recent work on modifications of SQP methods intended to guarantee fast local convergence on degenerate problems [9, 25]. The modification ensures that the KKT matrix remains nonsingular. The step is obtained by solving the linear system

$$(3.7) \quad \begin{bmatrix} \nabla_{zz}\mathcal{L}(z^k, \lambda^k) & \nabla_z c(z^k; t) & -I \\ \nabla_z c(z^k; t)^T & 0 & 0 \\ \tilde{\Lambda}_z^k & 0 & \tilde{Z}^k \end{bmatrix} \begin{pmatrix} dz^k \\ d\lambda_c^k \\ d\lambda_z^k \end{pmatrix} = - \begin{pmatrix} \nabla_z \mathcal{L}(z^k, \lambda^k) \\ c(z^k; t) \\ Z^k \Lambda_z^k e_n - \mu e_n \end{pmatrix},$$

where $\tilde{\Lambda}_z^k$ and $\tilde{Z}^k$ are modifications of the matrices Λ_z^k and Z^k satisfying the following for all (z, λ) sufficiently close to a solution:

- (M1) $\tilde{z}_i, \tilde{\lambda}_{z,i} \geq 0$.
- (M2) $\frac{\tilde{\lambda}_{z,i}}{\tilde{z}_i} = \mathcal{O}(\rho(z, \lambda)) \forall i \in B$ and $\rho(z, \lambda) = \mathcal{O}(\frac{\tilde{z}_i}{\tilde{\lambda}_{z,i}}) \forall i \in N$.
- (M3) $|\tilde{z}_i - z_i| = \mathcal{O}(\rho(z, \lambda)) \forall i \in B$ and $|\tilde{\lambda}_{z,i} - \lambda_{z,i}| = \mathcal{O}(\rho(z, \lambda)) \forall i \in N$.

In the above, $\rho(z, \lambda)$ denotes an estimate of distance of the point $(z, \lambda) > 0$ from the solution set, $\{z^*\} \times \mathcal{S}_\lambda$. A discussion on a suitable estimate follows but we will first focus on the modification.

The following are choices for $\tilde{\Lambda}_z$ and $\tilde{Z}$ that satisfy the conditions (M):

- (i) $\tilde{z}_i = \max(z_i, \eta), \quad \tilde{\lambda}_{z,i} = \max(\lambda_{z,i}, \eta) \quad \forall i \in \{1, \dots, m\},$
- (ii) $\tilde{z}_i = z_i + \eta \lambda_{z,i}, \quad \tilde{\lambda}_{z,i} = \lambda_{z,i} \quad \forall i \in \{1, \dots, m\},$

where η is chosen proportional to the error at the current iterate,

$$(3.9) \quad \eta = \Theta(\rho(z, \lambda)).$$

The first choice, proposed by Vicente and Wright [21], modifies the primal and dual variables while the second modification perturbs only the primal variables. In either case we can show that the local analysis continues to hold true.

We denote the iterate obtained by taking the full step in (3.7) as $(z^{k,\text{full}}, \lambda^{k,\text{full}})$,

$$(3.10) \quad (z^{k,\text{full}}, \lambda^{k,\text{full}}) := (z^k, \lambda^k) + (dz^k, d\lambda^k).$$

The algorithm ensures strict satisfaction of bounds using the projection of the above iterate into the bounds as follows:

$$(3.11) \quad (z^{k+1}, \lambda^{k+1}) := (z^k, \lambda^k) + \tau^k (P(z^{k,\text{full}}, \lambda^{k,\text{full}}) - (z^k, \lambda^k)),$$

where $\tau^k \in (0, 1)$ is usually chosen close to 1 and the projection $P(z, \lambda)$ is given by

$$(3.12) \quad P(z, \lambda) := (\max(z, 0), \lambda_c, \max(\lambda_z, 0)).$$

It is easily shown that

$$z^{k+1} \geq (1 - \tau^k)z^k \quad \text{and} \quad \lambda_z^{k+1} \geq (1 - \tau^k)\lambda_z^k.$$

The parameter τ^k should converge to 1 in the limit to allow for superlinear convergence. We now define an estimate for the distance of (z, λ) from the solution set:

$$(3.13) \quad \rho(z, \lambda) := \|(\nabla_z \mathcal{L}(z, \lambda^{\text{HG}}), \tilde{c}(z), \min(z, \lambda))\|,$$

where

$$(3.14) \quad \lambda_c^{\text{HG}} = \lambda_c, \quad \lambda_{z,i}^{\text{HG}} = \begin{cases} \lambda_{z,i} & \text{if } \lambda_{z,i} \geq z_i \text{ and } \lambda_{z,i} > 0, \\ 0 & \text{otherwise,} \end{cases} \quad i = 1, \dots, n,$$

$$(3.15) \quad \tilde{c}(z) = \begin{bmatrix} h(x, w, y) \\ \min(w, y) \end{bmatrix}.$$

We have redefined the constraint residual calculation for NLP (2.5) so that $\rho(z, \lambda)$ serves as an estimate of the distance to the solution set even when $w_i^* = 0 = y_i^*$. From the above definition of the multiplier λ^{HG} , the following result holds:

$$(3.16) \quad \nabla_z \mathcal{L}(z, \lambda) = \nabla_z \mathcal{L}(z, \lambda^{\text{HG}}) - (\lambda_z - \lambda_z^{\text{HG}}).$$

Hager and Gowda [10] show that $\rho(z, \lambda)$ serves as a lower and an upper bound for the distance from the solution set even when constraint qualifications are not satisfied, but a sufficient second order condition holds and (z, λ) is sufficiently close to the solution set, i.e.,

$$(3.17) \quad \|z - z^*\| + \|\lambda - \hat{\lambda}\| = \Theta(\rho(z, \lambda)),$$

where $\hat{\lambda}$ is the element closest to λ in $\mathcal{S}_\lambda$. Assumptions (A) holding at (z^*, λ^*) ensure that the requirements are met for MPCCs. A summary of the algorithm follows.

ALGORITHM 1.

Choose (z^0, λ^0) with $(z^0, \lambda_z^0) > 0$, tolerance $\text{tol} > 0$, and $\chi_\mu, \chi_t > 0$.
 Calculate $\rho(z^0, \lambda^0)$. Set $k \leftarrow 0$.
while $\rho(z^k, \lambda^k) > \text{tol}$ **do**
 Choose $\eta^k = \Theta(\rho^k)$, $\mu^k = \chi_\mu(\rho^k)^2$, and $t^k = \chi_t \mu^k$.
 Compute $(dz^k, d\lambda^k)$ by solving (3.7).
 Choose τ^k such that $(1 - \tau^k) = \mathcal{O}(\rho(z^k, \lambda^k))$.
 Calculate (z^{k+1}, λ^{k+1}) using (3.11).
 Set $k \leftarrow k + 1$.
end while

4. Local convergence. We analyze the local convergence behavior of Algorithm 1 in this section. The limit point of the sequence of iterates generated by the algorithm will depend greatly on how t and μ are changed. Much of the analysis is along the lines of Vicente and Wright [21]. We establish that the length of the step obtained from the solution of the modified linear system (3.7) is $\mathcal{O}(\rho(z, \lambda))$ in the neighborhood of the solution to the MPCC. Subsequently, we characterize the decrease in the distance to the solution set by taking the step defined in (3.11) and then conclude with the convergence result.

Since our interest is in the local behavior, we will confine our analysis to a neighborhood around a solution point $(z^*, \lambda^*) \in \{\bar{z}\} \times \mathcal{S}_\lambda$,

$$(4.1) \quad \mathcal{N}(\epsilon) = \{(z, \lambda) \mid \|(z, \lambda) - (z^*, \lambda^*)\| \leq \epsilon\}.$$

With a choice of $t = \chi_t \mu$ for some $\chi_t > 0$, we can provide a positive lower bound on the NLP multipliers that could potentially violate strict complementarity ($z_i^* = \lambda_{z,i}^* = 0$) based on Lemma 2.7. Hence, we can make a stronger assumption that we enter a neighborhood $\mathcal{N}_\gamma(\epsilon)$ of a solution (z^*, λ^*) satisfying strict complementarity, i.e.,

$$(4.2) \quad \mathcal{N}_\gamma(\epsilon) := \{(z, \lambda) \mid \|(z, \lambda) - (z^*, \lambda^*)\| \leq \epsilon, \lambda_{z,N} \geq \gamma e_N\},$$

where $\gamma > 0$ is a constant and e_N is a vector of ones of dimension $|N|$. N is the index set partition defined in (3.4). Lemma 2.7 provides an obvious choice of γ ,

$$\gamma \stackrel{(2.20)}{=} \min \left(\frac{1}{2\chi_t} \min \left(1, \bar{w}_{i \in I_y(\bar{z}) \setminus I_w(\bar{z})}, \bar{y}_{i \in I_w(\bar{z}) \setminus I_y(\bar{z})} \right), \bar{\lambda}_{w, i \in I_w(\bar{z}) \cap I_y(\bar{z})}, \bar{\lambda}_{y, i \in I_w(\bar{z}) \cap I_y(\bar{z})}, \bar{\lambda}_{x, i: \bar{x}_i = 0} \right).$$

In what follows we will drop the superscripts on the variables and assume that they represent the current iterates k . Additionally, given a vector $v \in \mathbb{R}^n$, v_N will represent a vector in $\mathbb{R}^{|N|}$ formed from components of v corresponding to components with indices in set $N \subseteq \{1, \dots, n\}$, $v_{i:i \in N}$. In a similar manner, $\nabla_{zz} \mathcal{L}_{BN}$ will represent the submatrix of $\nabla_{zz} \mathcal{L}$ formed from rows and columns of $\nabla_{zz} \mathcal{L}$ with indices, respectively, in the sets B and N (3.4).

We begin by showing that conditions (M) hold for the modifications of Λ_z and Z in (3.8).

LEMMA 4.1. *Suppose Assumptions (A) hold, then $\tilde{z}$ and $\tilde{\lambda}_z$ defined in (3.8) satisfy conditions (M) for all $(z, \lambda) \in \mathcal{N}_\gamma(\epsilon)$ such that $(z, \lambda_z) \geq 0$.*

Proof. Consider the proposed modification (ii) in (3.8). The definition of $\tilde{z}$ and $\tilde{\lambda}_z$ in (ii) of (3.8) implies the satisfaction of (M1) provided $(z, \lambda_z) \geq 0$. To show that (M2) holds, consider the set B . For all $i \in B$ and $\epsilon > 0$ sufficiently small,

$$\begin{aligned} \tilde{\lambda}_{z,i} &\stackrel{(3.8)}{=} \lambda_{z,i} \stackrel{(3.17)}{=} \mathcal{O}(\rho(z, \lambda)) \quad \text{and} \quad \tilde{z}_i \stackrel{(3.8)}{=} z_i + \eta \lambda_{z,i} \geq (z_i^*/2) \\ &\Rightarrow (\tilde{\lambda}_{z,i}/\tilde{z}_i) = \mathcal{O}(\rho(z, \lambda)). \end{aligned}$$

Similarly, using $\eta = \Theta(\rho(z, \lambda))$, we have, for all $i \in N$ and $\epsilon > 0$ sufficiently small and a constant $\beta > 0$,

$$\begin{aligned} (\lambda_{z,i}^*/2) &\leq \lambda_{z,i} \leq 2\lambda_{z,i}^* \quad \text{and} \quad z_i + \eta \lambda_{z,i} \geq \eta \lambda_{z,i} \geq \beta(\lambda_{z,i}^*/2)\rho \\ &\Rightarrow \rho = \mathcal{O}(\tilde{z}_i/\lambda_{z,i}), \end{aligned}$$

which shows that (M2) holds. Satisfaction of (M3) follows from (3.8) and (3.9). The claim for modification (i) can be shown in a similar manner. $\square$

The following lemma bounds the step $(dz, d\lambda)$ obtained by solving the linear system in (3.7).

LEMMA 4.2 (bound on $(dz, d\lambda)$). *Suppose that Assumptions (A) and conditions (M) on the modifications hold and that for some constant $\chi_t > 0$ we have*

$$(4.3) \quad \mu = \mathcal{O}(\rho(z, \lambda)^2) \quad \text{and} \quad t = \chi_t \mu.$$

Then for any $\gamma > 0$ there exist positive constants ϵ and $\kappa_1(\epsilon, \gamma)$ such that

$$(4.4) \quad \|(dz, d\lambda)\| \leq \kappa_1(\epsilon, \gamma)\rho(z, \lambda)$$

for all points $(z, \lambda) \in \mathcal{N}_\gamma(\epsilon)$ with $(z, \lambda_z) > 0$.

Proof. We divide the proof into five sections labeled (A)–(E) and simplify notation by dropping the argument (z, λ) along the lines of Vicente and Wright [21].

(A) *Transforming (3.7) via a singular value decomposition (SVD) of the active constraint Jacobian.* Utilizing the partition of $\{1, \dots, n\} = B \cup N$ we can eliminate $d\lambda_{z,B}$ from (3.7) to obtain

$$(4.5) \quad d\lambda_{z,B} = -\tilde{Z}_B^{-1}\tilde{\Lambda}_{z,B}dz_B - \tilde{Z}_B^{-1}Z_B\Lambda_{z,B}e_B + \mu\tilde{Z}_B^{-1}e_B.$$

Substituting this in (3.7) yields the following reduced system:

$$(4.6) \quad \begin{aligned} &\begin{bmatrix} \nabla_{zz}\mathcal{L}_{BB} + \tilde{Z}_B^{-1}\tilde{\Lambda}_{z,B} & \nabla_{zz}\mathcal{L}_{BN} & \nabla_z c_B & 0 \\ \nabla_{zz}\mathcal{L}_{NB} & \nabla_{zz}\mathcal{L}_{NN} & \nabla_z c_N & -I \\ \nabla_z c_B^T & \nabla_z c_N^T & 0 & 0 \\ 0 & -I & 0 & -\tilde{Z}_N\tilde{\Lambda}_{z,N}^{-1} \end{bmatrix} \begin{pmatrix} dz_B \\ dz_N \\ d\lambda_c \\ d\lambda_{z,N} \end{pmatrix} \\ &= - \begin{pmatrix} \nabla_z \mathcal{L}_B + \tilde{Z}_B^{-1}Z_B\Lambda_{z,B}e_B - \mu\tilde{Z}_B^{-1}e_B \\ \nabla_z \mathcal{L}_N \\ c(z; t) \\ -Z_N\tilde{\Lambda}_{z,N}^{-1}\Lambda_{z,N}e_N + \mu\tilde{\Lambda}_{z,N}^{-1}e_N \end{pmatrix}, \end{aligned}$$

where e_B, e_N denote vectors of all ones with $|B|$ and $|N|$ elements, respectively, and the last block row has been scaled by $-\tilde{\Lambda}_{z,N}^{-1}$.

We now require the SVD of the Jacobian matrix of the active constraints at z^* ,

$$(4.7) \quad \begin{bmatrix} \nabla_z c_B^{*T} & \nabla_z c_N^{*T} \\ 0 & -I \end{bmatrix} = [U \quad V] \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{U}^T \\ \hat{V}^T \end{bmatrix},$$

where $U \in \mathbb{R}^{(q+|N|) \times r}$, $V \in \mathbb{R}^{(q+|N|) \times (q+|N|-r)}$, $S \in \mathbb{R}^{r \times r}$, $\hat{U}^T \in \mathbb{R}^{r \times n}$, $\hat{V}^T \in \mathbb{R}^{(n-r) \times n}$, and r is the rank of the Jacobian matrix. Partitioning rows of V in an obvious way,

$$(4.8) \quad 0 = \begin{bmatrix} V_1^T & V_2^T \end{bmatrix} \begin{bmatrix} \nabla_z c_B^{*T} & \nabla_z c_N^{*T} \\ 0 & -I \end{bmatrix} = \begin{bmatrix} V_1^T \nabla_z c_B^{*T} & (V_1^T \nabla_z c_N^{*T} - V_2^T) \end{bmatrix}.$$

We will apply a change of variables using the orthogonal bases $\hat{U}$, $\hat{V}$ and U , V ,

$$\begin{bmatrix} dz_B \\ dz_N \end{bmatrix} = \hat{U} c_{\hat{U}} + \hat{V} c_{\hat{V}}, \quad \begin{bmatrix} d\lambda_c \\ d\lambda_{z,N} \end{bmatrix} = U c_U + V c_V.$$

Substituting the change of variables in the linear system (4.6) results in

$$(4.9) \quad \begin{bmatrix} \hat{U}^T L \hat{U} & \hat{U}^T L \hat{V} & \hat{U}^T J^T U & \hat{U}^T J^T V \\ \hat{V}^T L \hat{U} & \hat{V}^T L \hat{V} & \hat{V}^T J^T U & \hat{V}^T J^T V \\ U^T J \hat{U} & U^T J \hat{V} & U^T M U & U^T M V \\ V^T J \hat{U} & V^T J \hat{V} & V^T M U & V^T M V \end{bmatrix} \begin{pmatrix} c_{\hat{U}} \\ c_{\hat{V}} \\ c_U \\ c_V \end{pmatrix} = \begin{pmatrix} l_{\hat{U}} \\ l_{\hat{V}} \\ l_U \\ l_V \end{pmatrix},$$

where the matrices L , J , and M are given by

$$(4.10) \quad \begin{aligned} L &= \begin{bmatrix} \nabla_{zz} \mathcal{L}_{BB} + \tilde{Z}_B^{-1} \tilde{\Lambda}_{z,B} & \nabla_{zz} \mathcal{L}_{BN} \\ \nabla_{zz} \mathcal{L}_{NB} & \nabla_{zz} \mathcal{L}_{NN} \end{bmatrix}, & J &= \begin{bmatrix} \nabla_z c_B^T & \nabla_z c_N^T \\ 0 & -I \end{bmatrix}, \\ M &= \begin{bmatrix} 0 & 0 \\ 0 & -\tilde{Z}_N \tilde{\Lambda}_{z,N}^{-1} \end{bmatrix} \end{aligned}$$

and the residuals $l_{\hat{U}}$, $l_{\hat{V}}$, l_U , and l_V are given by

$$(4.11) \quad \begin{aligned} l_{\hat{U}} &= -\hat{U}^T \begin{bmatrix} \nabla_z \mathcal{L}_B + \tilde{Z}_B^{-1} Z_B \Lambda_{z,B} e_B - \mu \tilde{Z}_B^{-1} e_B \\ \nabla_z \mathcal{L}_N \end{bmatrix}, \\ l_{\hat{V}} &= -\hat{V}^T \begin{bmatrix} \nabla_z \mathcal{L}_B + \tilde{Z}_B^{-1} Z_B \Lambda_{z,B} e_B - \mu \tilde{Z}_B^{-1} e_B \\ \nabla_z \mathcal{L}_N \end{bmatrix}, \\ l_U &= -U^T \begin{bmatrix} c(z; t) \\ -Z_N \tilde{\Lambda}_{z,N}^{-1} \Lambda_{z,N} e_N + \mu \tilde{\Lambda}_{z,N}^{-1} e_N \end{bmatrix}, \\ l_V &= -V^T \begin{bmatrix} c(z; t) \\ -Z_N \tilde{\Lambda}_{z,N}^{-1} \Lambda_{z,N} e_N + \mu \tilde{\Lambda}_{z,N}^{-1} e_N \end{bmatrix}. \end{aligned}$$

(B) *Examining $V^T M V$ and its inverse.* We can write the blocks of (4.9) involving M using (4.10) as

$$(4.12) \quad \begin{aligned} \begin{bmatrix} U^T \\ V^T \end{bmatrix} M \begin{bmatrix} U & V \end{bmatrix} &= \begin{bmatrix} U_1^T & U_2^T \\ V_1^T & V_2^T \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & -\tilde{Z}_N \tilde{\Lambda}_{z,N}^{-1} \end{bmatrix} \begin{bmatrix} U_1 & V_1 \\ U_2 & V_2 \end{bmatrix} \\ &= - \begin{bmatrix} U_2^T \tilde{Z}_N \tilde{\Lambda}_{z,N}^{-1} U_2 & U_2^T \tilde{Z}_N \tilde{\Lambda}_{z,N}^{-1} V_2 \\ V_2^T \tilde{Z}_N \tilde{\Lambda}_{z,N}^{-1} U_2 & V_2^T \tilde{Z}_N \tilde{\Lambda}_{z,N}^{-1} V_2 \end{bmatrix} =: \begin{bmatrix} \overline{M}_{11} & \overline{M}_{12} \\ \overline{M}_{21} & \overline{M}_{22} \end{bmatrix}. \end{aligned}$$

From the definition of the $\overline{M}_{22}$ block and the property of $\tilde{\Lambda}_z$ and $\tilde{z}$ in (M2), $\rho = \mathcal{O}(\tilde{Z}_N \tilde{\Lambda}_{z,N}^{-1})$. This implies that the matrix $\tilde{Z}_N \tilde{\Lambda}_{z,N}^{-1}$ is positive diagonal. It is easily shown from the full rank of $\nabla_z c^*$ (3.6) and partitioning of V (4.8) that V_2 has full

column rank [21]. Hence, the eigenvalues of $\overline{M}_{22}$ behave the same way as $\tilde{Z}_N \tilde{\Lambda}_{z,N}^{-1}$ and this yields that

$$(4.13) \quad \overline{M}_{22}^{-1} = \mathcal{O}(\rho^{-1}).$$

(C) *Estimating other blocks in the linear system (4.9).* The contribution of the Jacobian terms may be estimated using the twice Lipschitz continuously differentiable assumption on constraints (A1) as follows:

$$(4.14) \quad \begin{bmatrix} U \\ V \end{bmatrix}^T J \begin{bmatrix} \hat{U} & \hat{V} \end{bmatrix} = \begin{bmatrix} U \\ V \end{bmatrix}^T (J - J^*) \begin{bmatrix} \hat{U} & \hat{V} \end{bmatrix} + \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} \stackrel{(A1),(3.13)}{=} \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} + \mathcal{O}(\rho).$$

The Hessian of the Lagrangian can be similarly bounded,

$$\nabla_{zz} \mathcal{L}(z, \lambda) \stackrel{(A1),(3.17)}{=} \nabla_{zz} \mathcal{L}(z^*, \hat{\lambda}) + \mathcal{O}(\rho),$$

where $\hat{\lambda}$ is the element in $\mathcal{S}_\lambda$ closest to the current iterate λ . Using this,

$$(4.15) \quad \begin{bmatrix} \hat{U}^T L \hat{U} & \hat{U}^T L \hat{V} \\ \hat{V}^T L \hat{U} & \hat{V}^T L \hat{V} \end{bmatrix} \stackrel{(M2),(4.10)}{=} \begin{bmatrix} \hat{U}^T \nabla_{zz} \mathcal{L}(z^*, \hat{\lambda}) \hat{U} & \hat{U}^T \nabla_{zz} \mathcal{L}(z^*, \hat{\lambda}) \hat{V} \\ \hat{V}^T \nabla_{zz} \mathcal{L}(z^*, \hat{\lambda}) \hat{U} & \hat{V}^T \nabla_{zz} \mathcal{L}(z^*, \hat{\lambda}) \hat{V} \end{bmatrix} + \mathcal{O}(\rho).$$

Substituting estimates (4.14), (4.15) into the matrix in (4.9) we obtain

$$(4.16) \quad \begin{bmatrix} \hat{U}^T \nabla_{zz} \mathcal{L}(z^*, \hat{\lambda}) \hat{U} & \hat{U}^T \nabla_{zz} \mathcal{L}(z^*, \hat{\lambda}) \hat{V} & S & 0 \\ \hat{V}^T \nabla_{zz} \mathcal{L}(z^*, \hat{\lambda}) \hat{U} & \hat{V}^T \nabla_{zz} \mathcal{L}(z^*, \hat{\lambda}) \hat{V} & 0 & 0 \\ S & 0 & 0 & 0 \\ 0 & 0 & 0 & \overline{M}_{22} \end{bmatrix} + \begin{bmatrix} \mathcal{O}(\rho) & \mathcal{O}(\rho) & \mathcal{O}(\rho) & \mathcal{O}(\rho) \\ \mathcal{O}(\rho) & \mathcal{O}(\rho) & \mathcal{O}(\rho) & \mathcal{O}(\rho) \\ \mathcal{O}(\rho) & \mathcal{O}(\rho) & \mathcal{O}(\rho) & \mathcal{O}(\rho) \\ \mathcal{O}(\rho) & \mathcal{O}(\rho) & \mathcal{O}(\rho) & 0 \end{bmatrix}.$$

Eliminating c_V from the linear system (4.9) and using the estimate of $\overline{M}_{22}$ from (4.13),

$$(4.17) \quad c_V = \mathcal{O}(\rho^{-1})(l_V + \mathcal{O}(\rho)c_{\hat{U}} + \mathcal{O}(\rho)c_{\hat{V}} + \mathcal{O}(\rho)c_U).$$

Using c_V from (4.17), the linear system (4.9) after reordering reduces to

$$(4.18) \quad \left(\begin{bmatrix} S & \hat{U}^T \nabla_{zz} \mathcal{L}(z^*, \hat{\lambda}) \hat{V} & \hat{U}^T \nabla_{zz} \mathcal{L}(z^*, \hat{\lambda}) \hat{U} \\ 0 & \hat{V}^T \nabla_{zz} \mathcal{L}(z^*, \hat{\lambda}) \hat{V} & \hat{V}^T \nabla_{zz} \mathcal{L}(z^*, \hat{\lambda}) \hat{U} \\ 0 & 0 & S \end{bmatrix} + \mathcal{O}(\rho) \right) \begin{bmatrix} c_U \\ c_{\hat{V}} \\ c_{\hat{U}} \end{bmatrix} = \begin{bmatrix} l_U \\ l_{\hat{V}} \\ l_{\hat{U}} \end{bmatrix} + \mathcal{O}(\|l_V\|).$$

Lemma 2.6 is applied to yield that the matrix $\hat{V}^T \nabla_{zz} \mathcal{L}(z^*, \hat{\lambda}) \hat{V}$ is uniformly nonsingular for all $\hat{\lambda} \in \mathcal{S}_\lambda$ and the matrix in (4.18) is an $\mathcal{O}(\rho)$ perturbation of a nonsingular matrix.

(D) *Estimating the right-hand side components in the linear system (4.9).* From the boundedness of $\mathcal{N}_\gamma(\epsilon)$,

$$(4.19) \quad \begin{aligned} z_i \lambda_{z,i} &= \min(z_i, \lambda_{z,i}) \max(z_i, \lambda_{z,i}) = \mathcal{O}(\min(z_i, \lambda_{z,i})) \\ &\stackrel{(3.13)}{=} \mathcal{O}(\rho) \quad \forall i = 1, \dots, n. \end{aligned}$$

We have that $\nabla_z \mathcal{L}(z, \lambda) = \mathcal{O}(\rho)$ by

$$(4.20) \quad \nabla_z \mathcal{L}(z, \lambda) \stackrel{(3.16)}{=} \nabla_z \mathcal{L}(z, \lambda^{\text{HG}}) - (\lambda_z - \lambda_z^{\text{HG}}) \stackrel{(3.13), (3.14)}{=} \mathcal{O}(\rho).$$

Using these estimates, the bound on $\tilde{z}_B$ from (M3), and $\mu = \mathcal{O}(\rho^2)$ in (4.11),

$$(4.21) \quad l_{\hat{U}} = \mathcal{O}(\rho), \quad l_{\hat{V}} = \mathcal{O}(\rho).$$

For all $i \in N$, $z_N = z_N - z_N^* = \mathcal{O}(\rho)$. We have $c(z; 0) = \mathcal{O}(\rho)$ from (3.13) and $\mu, t = \mathcal{O}(\rho^2)$ from (4.3), yielding

$$(4.22) \quad l_U = \mathcal{O}(\rho).$$

We bound l_V using (3.17), (4.3), (4.8), (A1), and (M3) in (4.11),

$$(4.23) \quad \begin{aligned} l_V &\stackrel{(4.11)}{=} -V^T \begin{bmatrix} c(z; t) \\ -Z_N \tilde{\Lambda}_{z,N}^{-1} \Lambda_{z,N} e_N + \mu \tilde{\Lambda}_{z,N}^{-1} e_N \end{bmatrix} \\ &\stackrel{(A1)}{=} -V^T \begin{bmatrix} c(z^*; 0) + \nabla_z c(z^*)^T (z - z^*) + \mathcal{O}(\|z - z^*\|^2) + \Theta(t) \\ -(Z e - Z^* e)_N + \mu \tilde{\Lambda}_{z,N}^{-1} e_N + Z_N \tilde{\Lambda}_{z,N}^{-1} (\tilde{\Lambda}_{z,N} - \Lambda_{z,N}) e_N \end{bmatrix} \\ &\stackrel{(M3)}{=} -V^T \begin{bmatrix} \nabla_z c_B^{*T} & \nabla_z c_N^{*T} \\ 0 & -I \end{bmatrix} \begin{bmatrix} z_B - z_B^* \\ z_N - z_N^* \end{bmatrix} - V^T \begin{bmatrix} \mathcal{O}(\|z - z^*\|^2) + \Theta(t) \\ \mu \tilde{\Lambda}_{z,N}^{-1} e_N + \mathcal{O}(\rho^2) \end{bmatrix} \\ &\stackrel{(4.8)}{=} -V^T \begin{bmatrix} \mathcal{O}(\|z - z^*\|^2) + \Theta(t) \\ \mu \tilde{\Lambda}_{z,N}^{-1} e_N + \mathcal{O}(\rho^2) \end{bmatrix} \stackrel{(3.17), (4.3)}{=} \mathcal{O}(\rho^2). \end{aligned}$$

(E) *Estimating the size of the components c_U , c_V , $c_{\hat{U}}$, and $c_{\hat{V}}$.* From uniform nonsingularity of the matrix in (4.18) for ρ sufficiently small,

$$c_U, c_{\hat{U}}, c_{\hat{V}} \stackrel{(4.21), (4.22)}{=} \mathcal{O}(\rho).$$

The bound on c_V is obtained by using (4.23) and the above result in (4.17),

$$c_V = \mathcal{O}(\rho).$$

Since the step $(dz, d\lambda_c, d\lambda_{z,N})$ is an orthogonal transformation of $(c_{\hat{U}}, c_{\hat{V}}, c_U, c_V)$,

$$\|(dz, d\lambda_c, d\lambda_{z,N})\| \stackrel{(3.9)}{=} \mathcal{O}(\rho).$$

To complete the proof, we provide the estimate on $d\lambda_{z,B}$ using (3.9), (4.3), and (M3),

$$\begin{aligned} d\lambda_{z,B} &= -\tilde{Z}_B^{-1} \tilde{\Lambda}_{z,B} dz_B - \tilde{Z}_B^{-1} Z_B \Lambda_{z,B} e_B + \mu \tilde{Z}_B^{-1} e_B \\ &= \mathcal{O}(\rho) \Rightarrow \|(dz, d\lambda)\| = \mathcal{O}(\rho). \quad \square \end{aligned}$$

The following result states the progress resulting from a full step.

LEMMA 4.3 (bound on $\rho(z^{\text{full}}, \lambda^{\text{full}})$). *Suppose that Assumptions (A) and conditions (M) on the modifications hold and μ, t are given by (4.3). Then given any $\gamma > 0$, there exist positive constants ϵ and $\kappa_2(\epsilon, \gamma)$ such that*

$$(4.24) \quad \rho(z^{\text{full}}, \lambda^{\text{full}}) \leq \kappa_2(\epsilon, \gamma) \rho(z, \lambda)^2$$

for all $(z, \lambda) \in \mathcal{N}_\gamma(\epsilon)$ with $(z, \lambda_z) \geq 0$.

Proof. To establish the bound on $\rho(z^{\text{full}}, \lambda^{\text{full}})$, we will bound each of the components in the error estimate (3.13). We will begin by establishing bounds on z^{full} and λ_z^{full} . Consider dz_i for $i \in N$,

$$dz_N \stackrel{(3.7)}{=} -Z_N \tilde{\Lambda}_{z,N}^{-1} \Lambda_{z,N} e_N - Z_N \tilde{\Lambda}_{z,N}^{-1} d\lambda_{z,N} + \mu \tilde{\Lambda}_{z,N}^{-1} e_N,$$

and hence from (M3), (4.3), (4.4), and $z_N = \mathcal{O}(\rho)$,

$$(4.25) \quad \|z_N + dz_N\| \leq \|Z_N \tilde{\Lambda}_{z,N}^{-1} (\tilde{\Lambda}_{z,N} - \Lambda_{z,N}) e_N\| + \|\tilde{\Lambda}_{z,N}^{-1} Z_N d\lambda_{z,N}\| + \mu \|\tilde{\Lambda}_{z,N}^{-1} e_N\| \\ = \mathcal{O}(\rho^2).$$

Consider $d\lambda_{z,B}$,

$$d\lambda_{z,B} = -\tilde{Z}_B^{-1} Z_B \Lambda_{z,B} e_B - \tilde{Z}_B^{-1} \tilde{\Lambda}_{z,B} dz_B + \mu \tilde{Z}_B^{-1} e_B,$$

and hence from (M3), (4.3), (4.4), and $\lambda_{z,B} = \mathcal{O}(\rho)$,

$$(4.26) \quad \|\lambda_{z,B} + d\lambda_{z,B}\| \leq \|\tilde{Z}_B^{-1} (\tilde{Z}_B - Z_B) \Lambda_{z,B} e_B\| + \|\tilde{Z}_B^{-1} \tilde{\Lambda}_{z,B} dz_B\| + \mu \|\tilde{Z}_B^{-1} e_B\| \\ = \mathcal{O}(\rho^2).$$

The bounds on z_B^{full} and $\lambda_{z,N}^{\text{full}}$ may be derived for all $\epsilon > 0$ sufficiently small, as

$$(4.27) \quad z_B^{\text{full}} = z_B + dz_B \stackrel{(4.4)}{=} z_B + \mathcal{O}(\rho) \stackrel{(3.4)}{\geq} (1/2)z_B^* > 0, \\ \lambda_{z,N}^{\text{full}} = \lambda_{z,N} + d\lambda_{z,N} \stackrel{(4.4)}{=} \lambda_{z,N} + \mathcal{O}(\rho) \stackrel{(3.4)}{\geq} (\gamma/2)e_N > 0.$$

Consider each of the components of ρ starting with the last component in (3.13). The bound is established below:

$$(4.28) \quad |\min(z_i^{\text{full}}, \lambda_{z,i}^{\text{full}})| = \begin{cases} |\lambda_{z,i}^{\text{full}}| \stackrel{(4.26), (4.27)}{=} \mathcal{O}(\rho^2) & \forall i \in B, \\ |z_i^{\text{full}}| \stackrel{(4.25), (4.27)}{=} \mathcal{O}(\rho^2) & \forall i \in N. \end{cases}$$

We now derive a bound on $\nabla_z \mathcal{L}(z^{\text{full}}, \lambda^{\text{full}, \text{HG}})$. Expanding $\nabla_z \mathcal{L}(z^{\text{full}}, \lambda^{\text{full}})$ using Taylor's series and substituting in the first row of (3.7), we obtain

$$(4.29) \quad \begin{aligned} \nabla_z \mathcal{L}(z^{\text{full}}, \lambda^{\text{full}}) &= \nabla_z \mathcal{L}(z + dz, \lambda + d\lambda) \\ &\stackrel{(A1)}{=} \nabla_z \mathcal{L}(z, \lambda + d\lambda) + \nabla_{zz} f(z) dz + \mathcal{O}(\|dz\|^2) \\ &\quad + \sum_{i=1}^q (\lambda_{c,i} + d\lambda_{c,i}) \nabla_{zz} c_i(z) dz + \mathcal{O}(\|dz\|^2 \|\lambda + d\lambda\|) \\ &= \left(\nabla_z \mathcal{L}(z, \lambda) + \left(\nabla_{zz} f(z) + \sum_{i=1}^q \lambda_{c,i} \nabla_{zz} c_i(z) \right) dz \right. \\ &\quad \left. + \nabla_z c(z) d\lambda_c - d\lambda_z \right) + \sum_{i=1}^q d\lambda_{c,i} \nabla_{zz} c_i(z) dz \\ &\quad + \mathcal{O}(\|dz\|^2) + \mathcal{O}(\|dz\|^2 \|\lambda + d\lambda\|) \\ &\stackrel{(3.7)}{=} \sum_{i=1}^q d\lambda_{c,i} \nabla_{zz} c_i(z) dz + \mathcal{O}(\|dz\|^2) + \mathcal{O}(\|dz\|^2 \|\lambda + d\lambda\|) \\ &\stackrel{(A1)}{=} \mathcal{O}(\|d\lambda\| \|dz\|) + \mathcal{O}(\|dz\|^2) + \mathcal{O}(\|dz\|^2 \|\lambda + d\lambda\|) \\ &\stackrel{(4.4)}{=} \mathcal{O}(\rho^2). \end{aligned}$$

From the definition of λ^{HG} (3.14) and $(z^{\text{full}}, \lambda_z^{\text{full}}) > 0$, we have that

$$\begin{aligned} \nabla_z \mathcal{L}(z^{\text{full}}, \lambda^{\text{full, HG}}) &\stackrel{(3.16)}{=} \nabla_z \mathcal{L}(z^{\text{full}}, \lambda^{\text{full}}) + (\lambda_z^{\text{full}} - \lambda_z^{\text{full, HG}}) \\ &\stackrel{(4.29)}{=} \mathcal{O}(\rho^2) + \mathcal{O}(\min(z^{\text{full}}, \lambda_z^{\text{full}})) \stackrel{(4.28)}{=} \mathcal{O}(\rho^2). \end{aligned}$$

Similarly, for $c(z + dz; 0)$ we have that

$$\begin{aligned} c(z + dz; 0) &= c(z; 0) + \nabla_z c(z)^T dz + \mathcal{O}(\|dz\|^2) \\ &\stackrel{(3.7)}{=} c(z; 0) - c(z; t) + \mathcal{O}(\|dz\|^2) \\ &= \mathcal{O}(t) + \mathcal{O}(\|dz\|^2) \\ &\stackrel{(4.3), (4.4)}{=} \mathcal{O}(\rho^2). \end{aligned} \tag{4.30}$$

The bounds on $\min(w^{\text{full}}, y^{\text{full}})$ may be deduced as

$$\| \min(w^{\text{full}}, y^{\text{full}}) \| \leq \| z_N^{\text{full}} \| \stackrel{(4.25)}{=} \mathcal{O}(\rho^2). \tag{4.31}$$

Collecting results in (4.30) and (4.31), $\tilde{c}(z^{\text{full}}) = \mathcal{O}(\rho^2)$. From the above results we have that $\rho(z^{\text{full}}, \lambda^{\text{full}}) = \mathcal{O}(\rho^2)$. $\square$

We now show a similar result to hold for the new iterate (z^+, λ^+) .

LEMMA 4.4 (bound on $\rho(z^+, \lambda^+)$). *Suppose that Assumptions (A) and conditions (M) on the modifications hold, with μ, t given by (4.3) and τ satisfying*

$$1 - \tau \leq \zeta \rho \tag{4.32}$$

for some $\zeta > 0$. Then, given $\gamma, \zeta > 0$, there exist positive constants ϵ and $\kappa_3(\epsilon, \gamma, \zeta)$ such that

$$\rho(z^+, \lambda^+) \leq \kappa_3(\epsilon, \gamma, \zeta) \rho(z, \lambda)^2 \tag{4.33}$$

for all $(z, \lambda) \in \mathcal{N}_\gamma(\epsilon)$ with $(z, \lambda_z) \geq 0$.

Proof. We begin by providing a bound on the difference between $(z^{\text{full}}, \lambda^{\text{full}})$ and its projection onto the bounds, $P(z^{\text{full}}, \lambda^{\text{full}})$. The projection onto the bounds alters only those negative components of $(z^{\text{full}}, \lambda^{\text{full}})$ that have bounds. Hence,

$$\begin{aligned} \|P(z^{\text{full}}, \lambda^{\text{full}}) - (z^{\text{full}}, \lambda^{\text{full}})\| &= \|(\max(z^{\text{full}}, 0), \max(\lambda_z^{\text{full}}, 0)) - (z^{\text{full}}, \lambda_z^{\text{full}})\| \\ &\stackrel{(4.27)}{=} \|(\max(z_N^{\text{full}}, 0) - z_N^{\text{full}}, \max(\lambda_{z,B}^{\text{full}}, 0) - \lambda_{z,B}^{\text{full}})\| \\ &\leq \|z_N^{\text{full}}\| + \|\lambda_{z,B}^{\text{full}}\| \stackrel{(4.25), (4.26)}{=} \mathcal{O}(\rho^2). \end{aligned}$$

The distance between (z^+, λ^+) and $(z^{\text{full}}, \lambda^{\text{full}})$ can be bounded using the above as

$$\begin{aligned} \|(z^+, \lambda^+) - (z^{\text{full}}, \lambda^{\text{full}})\| &\stackrel{(3.12)}{=} \|(1 - \tau)((z, \lambda) - P(z^{\text{full}}, \lambda^{\text{full}})) \\ &\quad + (P(z^{\text{full}}, \lambda^{\text{full}}) - (z^{\text{full}}, \lambda^{\text{full}}))\| \\ &\leq |1 - \tau| \|(dz, d\lambda)\| + \mathcal{O}(\rho^2) \stackrel{(4.4), (4.32)}{=} \mathcal{O}(\rho^2). \end{aligned} \tag{4.34}$$

This sets the stage for bounding the distance of (z^+, λ^+) from the solution set:

$$\begin{aligned} \|(z^+, \lambda^+) - (z^*, \widehat{\lambda^+})\| &\leq \|(z^+, \lambda^+) - (z^{\text{full}}, \lambda^{\text{full}})\| + \|(z^{\text{full}}, \lambda^{\text{full}}) - (z^*, \widehat{\lambda^{\text{full}}})\| \\ &\quad + \|(z^*, \widehat{\lambda^{\text{full}}}) - (z^*, \widehat{\lambda^+})\| \\ &\leq \|(z^+, \lambda^+) - (z^{\text{full}}, \lambda^{\text{full}})\| + \|(z^{\text{full}}, \lambda^{\text{full}}) - (z^*, \widehat{\lambda^{\text{full}}})\| \\ &\quad + \|\widehat{\lambda^{\text{full}}} - \widehat{\lambda^+}\| \\ &\stackrel{(4.34)}{\leq} \mathcal{O}(\rho^2) + \mathcal{O}(\rho(z^{\text{full}}, \lambda^{\text{full}})) + \|\lambda^{\text{full}} - \lambda^+\| \\ &\stackrel{(4.24)}{\leq} \mathcal{O}(\rho^2) + \|\lambda^{\text{full}} - \lambda^+\| \stackrel{(4.34)}{=} \mathcal{O}(\rho^2), \end{aligned}$$

where $\widehat{\lambda}^+$, $\widehat{\lambda}^{\text{full}}$, respectively, refer to the projection of λ^+ , λ^{full} onto $\mathcal{S}_\lambda$. The penultimate inequality follows from the nonexpansive property of the projection onto the set of NLP multipliers, $\mathcal{S}_\lambda$. The above result with (3.17) shows the claim. $\square$

From Lemmas 4.2 and 4.4 we have that, respectively, the steps from Algorithm 1 at each iteration are bounded by ρ and furthermore make excellent progress toward the solution. Using these lemmas we are in a position to state the convergence result. We omit the proof of the following result for brevity, as it follows from Theorem 4.3 of Vicente and Wright [21] and the compactness of $\mathcal{N}_\gamma(\epsilon)$.

THEOREM 4.5 (quadratic convergence). *Suppose that Assumptions (A) and conditions (M) on the modifications hold and μ , t are given by (4.3). Let $\gamma, \zeta > 0$ be given and consider a value of ϵ such that Lemmas 4.2 and 4.4 are applicable. If the initial point is $(z^0, \lambda^0) \in \mathcal{N}_{2\gamma}(\epsilon/2)$ and*

$$\rho(z^0, \lambda^0) \leq \min \left\{ \frac{1}{2\kappa_3(\epsilon, \gamma, \zeta)}, \frac{\epsilon}{4\kappa_1(\epsilon, \gamma)}, \frac{\gamma}{2\kappa_1(\epsilon, \gamma)} \right\},$$

then the iterates (z, λ) generated by Algorithm 1 remain inside the neighborhood $\mathcal{N}_\gamma(\epsilon)$ and converge q -quadratically to a point (z^*, λ^{**}) where $\lambda^{**} \in \mathcal{S}_\lambda$.

5. Numerical experience. Our motivation for the numerical study stems from a desire to understand the practical performance of the proposed interior point algorithm, Algorithm 1. In order to assume that the iterates eventually enter the neighborhood $\mathcal{N}_\gamma$ we need to bound μ/t away from zero but we cannot guarantee bounded multipliers. It will be interesting to see whether we converge to bounded multipliers in this case. We provide a brief description of the implementation of the algorithm and also provide a comparison with two other algorithms, LOQO [2] and filtermpec [6].

5.1. Implementation with IPOPT. The algorithm proposed in section 3 can be easily incorporated within most existing interior point NLP algorithms. Here we present preliminary numerical experience by incorporating the above modification within such an NLP algorithm, IPOPT [22, 23]. The basic steps of the algorithm with the proposed step modification given an NLP, such as in (3.1), are provided in Algorithm IPOPT-C.

IPOPT [22, 23] is an interior point NLP algorithm employing a filter to promote global convergence from initial iterates (z^0, λ^0) that are far removed from a solution to the problem. A search direction is obtained by solving a linear system, such as (3.7). The step size, $\alpha^{\max}$ is calculated so that the resulting iterate is within the bounds, i.e.,

$$z + \alpha^{\max} dz > 0 \quad \text{and} \quad \lambda_z + \alpha^{\max} d\lambda_z > 0.$$

We have employed this instead of the projection step (3.11), as global convergence cannot be guaranteed in the latter case for the current implementation in IPOPT. Our experience has been that in this case full steps are taken in the limit, i.e.,

$$(1 - \alpha^{\max}) = \mathcal{O}(\rho).$$

Using the above, quadratic convergence of the implementation within IPOPT can be shown using the analysis in section 4. A backtracking line search is performed to obtain a step size, α^k , and an iterate that achieves sufficient decrease in either the objective function or the constraint violation. Further information on the filter can be obtained from Wächter [22] and Wächter and Biegler [23]. In certain iterations,

ALGORITHM IPOPT-C.

Choose (z^0, λ^0) with $(z^0, \lambda^0) > 0$, tolerance $\text{tol} > 0$. Set $l \leftarrow 0$. Choose $\tau := 0.99$, $\mu^0, \mu_{\text{thresh}}, \delta_\mu, \chi_t, \chi_\mu > 0$, $\nu_\mu \geq 1$, and $\nu_\epsilon \in (0, 1]$ such that $\nu_\epsilon \geq (1/\nu_\mu)$. Set $t^l \leftarrow \chi_t \mu^l$.

while $\left\| \begin{bmatrix} \nabla_z \mathcal{L}(z^l, \lambda^l) \\ c(z^l; 0) \\ Z^l \lambda_z^l \end{bmatrix} \right\|_\infty > \text{tol}$ **do**

Set $k \leftarrow 0$, $(z^{l,k}, \lambda^{l,k}) \leftarrow (z^l, \lambda^l)$.

repeat

if $\mu^l \leq \mu_{\text{thresh}}$ **then**

Set $\eta^k \leftarrow 0.1 \frac{\mu^{l-1}}{1 + \|\lambda^{l,k}\|_\infty}$.

else

Set $\eta^k \leftarrow 0$.

end if

Compute $(dz^k, d\lambda^k)$ by solving (3.7) with $(z^{l,k}, \lambda^{l,k}, t^l, \mu^l)$ replacing (z^k, λ^k, t, μ) and $\tilde{z}_i^{l,k}, \tilde{\lambda}_{z,i}^{l,k}$ defined by (ii) in (3.8) and (3.7).

Calculate $\alpha^{\max} := \max\{\alpha \in (0, 1] \mid \alpha dz^k \geq -\tau z^{l,k}, \alpha d\lambda_z^{l,k} \geq -\tau \lambda_z^{l,k}\}$.

Perform backtracking line search to obtain $\alpha^k \in (0, \alpha^{\max}]$ such that

$(z^{l,k+1}, \lambda^{l,k+1}) \leftarrow (z^{l,k}, \lambda^{l,k}) + \alpha^k (dz^k, d\lambda^k)$ achieves progress in solving the barrier problem of (3.1) for (t^l, μ^l) as gauged by a filter [22, 23].

if line search fails **then**

Invoke a restoration phase to find $(z^{l,k+1}, \lambda^{l,k+1})$ that decreases infeasibility [22, 23].

end if

Set $k \leftarrow k + 1$.

until $\left\| \begin{bmatrix} \nabla_z \mathcal{L}(z^{l,k}, \lambda^{l,k}) \\ c(z^{l,k}; t^l) \\ Z^{l,k} \lambda_z^{l,k} - \mu^l e_n \end{bmatrix} \right\|_\infty \leq \delta_\mu (\mu^l)^{\nu_\epsilon}$

Set $(z^{l+1}, \lambda^{l+1}) \leftarrow (z^{l,k}, \lambda^{l,k})$, $\mu^{l+1} \leftarrow \chi_\mu (\mu^l)^{\nu_\mu}$, $t^{l+1} \leftarrow \chi_t \mu^{l+1}$, $\tau = \max(\tau, 1 - \mu^{l+1})$ and $l \leftarrow l + 1$.

end while

even sufficiently close to a solution, we can fail to find a step size, $\alpha^k > \alpha^{\min}$, that accomplishes this. IPOPT then invokes a restoration phase to locate an iterate that has lower constraint infeasibility. We will not elaborate further on the particular restoration phase algorithm but emphasize that successful termination of the restoration phase algorithm is crucial to the convergence of the algorithm [22, 23]. In our experience, as noted below, there have been instances in which convergence has been impeded due to failure of the restoration phase.

In the algorithm IPOPT-C, a choice of $\nu_\mu > 1$ may be used to reduce the barrier parameter at a superlinear rate as required by our analysis in the previous section. The analysis in the previous section applies only in the neighborhood of the solution and we use μ_{thresh} as a threshold for μ below which the step is modified. We obtain the original algorithm IPOPT, with no modification of the step, if we set $\eta^k \leftarrow 0$ for all k . For the modification, we have employed (ii) in (3.8). Our experience with other choices for the modification has been similar. In checking for the error in satisfying the NLP stationary conditions for the MPCC we have used $c(\cdot; 0)$ instead of $\tilde{c}(\cdot)$, as these are equivalent when the upper level strict complementarity condition holds for the MPCC (A3). The modification η^k has been made proportional to the barrier

parameter for the previous subproblem instead of the error ρ , and this has worked well in practice. Also, note that the algorithm solves each barrier problem to a tolerance of $\delta_\mu(\mu^l)^{\nu_\epsilon}$ for $\nu_\epsilon < 1$, whenever $\nu_\mu > 1$, instead of $\nu_\epsilon = 1$ as is generally used in NLP interior point methods. This is due to the fact that the algorithm yields an iterate with an error on the order of μ^l , and the choice of $\nu_\epsilon < 1$ serves as a safeguard.

5.2. AMPL interface. The algorithm IPOPT has been interfaced to AMPL [8], a modeling language. AMPL provides for easy specification of the complementarity constraint using the *complements* operator [8]. The canonical MPCC that AMPL conveys to the solver is in the following form:

$$\begin{aligned} \min_{x,y} \quad & f(x,y) \\ \text{s.t.} \quad & g(x,y) \leq 0, \\ & h(x,y) = 0, \\ & l_{\tilde{g},i} \leq \tilde{g}_i(x,y) \leq u_{\tilde{g},i} \perp l_{y,i} \leq y_i \leq u_{y,i}, \quad i = 1, \dots, m, \end{aligned}$$

where $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$, $g : \mathbb{R}^{n+m} \rightarrow \mathbb{R}^p$, $\tilde{g} : \mathbb{R}^{n+m} \rightarrow \mathbb{R}^m$, $h : \mathbb{R}^{n+m} \rightarrow \mathbb{R}^q$. In the above program, for each i , exactly one of $\{l_{\tilde{g},i}, u_{\tilde{g},i}\}$ and one of $\{l_{y,i}, u_{y,i}\}$ is finite. The interface will reformulate by addition of slack variables, $s_g \in \mathbb{R}^p$, $w \in \mathbb{R}^m$ for g , $\tilde{g}$, respectively, and $s_{cc} \in \mathbb{R}^m$ for the inequality relaxation of the complementarity constraints to yield the NLP,

$$\begin{aligned} \min_{x,w,y,s_g,s_{cc}} \quad & f(x,y) \\ \text{s.t.} \quad & g(x,y) + s_g = 0, \\ & h(x,y) = 0, \\ & \tilde{g}(x,y) - w = 0, \\ & (w_i - l_{\tilde{g},i})(y_i - l_{y,i}) + s_{cc,i} = t, \quad i = 1, \dots, m, \\ & w_i \geq l_{\tilde{g},i}, y_i \geq l_{y,i}, \quad i = 1, \dots, m, \\ & s_g, s_{cc} \geq 0 \end{aligned}$$

for some $t > 0$ approaching zero. A similar reformulation is easily done when we have a few or all of upper bounds $u_{\tilde{g},i}$, $u_{y,i}$ being finite instead of the lower bounds. Due to the reformulation, IPOPT-C sees a problem with $(n + 3m + p)$ variables, $(p + q + 2m)$ constraints including m complementarity constraints plus $(p + 3m)$ bounds. The AMPL interface of IPOPT-C conveys the appropriate function and derivative information to the solver.

5.3. IPOPTC: Results and discussion. We present results from solving 136 problems of the MacMPEC test suite [12] for different μ/t ratios. The suite contains 140 problems, of which we have excluded **ex9.2.2**, **qpec2**, **ralph1**, and **scholtes4**, as these problems do not possess a strongly stationary point. In the full set, there are a number of problems that do not satisfy strict complementarity or the MPCC-SOSC at the solution, and so we opt for a conservative choice of the constants in the algorithm IPOPT-C. In particular we choose the constants as $\mu^0 = 0.1$, $\chi_\mu = 1$, $\nu_\mu = 1.3$, $\delta_\mu = 5$, $\mu_{\text{thresh}} = 5 \cdot 10^{-6}$, $\text{tol} = 5 \cdot 10^{-7}$. The analysis of Wright [26] on the effect of machine precision for interior point methods can be readily shown to hold for MPCCs as well. Wright [26] suggests terminating the interior point algorithm at $\mu > \sqrt{\mathbf{u}} = 10^{-8}$, where $\mathbf{u}$ is the floating point machine precision. Due to this and the nonsatisfaction of Assumptions (A) on some of the MacMPEC problems, we have chosen a higher value ($\text{tol} = 5 \cdot 10^{-7} = 50 \mathbf{u}$) for our termination criterion. The initial points for all the problems are those provided in the AMPL model. The slacks for the

TABLE 5.1
Number of MacMPEC problems solved by different μ/t ratios.

$\frac{\mu}{t}$	# solved
10	129
1	130
0.1	128
$\mu^{0.1}$	126

complementarity constraints are set to 0.1 while those for the inequalities are set to the evaluation of the inequality constraints. The initial point may be perturbed by IPOPT-C to ensure strict satisfaction of the bounds [22].

We compare the performance on MacMPEC for different μ/t ratios. The number of problems solved by each of the four ratios that we compare are provided in Table 5.1. We do not report the times for each problem but just mention that all the options completed the entire set of 136 problems within 45 CPU minutes on a 2.4 GHz Intel Xeon processor with LINUX as the operating system. The appendix provides complete information on problem size and computational requirements for solving the MacMPEC problems for the choice of $\chi_t = 1$.

We provide a comparison of the different options using the Dolan and Moré [4] performance profiles. Let $t_{pr,s}$ denote a performance characteristic, such as function evaluations required, of an option s in solving a problem pr . Define the ratios

$$r_{pr,s} = \frac{t_{pr,s}}{\min_{s'} t_{pr,s'}}.$$

Then the performance profile for each option s is defined as

$$\pi_s(\tau) = \frac{\# \text{ of problems with } r_{pr,s} \leq \tau}{\text{total } \# \text{ of problems}}, \quad \tau \geq 0.$$

This performance profile will be used to compare the magnitudes of the multipliers at the solution, function evaluation, and iteration count for different options.

Figures 5.1, 5.2, and 5.3, respectively, plot the performance profiles in terms of magnitude of multipliers, iterations, and function evaluations for different ratios of μ/t . We have considered only 111 problems on which all options converged to a solution with same objective value. The legend in the plots indicates the ratio of μ/t .

Decreasing μ/t may result in convergence to smaller multipliers as indicated by Lemma 2.7. Figure 5.1 confirms this observation. We obtain the smallest multipliers when $\mu/t = 0.1$. On the other hand, smaller multipliers imply that the limiting multipliers are close to those in $\mathcal{S}_\lambda$ that do not satisfy strict complementarity. This leads to increased function evaluations and iterations; refer to Figures 5.2 and 5.3. It is interesting that even in the case of $t = \mu^{0.9}$, for which the assumption of iterates entering the neighborhood $\mathcal{N}_\gamma$ may fail to hold, we can solve most of the problems. The results indicate that the choice of $t = \chi_t \mu$ with $\chi_t \geq 1$ should be fairly robust.

5.4. Discussion on IPOPT-C failures. The algorithm can fail to converge to a solution due to one or more of the following reasons.

1. *Failure of Assumption (A2) or (A3):* This can occur if the limit point does not satisfy strong stationarity, MPCC-LICQ, MPCC-SOSC, or strict complementarity. This often results in unbounded multipliers and breakdown of the algorithm.

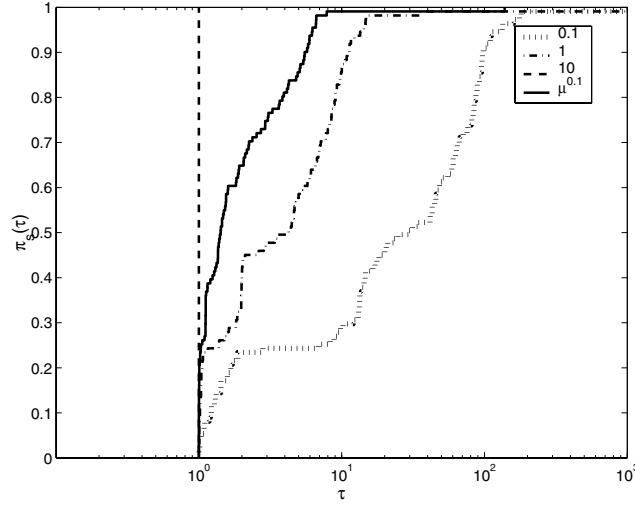


FIG. 5.1. Performance of IPOPT-C in terms of $\|(\lambda_w, \lambda_y, \lambda_{cc})\|_1$ for different μ/t on the MacMPEC suite.

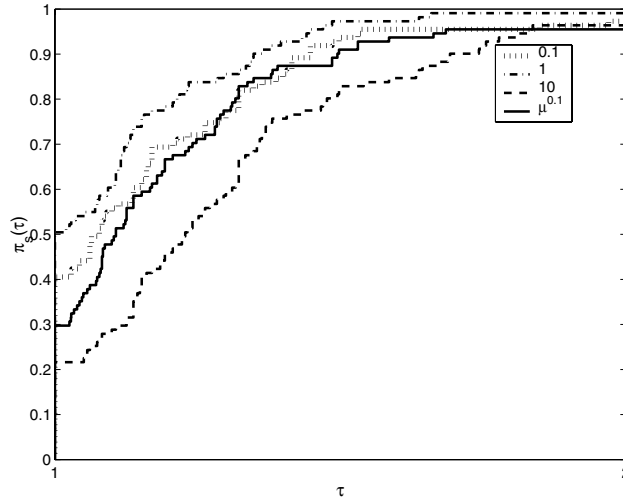


FIG. 5.2. Performance of IPOPT-C in terms of iterations for different μ/t on the MacMPEC suite.

2. *Failure of restoration phase:* As mentioned previously, IPOPT employs a filter with a backtracking line search procedure to aid convergence from poor initial points. If the algorithm does not make sufficient progress, then it invokes a restoration phase algorithm which attempts to locate a point with lower infeasibility. Restoration can be invoked arbitrarily close to the solution and if unsuccessful can prevent convergence of the algorithm.

The options presented in Table 5.1 all fail on **siouxfls1**, **water-FL**, and **water-net**. On these problems, the algorithm terminates without solving even the initial few barrier problems, attributable to failure mode (2).

The failure of Assumption (A2) does not automatically imply failure of the algorithm. The nonsatisfaction of MPCC-LICQ or strict complementarity is not as severe

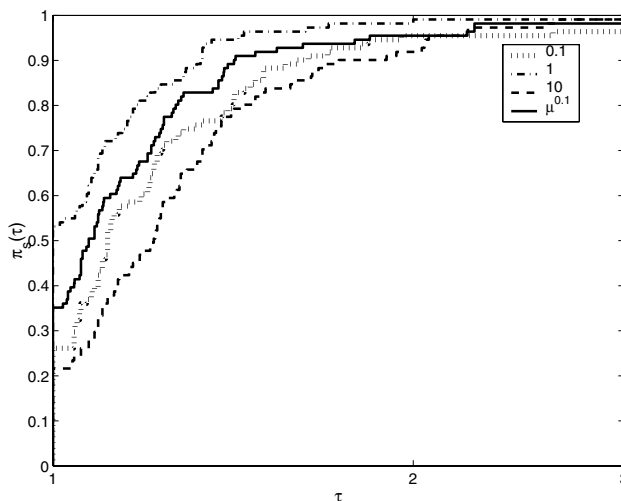


FIG. 5.3. Performance of IPOPT-C in terms of function evaluations for different μ/t on the MacMPEC suite.

as failure of strong stationarity. We have observed convergence, albeit at a linear rate, even when strict complementarity does not hold, for example, **bard1**, **bard3**, etc. In some cases there is a combined effect of failure of strict complementarity and an unsuccessful restoration phase that have caused failure. The analysis of the algorithm in section 4 can be readily applied for the case where MPCC-MFCQ holds in place of MPCC-LICQ provided that we generate iterates that enter $\mathcal{N}_\gamma$. We have observed this on **scholtes5** for which only the MPCC-MFCQ holds.

Convergence of algorithms depends critically on both effective modeling as well as algorithmic robustness. The preceding paragraph earmarked the scope for improvement of the algorithm. However, there are certain problem formulations in the MacMPEC test suite for which interior point methods do not seem a natural choice. In certain cases, there exists a reliable work-around that can redress it. The problem **qpec2** is a fit case. The presence of constraints such as

$$0 \leq y \perp y \geq 0$$

renders an MPCC whose limit does not satisfy MPCC-LICQ although MPCC-MFCQ may hold. Replacing this constraint with $y = 0$ results in the limit satisfying Assumptions (A), and the algorithm converges. An effectual redress does not exist in the optimum packaging problems, **pack-comp1***, **pack-comp2***, **packcomp1c***, **pack-comp2c***, **pack-rig1***, **pack-rig2***, **pack-rig1c***, and **pack-rig2c***. The problems possess constraints other than complementarity constraints that do not have a strict interior, rendering the interior point algorithm an ineffectual choice. A reformulation has been proposed in the form of penalization of some of the inequalities in the objective so as to present a problem with a strict interior. The reformulated problems are named **pack-comp1p***, **pack-comp2p***, **pack-comp1cp***, **pack-comp2cp***, **pack-rig1p***, **pack-rig2p***, **pack-rig1cp***, and **pack-rig2cp***.

5.5. Comparison with LOQO and filtermpec. Finally, we compare our algorithm to the approaches of Benson, Shanno, and Vanderbei [2] and Fletcher and

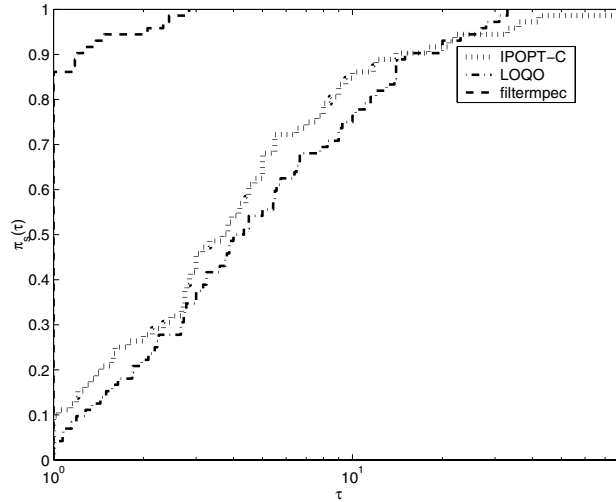


FIG. 5.4. Performance of IPOPT-C, LOQO, and filtermpec on MacMPEC in terms of number of iterations taken.

Leyffer [6]. Benson, Shanno, and Vanderbei [2] employ an interior point algorithm, LOQO [20], with no reformulation of the complementarity constraints to solve MPECs. They also suggest heuristics to handle degeneracy in the complementarity constraints and provide results for their approach. Fletcher and Leyffer [6] present results from using an SQP algorithm, filtermpec, to solve MPECs. Figure 5.4 compares the iteration count of filtermpec [6], LOQO [2], and IPOPT-C. The comparison considers only 77 of the MacMPEC problems on which all the algorithms converged to the same objective value. The results for LOQO and filtermpec are obtained from the respective papers. We also acknowledge the fact that results for the algorithms vary in tolerances but we can draw some general conclusions from the results.

The plot in Figure 5.4 indicates that filtermpec outperforms the interior point algorithms when comparing iterations. SQP algorithms generally require fewer iterations than interior point algorithms especially when NLPs have few inequalities. Further, SQP algorithms are more robust during failure of strict complementarity Assumption (A3) than interior point algorithms. We present iteration counts of the algorithms on a few representative problems in Table 5.2. Problems **bard1**, **bard2**, **jr1**, **jr2**, and **nash1** have linear equality, inequality constraints, and quadratic objectives. Table 5.2 indicates that filtermpec performs much better than the interior point algorithms on the above problems. The remaining problems possess a larger number of complementarity constraints and as a result more bounds. On these problems, filtermpec is sometimes better, sometimes worse than IPOPT-C and LOQO. A more interesting comparison would be the solution times of the algorithms on the larger problems. Our observation has been that interior point algorithms are faster than SQP algorithms when the number of complementarity constraints increases. The solution time for SQP algorithms can be larger due to the combinatorially intensive step of identifying the active-set. The results in Fletcher and Leyffer [6] also reveal that filtermpec requires more CPU time than LOQO. We avoid comparing solution times, as the papers [2, 6] do not provide the specifications on the processor speed and operating system on which results have been obtained.

TABLE 5.2

Iteration summary of IPOPT-C, LOQO, and filtermpec on representative MacMPEC problems.

	n	$p + q$	m	IPOPT-C	LOQO	filtermpec
bard1	5	1	3	11	14	1
bard2	12	5	4	23	34	2
jr1	2	1	1	8	9	1
jr2	2	1	1	8	11	6
nash1	6	2	2	10	14	1
incid-set1c-16	485	491	225	49	75	108
incid-set1c-32	1989	2003	961	60	164	41
incid-set2-16	485	491	225	54	59	69
incid-set2-32	1989	2003	961	93	151	39
incid-set2c-16	485	506	225	34	59	66
incid-set2c-32	1989	2034	961	57	456	33

6. Conclusions. For the convergence of the algorithm, the assumption that the iterates eventually enter the neighborhood of a solution satisfying strict complementarity, $\mathcal{N}_\gamma(\epsilon)$, is critical. This assumption is also made by Vicente and Wright [21] for interior point methods and Wright [25] for SQP methods for the solution of degenerate NLPs satisfying MFCQ. Such an assumption, also noted by these authors, cannot be guaranteed by their algorithms. Wright [27] has outlined an algorithm for identifying the strongly active indices and ensuring that the iterates enter a neighborhood, $\mathcal{N}_\gamma(\epsilon)$, if it exists. This requires the solution of linear programs. In contrast, the assumption that our iterates enter the neighborhood, $\mathcal{N}_\gamma(\epsilon)$, can be easily satisfied by appropriate choice of the complementarity relaxation parameter, t , and the barrier parameter, μ , under the Assumptions (A2)–(A3). Numerical experience also confirms the above observation. The convergence analysis of our algorithm will continue to hold if the MPCC-LICQ is relaxed to MPCC-MFCQ but will need an algorithm such as that of Wright [27] to guarantee that iterates enter $\mathcal{N}_\gamma(\epsilon)$.

Global convergence and improving robustness of the restoration phase algorithm will be investigated in the future. It remains to be seen whether a mechanism for ensuring bounded multipliers can be developed. Additionally, there are a number of MacMPEC problems on which our algorithm converges to solutions not satisfying some of Assumptions (A). The numerical behavior of these problems merits special attention.

Appendix. Summary of numerical results. This section lists the problem size and performance of algorithm IPOPT-C on all MacMPEC problems. The results provided correspond to choosing $\chi_t = \mu/t = 1$ and all other constants as described in section 5.3. The columns of the table correspond to the name of the problem (Name), number of variables (# var), number of constraints including complementarity constraints (# con), number of complementarity constraints (# ccon), number of iterations (# it), number of function evaluations (# fe), CPU seconds required for completion (CPU sec.), and objective value at convergence (Objective). The number of variables and constraints are those obtained after introduction of slacks by the AMPL interface as described in section 5.2. For a few problems, the table lists IL in the objective column indicating that the problem was not solved within the iteration limit of 1000 iterations. The problem **tap-09** has RF in the objective column to denote that the restoration phase failure caused termination of the algorithm. The results have been obtained on a 2.4 GHz Intel Xeon processor running LINUX as the operating system.

Name	# var	# con	# ccon	# it	# fe	CPU sec.	Objective
bar-truss-3	47	40	6	17	18	0.01	1.01666e+04
bard1	11	7	3	11	12	0.01	1.70000e+01
bard2	19	12	3	23	24	0.01	6.59800e+03
bard3	9	6	1	20	21	0.01	-1.26787e+01
bard1m	12	7	3	21	26	0.01	1.70000e+01
bard2m	19	12	3	23	24	0.01	-6.59800e+03
bard3m	13	8	3	20	21	0.01	-1.26787e+01
bem-milanc30-s	6365	4897	1464	1000	11836	243.68	IL
bilevel1	23	15	6	14	15	0.01	5.00000e+00
bilevel2	33	21	8	33	34	0.02	-6.60000e+03
bilevel3	18	13	3	22	23	0.01	-1.26787e+01
bilin	21	13	6	10	11	0	1.30000e+01
dempe	5	3	1	51	156	0.01	2.82503e+01
design-cent-1	20	14	3	8	9	0	1.86065e+00
design-cent-2	25	18	3	9	10	0	3.48382e+00
design-cent-3	23	14	3	17	19	0.02	3.72337e+00
design-cent-4	40	28	8	13	14	0.01	3.07920e+00
desilva	10	6	2	17	18	0	-9.99999e-01
df1	6	4	1	14	15	0.01	6.78632e-08
ex9.1.1	23	17	5	15	16	0	-1.30000e+01
ex9.1.2	12	9	2	9	10	0	-6.25000e+00
ex9.1.3	35	27	6	22	24	0.01	-2.92000e+01
ex9.1.4	12	9	2	20	26	0	-3.70000e+01
ex9.1.5	23	17	5	13	14	0.01	-1.00000e-00
ex9.1.6	26	19	6	18	19	0.01	-2.10000e+01
ex9.1.7	29	21	6	36	49	0.01	-2.60000e+01
ex9.1.8	14	10	2	11	12	0.01	-3.25000e+00
ex9.1.9	22	16	5	19	26	0.01	3.11111e+00
ex9.1.10	14	10	2	11	12	0	-3.25000e+00
ex9.2.1	18	13	4	18	19	0.01	1.70000e+01
ex9.2.3	23	17	4	14	16	0	5.00000e+00
ex9.2.4	12	9	2	14	15	0.01	5.00000e-01
ex9.2.5	14	10	3	17	18	0.01	5.00000e+00
ex9.2.6	28	18	6	14	15	0.01	-9.99999e-01
ex9.2.7	18	13	4	18	19	0	1.70000e+01
ex9.2.8	10	7	2	10	12	0	1.50000e+00
ex9.2.9	15	11	3	14	15	0.01	2.00000e+00
flp2	8	4	2	16	17	0	7.82036e-10
flp4-1	170	90	30	27	28	0.16	1.55380e-06
flp4-2	280	170	60	28	29	0.39	2.15161e-08
flp4-3	380	240	70	27	28	1.76	2.67991e-06
flp4-4	550	350	100	29	30	2.62	4.56477e-06
gauvin	7	4	2	12	15	0.01	2.00000e+01
gnash10	29	20	8	17	18	0.01	-2.30823e+02
gnash11	29	20	8	19	20	0.01	-1.29912e+02
gnash12	29	20	8	18	19	0.01	-3.69331e+01
gnash13	29	20	8	17	18	0	-7.06178e+00
gnash14	29	20	8	17	18	0	-1.79046e-01
gnash15	29	20	8	23	24	0.01	-3.54699e+02
gnash16	29	20	8	21	24	0	-2.41442e+02
gnash17	29	20	8	27	28	0.01	-9.07491e+01
gnash18	29	20	8	26	30	0.01	-2.56982e+01
gnash19	29	20	8	16	17	0	-6.11671e+00
hakonsen	18	12	4	17	19	0.01	2.43668e+01

Name	# var	# con	# ccon	# it	# fe	CPU sec.	Objective
hs044-i	40	24	10	20	22	0.01	1.56178e+01
incid-set1-08	236	168	49	24	28	0.23	6.84555e-06
incid-set1-16	976	716	225	43	47	2.16	3.13668e-05
incid-set1-32	3992	2964	961	145	286	54.83	1.35960e-04
incid-set1c-08	243	175	49	25	28	0.21	6.85070e-06
incid-set1c-16	991	731	225	49	61	4.91	3.70241e-05
incid-set1c-32	4023	2995	961	60	92	19.2	1.51486e-04
incid-set2-08	236	168	49	32	41	0.34	4.52651e-03
incid-set2-16	976	716	225	54	67	2.86	3.03272e-03
incid-set2-32	3992	2964	961	93	120	30.45	1.86717e-03
incid-set2c-08	243	175	49	25	27	0.23	5.48088e-03
incid-set2c-16	991	731	225	34	36	1.52	3.63571e-03
incid-set2c-32	4023	2995	961	57	65	14.79	2.57590e-03
jr1	4	2	1	8	9	0.01	5.00000e-01
jr2	4	2	1	8	9	0	5.00000e-01
kth1	4	2	1	10	11	0	2.83376e-07
kth2	4	2	1	9	10	0	5.38925e-08
kth3	4	2	1	8	9	0	5.00000e-01
liswet1-050	253	153	50	38	39	0.06	1.40025e-02
liswet1-100	503	303	100	51	52	0.15	1.37495e-02
liswet1-200	1003	603	200	74	75	0.44	1.70386e-02
nash11	10	6	2	10	11	0.01	7.43400e-15
nash12	10	6	2	10	11	0.01	8.01999e-15
nash13	10	6	2	10	11	0.01	8.20657e-15
nash14	10	6	2	18	19	0.01	2.75803e-09
nash15	10	6	2	10	11	0	8.12723e-15
outrata31	13	8	4	20	24	0	3.20770e+00
outrata32	13	8	4	18	19	0	3.44940e+00
outrata33	13	8	4	16	17	0	4.60425e+00
outrata34	13	8	4	16	17	0.01	6.59268e+00
pack-comp1cp-08	220	162	49	59	63	0.31	6.00002e-01
pack-comp1cp-16	948	706	225	44	46	1.31	6.23065e-01
pack-comp1cp-32	3940	2946	961	50	51	8.28	6.61555e-01
pack-comp1p-08	213	155	49	57	72	0.3	6.00005e-01
pack-comp1p-16	933	691	225	39	47	1.33	6.16980e-01
pack-comp1p-32	3909	2915	961	57	60	11.54	6.53089e-01
pack-comp2cp-08	220	162	49	80	97	0.45	6.73074e-01
pack-comp2cp-16	948	706	225	32	34	122.85	7.27545e-01
pack-comp2cp-32	3940	2946	961	51	55	9.52	7.83058e-01
pack-comp2p-08	213	155	49	49	52	0.28	6.72532e-01
pack-comp2p-16	933	691	225	39	41	1.81	7.27276e-01
pack-comp2p-32	3909	2915	961	42	48	10.22	7.82717e-01
pack-rig1cp-08	214	158	47	29	31	0.13	7.88300e-01
pack-rig1cp-16	882	662	203	35	36	0.96	8.26530e-01
pack-rig1cp-32	3640	2746	861	48	49	6.47	8.51760e-01
pack-rig1p-08	207	151	47	27	29	0.12	7.87932e-01
pack-rig1p-16	867	647	203	33	34	1.1	8.26044e-01
pack-rig1p-32	3609	2715	861	53	60	502.22	8.51010e-01
pack-rig2cp-08	208	154	45	27	29	0.11	7.99314e-01
pack-rig2cp-16	855	644	194	40	43	1.14	9.49266e-01
pack-rig2cp-32	3499	2652	814	62	64	9.34	1.30423e+00
pack-rig2p-08	201	147	45	28	29	0.14	7.80413e-01
pack-rig2p-16	840	629	194	49	50	1.33	1.08517e+00
pack-rig2p-32	3468	2621	814	57	58	8.23	1.13600e+00

Name	# var	# con	# ccon	# it	# fe	CPU sec.	Objective
portfl-i-1	111	37	12	24	25	0.05	1.90844e-05
portfl-i-2	111	37	12	22	23	0.04	1.89189e-05
portfl-i-3	111	37	12	21	22	0.04	1.14944e-05
portfl-i-4	111	37	12	23	24	0.06	6.75604e-06
portfl-i-6	111	37	12	22	23	0.03	2.68413e-06
qpec-100-1	307	202	100	24	25	1.1	1.09237e-01
qpec-100-2	312	202	100	45	50	2.01	-6.44521e+00
qpec-100-3	314	204	100	28	29	1	-5.48288e+00
qpec-100-4	324	204	100	25	27	0.99	-3.98212e+00
qpec1	70	40	20	10	11	0.01	8.00000e+01
qpec-200-1	614	404	200	38	44	14.68	-1.93490e+00
qpec-200-2	624	404	200	44	46	16.33	-2.40774e+01
qpec-200-3	628	408	200	1000	11295	292.23	IL
qpec-200-4	648	408	200	27	34	12.15	-6.23052e+00
ralph2	4	2	1	14	15	0	-6.07653e-07
ralphmod	304	200	100	85	91	3.23	-6.83033e+02
scholtes1	5	2	1	14	17	0	2.00000e+00
scholtes2	5	2	1	15	18	0.01	1.50000e+01
scholtes3	4	2	1	12	14	0	5.00000e-01
scholtes5	7	4	2	8	9	0	1.00000e+00
siouxfls	5899	4124	1748	655	702	397.04	2.08258e+02
siouxfls1	5899	4124	1748	1000	7200	342.74	IL
sl1	14	8	3	26	27	0.01	1.00296e-04
stackelberg1	5	3	1	11	12	0.01	-3.26667e+03
tap-09	154	100	32	148	207	0.75	RF
tap-15	376	250	83	631	684	11.08	2.53404e+02
water-FL	301	204	44	1000	7797	5.44	IL
water-net	94	64	14	1000	9509	1.89	IL

Acknowledgments. The authors are grateful to Andreas Wächter for help in implementing the step modification within the algorithm, IPOPT, and for many interesting discussions which helped improve the performance of the proposed algorithm, IPOPT-C. We are also thankful to Mihai Anitescu and Steve Wright for careful reading of an earlier version of the manuscript. Finally, we gratefully acknowledge two anonymous referees whose suggestions helped improve the presentation and also strengthen the paper.

REFERENCES

- [1] M. ANITESCU, *On Using the Elastic Mode in Nonlinear Programming Approaches to Mathematical Programs with Complementarity Constraints*, Tech. report ANL/MCS-P864-1200, Argonne National Laboratories, Argonne, IL, 2000.
- [2] H. Y. BENSON, D. F. SHANNO, AND R. J. VANDERBEI, *Interior-Point Methods for Nonconvex Nonlinear Programming: Complementarity Constraints*, Tech. report, Research Report ORFE-02-02, Operations Research and Financial Engineering, Princeton University, Princeton, NJ, 2002.
- [3] A. V. DE MIGUEL, M. P. FREIDLANDER, F. J. NOGALES, AND S. SCHOLTES, *An interior-point method for mathematical programs with equilibrium constraints*, LBS Working Paper, 2003.
- [4] E. D. DOLAN AND J. J. MORÉ, *Benchmarking optimization software with performance profiles*, Math. Program., 91A, (2002), pp. 201–213.
- [5] F. FACCHINEI, H. JIANG, AND L. QI, *A smoothing method for mathematical programs with equilibrium constraints*, Math. Program., 85A, (1999), pp. 107–134.

- [6] R. FLETCHER AND S. LEYFFER, *Numerical Experience with Solving MPECs as NLPs*, Tech. report Numerical Analysis Report NA/210, Department of Mathematics, University of Dundee, Dundee, UK, 2002.
- [7] R. FLETCHER, S. LEYFFER, S. SCHOLTES, AND D. RALPH, *Local Convergence of SQP Methods for Mathematical Programs with Equilibrium Constraints*, Tech. report, Numerical Analysis Report NA/183, Department of Mathematics, University of Dundee, Dundee, UK, 2002.
- [8] R. FOURER, D. M. GAY, AND B. W. KERNIGHAN, *AMPL A Modeling Language for Mathematical Programming*, 2nd ed., Thomson Brooks/Cole, Stamford, CT, 2003.
- [9] W. W. HAGER, *Stabilized sequential quadratic programming*, Comput. Optim. Appl., 12 (1999), pp. 253–273.
- [10] W. W. HAGER AND M. S. GOWDA, *Stability in the presence of degeneracy and error estimation*, Math. Program., 85A (1999), pp. 181–192.
- [11] H. JIANG AND D. RALPH, *Smooth SQP methods for mathematical programs with nonlinear complementarity constraints*, SIAM J. Optim., 10 (2000), pp. 779–808.
- [12] S. LEYFFER, *MacMPEC—www.unix.mcs.anl.gov/leyffer/macmpec/*, 2002.
- [13] X. LIU AND J. SUN, *Generalized Stationary Points and an Interior Point Method for Mathematical Programs with Equilibrium Constraints*, Tech. report, Department of Decision Sciences and Singapore-MIT Alliance, National University of Singapore, Republic of Singapore, 2002.
- [14] Z.-Q. LUO, J.-S. PANG, AND D. RALPH, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, Cambridge, UK, 1996.
- [15] Z.-Q. LUO, J.-S. PANG, AND D. RALPH, *Piecewise sequential quadratic programming for mathematical programs with nonlinear complementarity constraints*, in Multilevel Optimization: Algorithms and Applications, A. Migdalas, P. Pardalos, and P. Värbrand, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 209–229.
- [16] J. OUTRATA, M. KOČVARA, AND J. ZOWE, *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints: Theory, Applications and Numerical Results*, Kluwer Academic Publishers, Dordrecht, Boston, London, 1998.
- [17] H. SCHEEL AND S. SCHOLTES, *Mathematical programs with complementarity constraints: Stationarity, optimality and sensitivity*, Math. Oper. Res., 25 (2000), pp. 1–22.
- [18] S. SCHOLTES, *Convergence properties of a regularization scheme for mathematical programs with complementarity constraints*, SIAM J. Optim., 11 (2001), pp. 918–936.
- [19] S. SCHOLTES AND M. STÖHR, *Exact penalization of mathematical programs with equilibrium constraints*, SIAM J. Control Optim., 37 (1999), pp. 617–652.
- [20] R. J. VANDERBEI AND D. F. SHANNO, *An interior-point algorithm for nonconvex nonlinear programming*, Comput. Optim. Appl., 13 (1999), pp. 231–252.
- [21] L. N. VICENTE AND S. WRIGHT, *Local convergence of a primal-dual method for degenerate nonlinear programming*, Comput. Optim. Appl., 22 (2002), pp. 311–328.
- [22] A. WÄCHTER, *Interior point methods for large-scale nonlinear programming with applications*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, 2002.
- [23] A. WÄCHTER AND L. T. BIEGLER, *Global and Local Convergence of Line Search Filter Methods for Nonlinear Programming*, Tech. report, Dept. of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA, 2002.
- [24] M. H. WRIGHT, *Interior point methods for constrained optimization*, Acta Numer., 1 (1992), pp. 341–407.
- [25] S. J. WRIGHT, *Superlinear convergence of a stabilized SQP method to a degenerate solution*, Comput. Optim. Appl., 11 (1998), pp. 253–275.
- [26] S. J. WRIGHT, *Effect of Finite Precision Arithmetic on Interior-Point Methods for Nonlinear Programming*, SIAM J. Optim., 12 (2001), pp. 36–78.
- [27] S. J. WRIGHT, *Constraint identification and algorithm stabilization for degenerate nonlinear programming*, Math. Program., 95B, (2003), pp. 137–160.

A ROBUST GRADIENT SAMPLING ALGORITHM FOR NONSMOOTH, NONCONVEX OPTIMIZATION*

JAMES V. BURKE[†], ADRIAN S. LEWIS[‡], AND MICHAEL L. OVERTON[§]

Abstract. Let f be a continuous function on $\mathbf{R}^n$, and suppose f is continuously differentiable on an open dense subset. Such functions arise in many applications, and very often minimizers are points at which f is not differentiable. Of particular interest is the case where f is not convex, and perhaps not even locally Lipschitz, but is a function whose gradient is easily computed where it is defined. We present a practical, robust algorithm to locally minimize such functions, based on *gradient sampling*. No subgradient information is required by the algorithm.

When f is locally Lipschitz and has bounded level sets, and the sampling radius ϵ is fixed, we show that, with probability 1, the algorithm generates a sequence with a cluster point that is Clarke ϵ -stationary. Furthermore, we show that if f has a unique Clarke stationary point $\bar{x}$, then the set of all cluster points generated by the algorithm converges to $\bar{x}$ as ϵ is reduced to zero.

Numerical results are presented demonstrating the robustness of the algorithm and its applicability in a wide variety of contexts, including cases where f is not locally Lipschitz at minimizers. We report approximate local minimizers for functions in the applications literature which have not, to our knowledge, been obtained previously. When the termination criteria of the algorithm are satisfied, a precise statement about nearness to Clarke ϵ -stationarity is available. A MATLAB implementation of the algorithm is posted at <http://www.cs.nyu.edu/overton/papers/gradsamp/alg>.

Key words. generalized gradient, nonsmooth optimization, subgradient, gradient sampling, nonconvex

AMS subject classifications. 65K10, 90C26

DOI. 10.1137/030601296

1. Introduction. The analysis of nonsmooth, nonconvex functions has been a rich area of mathematical research for three decades. Clarke introduced the notion of generalized gradient in [Cla73, Cla83]; comprehensive studies of more recent developments may be found in [CLSW98, RW98]. The generalized gradient of a function f at a point x reduces to the gradient if f is smooth at x and to the subdifferential if f is convex; hence, we follow common usage in referring to the generalized gradient as the (Clarke) subdifferential, or set of (Clarke) subgradients. Its use in optimization algorithms began soon after its appearance in the literature. In particular, the concept of the ϵ -steepest descent direction for locally Lipschitz functions was introduced by Goldstein in [Gol77]; another early paper is [CG78]. It is well known that the ordinary steepest descent algorithm typically fails by converging to a nonoptimal point when applied to nonsmooth functions, whether convex or not. The fundamental difficulty is that most interesting nonsmooth objective functions have minimizers where the gradient is not defined.

An extensive discussion of several classes of algorithms for the minimization of

*Received by the editors October 20, 2003; accepted for publication June 2, 2004; published electronically April 8, 2005.

<http://www.siam.org/journals/siopt/15-3/60129.html>

[†]Department of Mathematics, University of Washington, Seattle, WA 98195 (burke@math.washington.edu). Research supported in part by National Science Foundation grant DMS-0203175.

[‡]School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14853 (aslewis@orie.cornell.edu). Research supported in part by National Science and Engineering Research Council of Canada at Simon Fraser University, Burnaby, BC, Canada.

[§]Courant Institute of Mathematical Sciences, New York University, New York, NY 10012 (overton@cs.nyu.edu). Research supported in part by National Science Foundation grant DMS-0412049.

nonsmooth, nonconvex, locally Lipschitz functions, complete with convergence analysis, may be found in Kiwiel's book [Kiw85]. What these algorithms have in common is that, at each iteration, they require the computation of a single subgradient (not the entire subdifferential set) in addition to the value of the function. The algorithms then build up information about the subdifferential properties of the function using ideas known as bundling and aggregation. Such "bundle" algorithms, as they are generally known, are especially effective for nonsmooth, convex optimization because of the global nature of convexity, and the ideas in this case trace back to [Lem75, Wol75]; for a comprehensive discussion of the convex case, see [HUL93]. However, for nonconvex functions, subgradient information is meaningful only locally and must be discounted when no longer relevant. The consequence is that bundle algorithms are necessarily much more complicated in the nonconvex case. Other contributions to nonconvex bundle methods since Kiwiel's book was published in 1985 include [FGG02, Gro02, LSB91, LV98, MN92, OKZ98, SZ92]. Despite this activity in the field, the only publicly available nonconvex bundle software of which we are aware are the Bundle Trust (BT) FORTRAN code dating from 1991 [SZ92] and some more recent FORTRAN codes of [LV98].

In addition to this body of work on general nonsmooth, nonconvex optimization, there is a large literature on more specialized problems, including nonconvex polyhedral functions [Osb85], compositions of convex and smooth functions [Bur85, Fle87], and quasi-differentiable functions [DR95].

There are many reasons why most algorithms for nonsmooth optimization do not ask the user to provide a description of the entire subdifferential set at each iterate. One is that this would demand a great deal of the user in all but the simplest applications. More fundamentally, it is not clear how one would represent such a set in general since it is already a formidable task in the polyhedral setting [Osb85]. Even if this were resolved, implementation would be difficult for the user given the inherent complexity of the continuity properties of these set-valued mappings. Asking the user to provide only one subgradient at a point resolves these difficulties.

In virtually all interesting applications, the function being minimized is continuously differentiable almost everywhere, although it is often not differentiable at minimizers. Under this assumption, when a user is asked to provide a subgradient at a randomly selected point, with probability 1 the subgradient is unique, namely, the gradient. This observation led us to consider a simple gradient sampling algorithm, first presented without any analysis in [BLO02b]. At a given iterate, we compute the gradient of the objective function on a set of randomly generated nearby points, and use this information to construct a local search direction that may be viewed as an approximate ϵ -steepest descent direction, where ϵ is the sampling radius. As is standard for algorithms based on subgradients, we obtain the descent direction by solving a quadratic program. Gradient information is not saved from one iteration to the next, but discarded once a lower point is obtained from a line search. A key motivating factor is that, in many applications, computing the gradient when it exists is little additional work once the function value is computed. Often, well-known formulas for the gradient are available; alternatively, automatic differentiation might be used. No subgradient information is required from the user. We have found the gradient sampling algorithm to be very effective for approximating local minimizers of a wide variety of nonsmooth, nonconvex functions, including non-Lipschitz functions.

In a separate work [BLO02a], we analyzed the extent to which the Clarke subdifferential at a point can be approximated by random sampling of gradients at nearby

points, justifying the notion that the convex hull of the latter set can serve as a surrogate for the former.

This paper is organized as follows. The gradient sampling (GS) algorithm is presented in section 2. The sampling radius ϵ may be fixed for all iterates or may be reduced dynamically; this is controlled by the choice of parameters defining the algorithm.

A convergence analysis is given in section 3, making the assumption that the function $f : \mathbf{R}^n \rightarrow \mathbf{R}$ is locally Lipschitz, has bounded level sets, and, in addition, is continuously differentiable on an open dense subset of $\mathbf{R}^n$. Our first convergence result analyzes the GS algorithm with fixed ϵ , and establishes that, with probability 1, it generates a sequence with a cluster point that is Clarke ϵ -stationary, in a sense that will be made precise. A corollary shows that if f has a unique Clarke stationary point $\bar{x}$, then the sets of all cluster points generated by the GS algorithm converge to $\bar{x}$ as ϵ is reduced to zero. These results are then strengthened for the case where f is either convex or smooth. In all cases, when the termination criteria of the GS algorithm are satisfied, a precise statement about nearness to Clarke ϵ -stationarity is available.

We should emphasize that although Clarke stationarity is a first-order optimality condition, there are two considerations that allow us to expect that, in practice, cluster points of the GS algorithm are more than just approximate stationary points, but are in fact approximate local minimizers. The first consideration is a very practical one: the line search enforces a descent property for the sequence of iterates. The second consideration is more theoretical: we are generally interested in applying the algorithm to a nonsmooth function that, although not convex, is *subdifferentially regular* [RW98] (equivalently, its epigraph is regular in the sense of Clarke [Cla83]). Clarke stationarity at a point of subdifferential regularity implies the nonnegativity of the usual directional derivative in all directions. This is much stronger than Clarke stationarity in the absence of regularity. For example, 0 is a Clarke stationary point of the function $f(x) = -|x|$, but f is not subdifferentially regular at 0.

In section 4, we present numerical results that demonstrate the effectiveness and robustness of the GS algorithm and its applicability in a variety of contexts. We set the parameters defining the GS algorithm so that the sampling radius ϵ is reduced dynamically. We begin with a classical problem: Chebyshev exponential approximation. Our second example involves minimizing a product of eigenvalues of a symmetric matrix; it arises in an environmental data analysis application. We then turn to some important functions arising in nonsymmetric matrix analysis and robust control, including non-Lipschitz spectral functions, pseudospectral functions, and the distance to instability. We conclude with a stabilization problem for a model of a Boeing 767 at a flutter condition. As far as we know, none of the problems that we present has been solved previously by any method.

Finally, we make some concluding remarks in section 5. Our MATLAB implementation of the GS algorithm is freely available on the Web.¹

2. The gradient sampling algorithm. The GS algorithm is conceptually very simple. Basically, it is a stabilized steepest descent algorithm. At each iteration, a descent direction is obtained by evaluating the gradient at the current iterate and at additional nearby points and then computing the vector in the convex hull of these gradients with smallest norm. A standard line search is then used to obtain a lower

¹<http://www.cs.nyu.edu/overton/papers/gradsamp/alg/>.

point. Thus, stabilization is controlled by the sampling radius used to sample the gradients. In practice, we begin with a large sampling radius and then reduce this according to rules that are set out in section 4, where we show, using various examples, how well the algorithm works.

Despite the simplicity and power of the algorithm, its analysis is not so simple. One difficulty is that it is inherently probabilistic, since the gradient is not defined on the whole space. Another is that analyzing its convergence for a fixed sampling radius is already challenging, and extending our results in that case to a version of the algorithm that reduces the sampling radius dynamically presents additional difficulties. In order to take care of both fixed and dynamically changing sampling radius, the statement of the algorithm becomes a little more complicated. We set out the algorithmic details in this section and present the convergence analysis in the next section.

The algorithm may be applied to any function $f : \mathbf{R}^n \rightarrow \mathbf{R}$ that is continuous on $\mathbf{R}^n$ and differentiable almost everywhere. However, all our theoretical results assume that f is locally Lipschitz continuous and continuously differentiable on an open dense subset D of $\mathbf{R}^n$. In addition, we assume that there is a point $\tilde{x} \in \mathbf{R}^n$ for which the set $\mathcal{L} = \{x \mid f(x) \leq f(\tilde{x})\}$ is compact.

The local Lipschitz hypothesis allows us to approximate the Clarke subdifferential [Cla83] as follows. For each $\epsilon > 0$, define the multifunction $G_\epsilon : \mathbf{R}^n \rightrightarrows \mathbf{R}^n$ by

$$G_\epsilon(x) = \text{cl conv } \nabla f((x + \epsilon\mathbb{B}) \cap D),$$

where $\mathbb{B} = \{x \mid \|x\| \leq 1\}$ is the closed unit ball and $\|\cdot\|$ is the 2-norm. The sets $G_\epsilon(x)$ can be used to give the following representation of the Clarke subdifferential of f at a point x :

$$\bar{\partial}f(x) = \bigcap_{\epsilon > 0} G_\epsilon(x).$$

We also make use of the ϵ -subdifferential introduced by Goldstein [Gol77]. For each $\epsilon > 0$, the Clarke ϵ -subdifferential is given by

$$\bar{\partial}_\epsilon f(x) = \text{cl conv } \bar{\partial}f(x + \epsilon\mathbb{B}).$$

Clearly, $G_\epsilon(x) \subset \bar{\partial}_\epsilon f(x)$, and for $0 < \epsilon_1 < \epsilon_2$ we have $\bar{\partial}_{\epsilon_1} f(x) \subset G_{\epsilon_2}(x)$. In addition, it is easily shown that the multifunction $\bar{\partial}_\epsilon f$ has closed graph.

We say that a point x is a Clarke ϵ -stationary point for f if $0 \in \bar{\partial}_\epsilon f(x)$. This notion of ϵ -stationarity is key to our approach. Indeed, the algorithm described below is designed to locate Clarke ϵ -stationary points. For this reason we introduce the following scalar measure of proximity to Clarke ϵ -stationarity:

$$(1) \quad \rho_\epsilon(x) = \text{dist}(0 \mid G_\epsilon(x)).$$

We now state the GS algorithm. Scalar parameters are denoted by lowercase Greek letters. A superscript on a scalar parameter indicates taking that scalar to the power of the superscript.

In order to facilitate the reading and analysis of the algorithm, we provide a partial glossary of the notation used in its statement.

Glossary of Notation	
k : Iteration counter.	μ : Sampling radius reduction factor.
x^k : Current iterate.	θ : Optimality tolerance reduction factor.
γ : Backtracking reduction factor.	m : Sample size.
$\mathcal{L}$: $\{x \mid f(x) \leq f(\bar{x})\}$.	D : Points of differentiability.
u^{kj} : Unit ball samples.	x^{kj} : Sampling points.
β : Armijo parameter.	g^k : Shortest approximate subgradient.
ϵ_k : Sampling radius.	d^k : Search direction.
ν_k : Optimality tolerance.	t_k : Step length.

The GS algorithm.**Step 0:** (Initialization)

Let $x^0 \in \mathcal{L} \cap D$, $\gamma \in (0, 1)$, $\beta \in (0, 1)$, $\epsilon_0 > 0$, $\nu_0 \geq 0$, $\mu \in (0, 1]$, $\theta \in (0, 1]$, $k = 0$, and $m \in \{n + 1, n + 2, \dots\}$.

Step 1: (Approximate the Clarke ϵ -subdifferential by gradient sampling)

Let $u^{k1}, \dots, u^{km}$ be sampled independently and uniformly from $\mathbb{B}$, and set

$$x^{k0} = x^k \quad \text{and} \quad x^{kj} = x^k + \epsilon_k u^{kj}, \quad j = 1, \dots, m.$$

If for some $j = 1, \dots, m$ the point $x^{kj} \notin D$, then STOP; otherwise, set

$$G_k = \text{conv} \{ \nabla f(x^{k0}), \nabla f(x^{k1}), \dots, \nabla f(x^{km}) \},$$

and go to Step 2.

Step 2: (Compute a search direction)

Let $g^k \in G_k$ solve the quadratic program $\min_{g \in G_k} \|g\|^2$, i.e.,

$$\|g^k\| = \text{dist}(0 \mid G_k) \quad \text{and} \quad g^k \in G_k.$$

If $\nu_k = \|g^k\| = 0$, STOP. If $\|g^k\| \leq \nu_k$, set $t_k = 0$, $\nu_{k+1} = \theta \nu_k$, and $\epsilon_{k+1} = \mu \epsilon_k$, and go to Step 4; otherwise, set $\nu_{k+1} = \nu_k$, $\epsilon_{k+1} = \epsilon_k$, and $d^k = -g^k / \|g^k\|$, and go to Step 3.

Step 3: (Compute a step length)

Set

$$t_k = \max \gamma^s \\ \text{subject to} \quad s \in \{0, 1, 2, \dots\} \text{ and} \\ f(x^k + \gamma^s d^k) < f(x^k) - \beta \gamma^s \|g^k\|,$$

and go to Step 4.

Step 4: (Update)

If $x^k + t_k d^k \in D$, set $x^{k+1} = x^k + t_k d^k$, $k = k + 1$, and go to Step 1. If $x^k + t_k d^k \notin D$, let $\hat{x}^k$ be any point in $x^k + \epsilon_k \mathbb{B}$ satisfying $\hat{x}^k + t_k d^k \in D$ and

$$(2) \quad f(\hat{x}^k + t_k d^k) < f(x^k) - \beta t_k \|g^k\|$$

(such an $\hat{x}^k$ exists due to the continuity of f). Then set $x^{k+1} = \hat{x}^k + t_k d^k$, $k = k + 1$, and go to Step 1.

The algorithm is designed so that every iterate x^k is an element of the set $\mathcal{L} \cap D$. We now show that the line search defined in Step 3 of the algorithm is well defined in the sense that the value of t_k can be determined by a finite process. Let $N_C(x)$ denote the normal cone to the set C at a point x [Roc70], and recall from convex analysis

that g^k solves $\inf_{g \in G_k} \|g\|$ if and only if $-g^k \in N_{G_k}(g^k)$, that is, $\langle g - g^k, -g^k \rangle \leq 0$ for all $g \in G_k$. Therefore, if $\|g^k\| = \text{dist}(0 | G_k) \neq 0$, then

$$\nabla f(x^k)^T d^k \leq \sup_{g \in G_k} \langle g, d^k \rangle \leq -\|g^k\|.$$

Since $x^k \in D$, we have $f'(x^k; d^k) = \nabla f(x^k)^T d^k$. Hence, there is a $\bar{t} > 0$ such that

$$f(x^k + td^k) \leq f(x^k) + t\beta \nabla f(x^k)^T d^k \leq f(x^k) - t\beta \|g^k\| \quad \forall t \in (0, \bar{t}).$$

The choice of search direction used in the GS algorithm is motivated by the direction of steepest descent in nonsmooth optimization. Recall that the Clarke directional derivative for f at a point x is given by the support functional for the Clarke subdifferential at x :

$$f^\circ(x; d) = \max_{z \in \partial f(x)} \langle z, d \rangle.$$

Therefore, the direction of steepest descent is obtained by solving the problem

$$\min_{\|d\| \leq 1} f^\circ(x; d) = \min_{\|d\| \leq 1} \max_{z \in \partial f(x)} \langle z, d \rangle.$$

The next lemma, which is essentially well known, shows that the search direction in the GS algorithm is an approximate direction of steepest descent.

LEMMA 2.1. *Let G be any compact convex subset of $\mathbf{R}^n$; then*

$$(3) \quad -\text{dist}(0 | G) = \min_{\|d\| \leq 1} \max_{g \in G} \langle g, d \rangle.$$

Moreover, if $\bar{g} \in G$ satisfies $\|\bar{g}\| = \text{dist}(0 | G)$, then $\bar{d} = -\bar{g}/\|\bar{g}\|$ solves the problem on the right-hand side of (3).

Proof. The result is an elementary consequence of the von Neumann minimax theorem. Indeed, one has

$$\begin{aligned} -\text{dist}(0 | G) &= -\min_{g \in G} \|g\| \\ &= -\min_{g \in G} \max_{\|d\| \leq 1} \langle g, d \rangle \\ &= -\max_{\|d\| \leq 1} \min_{g \in G} \langle g, d \rangle \\ &= -\max_{\|d\| \leq 1} \min_{g \in G} \langle g, -d \rangle \\ &= \min_{\|d\| \leq 1} \max_{g \in G} \langle g, d \rangle, \end{aligned}$$

from which it easily follows that $d = -\bar{g}/\|\bar{g}\|$ solves the problem

$$\inf_{\|d\| \leq 1} \sup_{g \in G} \langle g, d \rangle,$$

where $\bar{g}$ is the least norm element of G . $\square$

By setting G equal to the sets G_k in the GS algorithm, we obtain the *approximate steepest descent property*:

$$-\text{dist}(0 | G_k) = \min_{\|d\| \leq 1} \max_{g \in G_k} \langle g, d \rangle.$$

The case $x^k + t_k d^k \notin D$ in Step 4 of the algorithm seems unlikely to occur, and we do not correct for this possibility in our numerical implementation. Nonetheless, we need to compensate for it in our theoretical analysis since we have not been able to show that it is a zero probability event. The GS algorithm is a nondeterministic algorithm, and in this spirit Step 4 is easily implemented in a nondeterministic fashion as follows. At step $\omega = 1, 2, \dots$, sample $\hat{x}$ from a uniform distribution on $x^k + (\epsilon_k/\omega)\mathbb{B}$ and check to see if $\hat{x} + t_k d^k \in D$ and the inequality (2) with $\hat{x}^k = \hat{x}$ is satisfied. If so, set $\hat{x}^k = \hat{x}$, $x^{k+1} = \hat{x}^k + t_k d^k$, $k = k + 1$, and return to Step 1; otherwise, increase ω and repeat. With probability 1 this procedure terminates finitely.

The GS algorithm can be run with $\nu_0 = 0$ and $\mu = 1$, so that $\nu_k = 0$ and $\epsilon_k = \epsilon_0$ for all k . This instance of the algorithm plays a prominent role in our convergence analysis. Indeed, all of our theoretical convergence results follow from the analysis in this case. In practice, however, the algorithm is best implemented with $\mu < 1$ and ν_0 positive. When $\nu_0 = 0$, the algorithm terminates at iteration k_0 if either $x^{k_0 j} \notin D$ for some $j = 1, \dots, m$ or $\|g^{k_0}\| = 0$. The probability that $x^{k_0 j} \notin D$ for some $j = 1, \dots, m$ is zero, while $\|g^{k_0}\| = 0$ is equivalent to $\rho_{\epsilon_{k_0}}(x^{k_0}) = 0$.

Before proceeding to the convergence analysis, we make a final observation concerning the stochastic structure of the algorithm, as it plays a key role in our analysis. Although the algorithm specifies that the points $u^{k_1}, \dots, u^{k_m}$ are sampled from $\mathbb{B}$ at each iteration, we may think of this sequence as a realization of a stochastic process $\{(\mathbf{u}^{k_1}, \dots, \mathbf{u}^{k_m})\}$ where the realization occurs before the initiation of the algorithm. In this regard, we consider only those realizations that are ergodic with respect to $\mathbb{B}^m$. Specifically, we consider only those processes that hit every positive measure subset of $\mathbb{B}^m$ infinitely often. We define this subset of events as $\mathcal{E}$ and note that with probability 1 the realization $\{u^{k_1}, \dots, u^{k_m}\}$ is in $\mathcal{E}$.

3. Convergence analysis. Throughout this section it is assumed that the function $f : \mathbf{R}^n \rightarrow \mathbf{R}$ is locally Lipschitz continuous on $\mathbf{R}^n$ and continuously differentiable on an open dense subset D of $\mathbf{R}^n$. We begin with two technical lemmas.

LEMMA 3.1. *Let $v \in C$, where C is a nonempty closed convex subset of $\mathbf{R}^n$ that does not contain the origin. If $\delta > 0$, $\eta > 0$, and $u, \bar{u} \in C$ are such that $\eta \leq \|\bar{u}\| = \text{dist}(0 | C)$ and $\|u\| \leq \|\bar{u}\| + \delta$, then*

$$\left\langle v - u, \frac{-u}{\|u\|} \right\rangle \leq \left[\|v\| \sqrt{\frac{2}{\eta}} + \sqrt{2\|\bar{u}\| + \delta} \right] \sqrt{\delta}.$$

Proof. Since $\|\bar{u}\| = \text{dist}(0 | C)$, we have $-\bar{u} \in N_C(\bar{u})$. Hence, for all $h \in C$

$$\langle h - \bar{u}, -\bar{u} \rangle \leq 0,$$

or, equivalently,

$$\|\bar{u}\|^2 \leq \langle h, \bar{u} \rangle.$$

Therefore,

$$1 - \left\langle \frac{u}{\|u\|}, \frac{\bar{u}}{\|\bar{u}\|} \right\rangle \leq 1 - \frac{\|\bar{u}\|^2}{\|u\| \|\bar{u}\|} = \frac{1}{\|u\|} [\|u\| - \|\bar{u}\|] \leq \frac{\delta}{\|u\|} \leq \frac{\delta}{\|\bar{u}\|},$$

$$\left\| \frac{u}{\|u\|} - \frac{\bar{u}}{\|\bar{u}\|} \right\|^2 = 2 \left[1 - \left\langle \frac{u}{\|u\|}, \frac{\bar{u}}{\|\bar{u}\|} \right\rangle \right] \leq \frac{2\delta}{\|\bar{u}\|},$$

$$\begin{aligned}
\|u - \bar{u}\|^2 &= \|u\|^2 - 2\langle u, \bar{u} \rangle + \|\bar{u}\|^2 \\
&\leq \|u\|^2 - \|\bar{u}\|^2 \\
&= (\|u\| + \|\bar{u}\|)(\|u\| - \|\bar{u}\|) \\
&\leq [2\|\bar{u}\| + \delta]\delta,
\end{aligned}$$

and

$$\begin{aligned}
\|v - \bar{u}\|^2 &= \|v\|^2 - 2\langle v, \bar{u} \rangle + \|\bar{u}\|^2 \\
&\leq \|v\|^2 - \|\bar{u}\|^2 \\
&\leq \|v\|^2.
\end{aligned}$$

Consequently,

$$\begin{aligned}
\left\langle v - u, \frac{-u}{\|u\|} \right\rangle &= \left\langle v - \bar{u}, \frac{-\bar{u}}{\|\bar{u}\|} \right\rangle + \left\langle v - \bar{u}, \frac{\bar{u}}{\|\bar{u}\|} - \frac{u}{\|u\|} \right\rangle + \left\langle u - \bar{u}, \frac{u}{\|u\|} \right\rangle \\
&\leq \left\langle v - \bar{u}, \frac{\bar{u}}{\|\bar{u}\|} - \frac{u}{\|u\|} \right\rangle + \left\langle u - \bar{u}, \frac{u}{\|u\|} \right\rangle \\
&\leq \|v - \bar{u}\| \left\| \frac{u}{\|u\|} - \frac{\bar{u}}{\|\bar{u}\|} \right\| + \|u - \bar{u}\| \\
&\leq \|v\| \sqrt{\frac{2\delta}{\|\bar{u}\|}} + \sqrt{[2\|\bar{u}\| + \delta]\delta} \\
&\leq \left[\|v\| \sqrt{\frac{2}{\eta}} + \sqrt{[2\|\bar{u}\| + \delta]} \right] \sqrt{\delta}. \quad \square
\end{aligned}$$

The next lemma establishes properties of the set of all points close to a given point x' that can be used to provide a δ -approximation to the element of $G_\epsilon(x')$ of least norm. Specifically, let $m \geq n + 1$, $\delta > 0$, and $x', x \in \mathbf{R}^n$ be given, and set

$$D_\epsilon^m(x) = \prod_1^m (D \cap (x + \epsilon\mathbb{B})) \subset \prod_1^m \mathbf{R}^n.$$

Consider the set $R_\epsilon(x', x, \delta) \subset \prod_1^{m+1} \mathbf{R}^n$ of all $(m+1)$ -tuples $(x^1, \dots, x^m, g)$ satisfying

$$(x^1, \dots, x^m) \in D_\epsilon^m(x) \quad \text{and} \quad g = \sum_{j=1}^m \lambda_j \nabla f(x^j)$$

for some $0 \leq \lambda_j$, $j = 1, 2, \dots, m$, with

$$\sum_{j=1}^m \lambda_j = 1 \quad \text{and} \quad \left\| \sum_{j=1}^m \lambda_j \nabla f(x^j) \right\| \leq \rho_\epsilon(x') + \delta.$$

We need to understand the local behavior of this set as well as its projections

$$V_\epsilon(x', x, \delta) = \{(x^1, \dots, x^m) : \exists g \text{ with } (x^1, \dots, x^m, g) \in R_\epsilon(x', x, \delta)\}$$

and

$$U_\epsilon(x', x, \delta) = \{g : \exists (x^1, \dots, x^m) \text{ with } (x^1, \dots, x^m, g) \in R_\epsilon(x', x, \delta)\}.$$

LEMMA 3.2. For $\epsilon > 0$, let ρ_ϵ be as defined in (1), and let $m \geq n + 1$, $\delta > 0$, and $\bar{x} \in \mathbf{R}^n$ be given with $0 < \rho_\epsilon(\bar{x})$.

- (i) There is a $\tau > 0$ such that the set $V_\epsilon(\bar{x}, x, \delta)$ contains a nonempty open subset whenever $\|x - \bar{x}\| \leq \tau$.
- (ii) $\tau > 0$ may be chosen so that there is a nonempty open set $\bar{V} \subset V_\epsilon(\bar{x}, \bar{x}, \delta)$ such that $\bar{V} \subset V_\epsilon(\bar{x}, x, \delta)$ for all $x \in \bar{x} + \tau\mathbb{B}$.
- (iii) By definition, we have $U_\epsilon(\bar{x}, x, \delta) \subset G_\epsilon(x)$ for all $x \in \mathbf{R}^n$, and so

$$\rho_\epsilon(x) \leq \|u\| \leq \rho_\epsilon(\bar{x}) + \delta \quad \forall u \in U_\epsilon(\bar{x}, x, \delta), \quad x \in \mathbf{R}^n.$$

In addition, if $\tau > 0$ is as given by statement (i) or (ii), then the set $U_\epsilon(\bar{x}, x, \delta)$ is guaranteed to be nonempty whenever $\|x - \bar{x}\| \leq \tau$.

- (iv) The function ρ_ϵ is upper semicontinuous, i.e.,

$$\limsup_{x \rightarrow \bar{x}} \rho_\epsilon(x) \leq \rho_\epsilon(\bar{x}).$$

- (v) Let $\eta > 0$. Then for every compact subset K of the set $\mathcal{L}$ for which $\inf_K \rho_\epsilon(x) \geq \eta$ there exist $\bar{\mu} > 0$, an integer $\ell \geq 1$, $\tau_j \in (0, \epsilon/3)$, $j = 1, \dots, \ell$, and a set of points $\{z^1, z^2, \dots, z^\ell\} \subset K$ such that the union $\cup_{j=1}^\ell (z^j + \tau_j \text{int } \mathbb{B})$ is an open cover of K and for each $j = 1, \dots, \ell$ there is a point $(z^{j1}, \dots, z^{jm}) \in V_\epsilon(z^j, z^j, \delta)$ such that

$$(z^{j1}, \dots, z^{jm}) + \bar{\mu}\mathbb{B}^m \subset V_\epsilon(z^j, x, \delta) \text{ whenever } x \in z^j + \tau_j\mathbb{B}.$$

Proof. Let $u \in \text{conv} \{\nabla f(x) \mid x \in (\bar{x} + \epsilon\mathbb{B}) \cap D\}$ be such that

$$\|u\| < \rho_\epsilon(\bar{x}) + \delta.$$

Then Carathéodory's theorem [Roc70] implies the existence of $(\bar{x}^1, \dots, \bar{x}^m) \in D_\epsilon^m(\bar{x})$ and $\bar{\lambda} \in \mathbf{R}_+^m$ with $\sum_{j=1}^m \bar{\lambda}_j = 1$ such that $u = \sum_{j=1}^m \bar{\lambda}_j \nabla f(\bar{x}^j)$. Since f is continuously differentiable on the open set D , there is an $\epsilon_0 > 0$ such that f is continuously differentiable on $\bar{x}^j + \epsilon_0 \text{int } \mathbb{B} \subset \bar{x} + \epsilon\mathbb{B}$ for $j = 1, \dots, m$. Define $F : (\bar{x}^1 + \epsilon_0 \text{int } \mathbb{B}) \times \dots \times (\bar{x}^m + \epsilon_0 \text{int } \mathbb{B}) \rightarrow \mathbf{R}^n$ by $F(x^1, \dots, x^m) = \sum_{j=1}^m \bar{\lambda}_j \nabla f(x^j)$. The mapping F is continuous on $(\bar{x}^1 + \epsilon_0 \text{int } \mathbb{B}) \times \dots \times (\bar{x}^m + \epsilon_0 \text{int } \mathbb{B})$. Next define $U = \{u \in \mathbf{R}^n \mid \|u\| < \rho_\epsilon(\bar{x}) + \delta\}$. Then, by definition, the set

$$V = F^{-1}(U) \cap ((\bar{x}^1 + \epsilon_0 \text{int } \mathbb{B}) \times \dots \times (\bar{x}^m + \epsilon_0 \text{int } \mathbb{B}))$$

is a nonempty open subset of $V_\epsilon(\bar{x}, \bar{x}, \delta)$. Now since the sets $x + \epsilon\mathbb{B}$ converge to the set $\bar{x} + \epsilon\mathbb{B}$ in the Hausdorff metric as $x \rightarrow \bar{x}$, we must have that $V \cap ((x + \epsilon \text{int } \mathbb{B}) \times \dots \times (x + \epsilon \text{int } \mathbb{B}))$ is open and nonempty for all x sufficiently close to $\bar{x}$. This proves statement (i).

To see that statement (ii) is true, observe that since $V_\epsilon(\bar{x}, \bar{x}, \delta)$ contains a nonempty open subset $\tilde{V}$ there must exist an $\tilde{\epsilon} \in (0, \epsilon)$ such that

$$\bar{V} = \tilde{V} \cap ((\bar{x} + \tilde{\epsilon} \text{int } \mathbb{B}) \times \dots \times (\bar{x} + \tilde{\epsilon} \text{int } \mathbb{B}))$$

is open and nonempty. Since the Hausdorff distance between $x + \epsilon\mathbb{B}$ and $\bar{x} + \epsilon\mathbb{B}$ is $\|x - \bar{x}\|$ whenever $\|x - \bar{x}\| < \epsilon/2$, we have that $\bar{V} \subset V_\epsilon(\bar{x}, x, \delta)$ whenever $\|x - \bar{x}\| \leq (\epsilon - \tilde{\epsilon})/2 = \tau$. This proves statement (ii).

Statement (iii) follows immediately from statement (ii), while statement (iv) follows from statement (iii) by letting $\delta \rightarrow 0$.

Statement (ii) and compactness imply statement (v). Indeed, statement (ii) implies that for each $x' \in K$ there is a $\tau(x') \in (0, \epsilon/3)$ and a nonempty open set

$V(x') \subset V_\epsilon(x', x', \delta)$ such that $V(x') \subset V_\epsilon(x', x, \delta)$ whenever $x \in x' + \tau(x')\mathbb{B}$. The sets $x' + \tau(x') \text{int } \mathbb{B}$ form an open cover of K . Since K is compact, this open cover contains a finite subcover $z^j + \tau_j \text{int } \mathbb{B}, j = 1, \dots, \ell$. Let $(z^{j_1}, \dots, z^{j_m}) \in V_\epsilon(z^j, z^j, \delta)$ and $\bar{\mu}_j > 0$ be such that $(z^{j_1}, \dots, z^{j_m}) + \bar{\mu}_j \mathbb{B}^m \subset V(z^j)$ for $j = 1, \dots, \ell$. By setting $\bar{\mu} = \min\{\bar{\mu}_1, \dots, \bar{\mu}_\ell\}$ we obtain the result. $\square$

We also need the following mean value inequality [Cla83, Theorem 2.3.7].

THEOREM 3.3 (Lebourg mean value theorem). *Let $f : \mathbf{R}^n \rightarrow \mathbf{R}$ be locally Lipschitz and let $x, y \in \mathbf{R}^n$. Then there exist $z \in [x, y]$ and $w \in \bar{\partial}f(z)$ such that*

$$f(y) - f(x) = \langle w, y - x \rangle.$$

The main convergence result follows.

THEOREM 3.4 (convergence for fixed sampling radius). *If $\{x^k\}$ is a sequence generated by the GS algorithm with $\epsilon_0 = \epsilon$, $\nu_0 = 0$, and $\mu = 1$, then with probability 1 either the algorithm terminates finitely at some iteration k_0 with $\rho_\epsilon(x^{k_0}) = 0$ or there is a subsequence $J \subset \mathbb{N}$ such that $\rho_\epsilon(x^k) \rightarrow_J 0$ and every cluster point $\bar{x}$ of the subsequence $\{x^k\}_J$ satisfies $0 \in \bar{\partial}_\epsilon f(\bar{x})$.*

Proof. We may assume that event $\mathcal{E}$ occurs (see the discussion at the end of section 2). That is, we may assume that the sequence $\{(u^{k_1}, \dots, u^{k_m})\}$ hits every positive measure subset of $\mathbb{B}^m$ infinitely often. As previously noted, this event occurs with probability 1.

We begin by considering the case where the algorithm terminates finitely. Let $x \in \mathcal{L}$ and $\epsilon > 0$, and let z be a realization of a random variable that is uniformly distributed on $\mathbb{B}$. Then the probability that $x + \epsilon z \notin D$ is zero. Hence, with probability 1 the algorithm does not terminate in Step 1. Therefore, if the algorithm terminates finitely at some iteration k_0 , then with probability 1 it did so in Step 2 with $\rho_\epsilon(x^{k_0}) = 0$.

We now restrict our attention to the set of events $\hat{\mathcal{E}} \subset \mathcal{E}$ where the algorithm does not terminate finitely. For such events, we have that $x^{kj} \in D$ for $j = 0, 1, \dots, m$ and $k = 0, 1, 2, \dots$. Conditioned on $\hat{\mathcal{E}}$ occurring, we show that with probability 1 there is a subsequence $J \subset \mathbb{N}$ such that $\rho_\epsilon(x^k) \rightarrow_J 0$ and every cluster point $\bar{x}$ of the subsequence $\{x^k\}_J$ satisfies $0 \in \bar{\partial}_\epsilon f(\bar{x})$.

Since the sequence $\{f(x^k)\}$ is decreasing and $\mathcal{L}$ is compact, it must be the case that there is a $\kappa > 0$ and an $\hat{f}$ such that κ is a Lipschitz constant for f on all of $\mathcal{L} + \mathbb{B}$ and $f(x^k) \downarrow \hat{f}$. Consequently, $(f(x^{k+1}) - f(x^k)) \rightarrow 0$, and so by Step 3 of the GS algorithm

$$t_k \|g^k\| \rightarrow 0.$$

If the result were false, then with positive probability there is an $\eta > 0$ such that

$$(4) \quad \eta = \inf_{k \in \mathbb{N}} \rho_\epsilon(x^k) \leq \inf_{k \in \mathbb{N}} \|g^k\|,$$

since by definition $\rho_\epsilon(x) = \text{dist}(0 | G_\epsilon(x))$ and $\|g^k\| = \text{dist}(0 | G_k)$. For such an event $t_k \downarrow 0$.

Let us assume that event (4) has occurred, and let K be the set of points $x \in \mathcal{L}$ having $\rho_\epsilon(x) \geq \eta$. Then

$$(5) \quad \{x^k\} \subset K.$$

The set K is closed by part (iv) of Lemma 3.2 and so K is compact since $\mathcal{L}$ is compact. Choose $\delta > 0$ so that

$$(6) \quad \left[\kappa \sqrt{\frac{2}{\eta}} + \sqrt{2\kappa + \delta} \right] \sqrt{\delta} < (1 - \beta) \frac{\eta}{2}.$$

Let $\bar{\mu} > 0$, $\tau_j \in (0, \epsilon/3)$, $z^j \in K$, and $(z^{j1}, \dots, z^{jm}) \in V_\epsilon(z^j, z^j, \delta)$ for $j = 1, \dots, \ell$ be associated with K as in part (v) of Lemma 3.2 for $\epsilon > 0$ and $\delta > 0$ as given above. Thus, in particular,

$$(7) \quad K \subset \bigcup_{j=1}^{\ell} (z^j + \tau_j \text{int } \mathbb{B}).$$

By (5) and (7), there exists for each $k = 1, 2, \dots$ a $j_k \in \{1, 2, \dots, \ell\}$ such that

$$x^k \in z^{j_k} + \tau_{j_k} \text{int } \mathbb{B}.$$

Hence, by part (v) of Lemma 3.2,

$$(z^{j_k1}, \dots, z^{j_km}) + \bar{\mu} \mathbb{B}^m \subset V_\epsilon(z^{j_k}, x^k, \delta) \subset D_\epsilon^m(x^k).$$

Since the x^{ki} ($i = 1, \dots, m$) are uniformly distributed random variables supported on $D_\epsilon^m(x^k)$, it follows that

$$\text{prob}\{x^{ki} \in z^{j_k i} + \bar{\mu} \mathbb{B}\} = \frac{\text{vol}(\bar{\mu} \mathbb{B})}{\text{vol}(\epsilon \mathbb{B})} = \left(\frac{\bar{\mu}}{\epsilon}\right)^n.$$

Hence,

$$\text{prob}\{(x^{k1}, \dots, x^{km}) \in (z^{j_k1}, \dots, z^{j_km}) + \bar{\mu} \mathbb{B}^m\} = \left(\frac{\bar{\mu}}{\epsilon}\right)^{nm}.$$

Consequently, since event (4) has occurred, we have with probability 1 that there is an infinite subsequence $J \subset \mathbb{N}$ and a $\bar{j} \in \{1, \dots, \ell\}$ such that for all $k \in J$

$$(8) \quad \gamma^{-1} t_k < \min \left\{ \gamma, \frac{\epsilon}{3} \right\},$$

$$(9) \quad x^k \in z^{\bar{j}} + \frac{\epsilon}{3} \mathbb{B}, \quad \text{and}$$

$$(x^{k1}, \dots, x^{km}) \in (z^{\bar{j}1}, \dots, z^{\bar{j}m}) + \bar{\mu} \mathbb{B}^m \subset V_\epsilon(z^{\bar{j}}, z^{\bar{j}}, \delta) \subset D_\epsilon^m(z^{\bar{j}}),$$

which implies that for all $k \in J$

$$(10) \quad \{x^{k0}, x^{k1}, \dots, x^{km}\} \subset z^{\bar{j}} + \epsilon \mathbb{B} \quad \text{and}$$

$$(11) \quad \eta \leq \|g^k\| \leq \text{dist} \left(0 \mid \hat{G}_k \right) \leq \rho_\epsilon(z^{\bar{j}}) + \delta,$$

where $\hat{G}_k = \text{conv} \{ \nabla f(x^{k1}), \dots, \nabla f(x^{km}) \}$.

By construction we have that

$$-\gamma^{-1} \beta t_k \|g^k\| \leq f(x^k + \gamma^{-1} t_k d^k) - f(x^k) \quad \forall k \in J.$$

Theorem 3.3 yields for each $k \in J$ the existence of $\tilde{x}^k \in [x^k + \gamma^{-1} t_k d^k, x^k]$ and $v^k \in \bar{\partial} f(\tilde{x}^k)$ such that

$$f(x^k + \gamma^{-1} t_k d^k) - f(x^k) = \gamma^{-1} t_k \langle v^k, d^k \rangle.$$

Since $\|d^k\| = 1$, relations (8) and (9) imply that $\tilde{x}^k \in z^{\bar{j}} + \epsilon\mathbb{B}$ and so $v^k \in G_\epsilon(z^{\bar{j}})$ for all $k \in J$. In addition, the Lipschitz continuity hypothesis implies that $\|v^k\| \leq \kappa$ for all $k \in J$. Hence,

$$-\gamma^{-1}\beta t_k \|g^k\| \leq \gamma^{-1}t_k \langle v^k, d^k \rangle \quad \forall k \in J,$$

or, equivalently,

$$-\beta \|g^k\| \leq -\|g^k\| + \langle v^k - g^k, d^k \rangle \quad \forall k \in J,$$

which in turn implies that

$$(12) \quad 0 \leq (\beta - 1)\eta + \left\langle v^k - g^k, \frac{-g^k}{\|g^k\|} \right\rangle \quad \forall k \in J.$$

By combining (10) and (11) with Lemma 3.1, where $C = G_\epsilon(z^{\bar{j}})$, and then applying the bound (6), we find that

$$\left\langle v^k - g^k, \frac{-g^k}{\|g^k\|} \right\rangle \leq \left[\kappa \sqrt{\frac{2}{\eta}} + \sqrt{2\kappa + \delta} \right] \sqrt{\delta} < (1 - \beta) \frac{\eta}{2}.$$

Plugging this into (12) yields the contradiction

$$0 < \frac{1}{2}(\beta - 1)\eta \quad \forall k \in J.$$

Consequently, the event (4) cannot occur with positive probability, which proves that if $\hat{\mathcal{E}}$ occurs, then with probability 1 there is a subsequence $J \subset \mathbb{N}$ such that $\rho_\epsilon(x^k) \rightarrow 0$.

Finally, if $\bar{x}$ is a cluster point of the subsequence $\{x^k\}_J$ with $J \subset \mathbb{N}$, then

$$\text{dist}(0 \mid \bar{\partial}_\epsilon f(x^k)) \rightarrow_J 0$$

since $0 \leq \text{dist}(0 \mid \bar{\partial}_\epsilon f(x^k)) \leq \rho_\epsilon(x^k) \rightarrow_J 0$. Hence $0 \in \bar{\partial}_\epsilon f(\bar{x})$ for every cluster point $\bar{x}$ of the subsequence $\{x^k\}_J$ since the multifunction $\bar{\partial}_\epsilon f$ is closed. $\square$

Theorem 3.4 tells us that when f has compact level sets then with probability 1 the GS algorithm generates a sequence of iterates having at least one cluster point that is Clarke ϵ -stationary. It is interesting to note that we have deduced this without having shown that the values $\|g^k\| = \text{dist}(0 \mid G_k)$ converge to zero. Indeed, if this sequence of values did converge to zero, then every cluster point of the sequence $\{x^k\}$ would be a Clarke ϵ -stationary point. A convergence result of the type where every cluster point satisfies some kind of approximate stationarity condition is what is usually obtained in the smooth case. We now show that the GS algorithm also has this property under a smoothness hypothesis. In the same vein, if one assumes convexity, then a much stronger result is possible, and we cover this case as well in our next result. This is introduced to provide intuition into the behavior of the GS algorithm in the familiar settings of smoothness and convexity. By no means are we suggesting that the GS algorithm is a useful method when either smoothness or convexity is present.

COROLLARY 3.5. *Let the hypotheses of Theorem 3.4 hold.*

1. *If the function f is everywhere continuously differentiable, then either the GS algorithm terminates finitely at a point $\bar{x}$ for which $0 \in G_\epsilon(\bar{x}) = \bar{\partial}_\epsilon f(\bar{x})$ or the sequences $\{x^k\}$ and $\{\|g^k\|\}$ are such that $\|g^k\| \rightarrow 0$ and every cluster point $\bar{x}$ of the sequence $\{x^k\}$ satisfies $0 \in G_\epsilon(\bar{x}) = \bar{\partial}_\epsilon f(\bar{x})$.*

2. If the function f is convex, then with probability 1 either the GS algorithm terminates finitely at a point $\bar{x}$ for which $0 \in \bar{\partial}_\epsilon f(\bar{x} + \epsilon \mathbb{B})$ or every cluster point $\bar{x}$ of the sequence $\{x^k\}$ satisfies

$$(13) \quad f(\bar{x}) \leq \min_{\mathbf{R}^n} f(x) + 2\kappa\epsilon,$$

where κ is any Lipschitz constant for f on $\mathcal{L}$.

Remark. Condition (13) is equivalent to the condition that $0 \in \partial_{2\kappa\epsilon} f(\bar{x})$, where $\partial_\epsilon f(x)$ denotes the ϵ -subdifferential from convex analysis:

$$\partial_\epsilon f(x) = \{z \mid f(y) \geq f(x) + \langle z, y - x \rangle - \epsilon \ \forall y \in \mathbf{R}^n\}.$$

Proof.

1. If the algorithm terminates finitely at iteration k_0 , then it must do so in Step 2, in which case $0 \in G_\epsilon(\bar{x})$.

Next suppose that the algorithm does not terminate finitely. Further, let us suppose to the contrary that the sequence of values $\|g^k\|$ contains a subsequence $J \subset \mathbb{N}$ that is bounded away from zero:

$$0 < \eta := \inf_J \|g^k\|.$$

Since the sequence $\{f(x^k)\}$ is decreasing and the set $\mathcal{L}$ is compact, we may assume with no loss of generality that there is a point $\bar{x} \in \mathcal{L}$ such that $x^k \rightarrow_J \bar{x}$ and $f(x^k) \downarrow f(\bar{x})$. In addition, we have

$$f(x^{k+1}) - f(x^k) < -t_k \beta \|g^k\|,$$

and so $t_k \|g^k\| \rightarrow 0$. For the subsequence J this implies that $t_k \rightarrow 0$. Thus, we may assume again with no loss of generality that $t_k \downarrow_J 0$ with $t_k < 1$ for all $k \in J$. Hence, for all $k \in J$ there exists $z^k \in [x^k, x^k + \gamma^{-1} t_k d^k]$ such that

$$\begin{aligned} -\gamma^{-1} t_k \beta \|g^k\| &\leq f(x^k + \gamma^{-1} t_k d^k) - f(x^k) \\ &= \gamma^{-1} t_k \langle \nabla f(z^k), d^k \rangle \\ &= \gamma^{-1} t_k \langle \nabla f(x^k), d^k \rangle + \gamma t_k^{-1} \langle \nabla f(z^k) - \nabla f(x^k), d^k \rangle \\ &\leq \gamma^{-1} t_k \left[\sup_{g \in G_k} \langle g, d^k \rangle \right] + \gamma t_k^{-1} \|\nabla f(z^k) - \nabla f(x^k)\| \\ &= -\gamma^{-1} t_k \|g^k\| + \gamma t_k^{-1} \|\nabla f(z^k) - \nabla f(x^k)\|, \end{aligned}$$

where the first inequality follows since $t_k < 1$, the first equality is an application of the mean value theorem, and the third equality follows by construction since d^k solves the problem $\inf_{\|d\| \leq 1} \sup_{g \in G_k} \langle g, d \rangle$. Dividing through by $\gamma^{-1} t_k$ and rearranging yields the inequality

$$0 \leq (\beta - 1) \|g^k\| + \|\nabla f(z^k) - \nabla f(x^k)\| \leq (\beta - 1)\eta + \|\nabla f(z^k) - \nabla f(x^k)\|.$$

Taking the limit over J and using the continuity of ∇f we obtain the contradiction $0 \leq (\beta - 1)\eta < 0$. Hence $\|g^k\| \rightarrow 0$.

The final statement in part 1 of the corollary now follows from the inequality $0 \leq \text{dist}(0 \mid G_\epsilon(\bar{x})) \leq \|g^k\|$ and the continuity of the gradient.

2. Theorem 3.4 states that with probability 1 either the GS algorithm terminates finitely at a point $\bar{x}$ satisfying $0 \in G_\epsilon(\bar{x}) \subset \bar{\partial}_\epsilon f(\bar{x})$ or there is a cluster point $\hat{x}$ that is a Clarke ϵ -stationary point of f . Hence we need only focus on the case where we obtain the cluster point $\hat{x}$. Since the GS algorithm is a descent method we know that $f(\hat{x}) = f(\bar{x})$, where $\bar{x}$ is any cluster point of the sequence generated by the algorithm. Hence we need only show that (13) holds for $\hat{x}$.

Since $0 \in \bar{\partial}_\epsilon f(\hat{x})$, Carathéodory's theorem states that there exist pairs $(z^1, w^1), (z^2, w^2), \dots, (z^{n+1}, w^{n+1}) \in \mathbf{R}^n \times \mathbf{R}^n$ and nonnegative scalars $0 \leq \lambda_j \leq 1$, $j = 1, \dots, n+1$, such that $z^j \in \hat{x} + \epsilon \mathbb{B}$, $w^j \in \bar{\partial} f(z^j)$, $j = 1, \dots, n+1$, $\sum_{j=1}^{n+1} \lambda_j = 1$, and $\sum_{j=1}^{n+1} \lambda_j w^j = 0$. The subdifferential inequality implies that for every $x \in \mathbf{R}^n$

$$(14) \quad f(x) \geq f(z^j) + \langle w^j, x - z^j \rangle, \quad j = 1, \dots, n+1.$$

Let $w \in \partial f(\hat{x})$. Multiply each of the inequalities in (14) by its associated λ_j and sum up. Then

$$\begin{aligned} f(x) &\geq \sum_{j=1}^{n+1} \lambda_j f(z^j) + \sum_{j=1}^{n+1} \lambda_j \langle w^j, x - z^j \rangle \\ &\geq \sum_{j=1}^{n+1} \lambda_j (f(\hat{x}) + \langle w, z^j - \hat{x} \rangle) + \left\langle \sum_{j=1}^{n+1} \lambda_j w^j, x - \hat{x} \right\rangle + \sum_{j=1}^{n+1} \lambda_j \langle w^j, \hat{x} - z^j \rangle \\ &\geq f(\hat{x}) + \sum_{j=1}^{n+1} \lambda_j \langle w^j - w, \hat{x} - z^j \rangle \\ &\geq f(\hat{x}) - \sum_{j=1}^{n+1} \lambda_j (\|w^j\| + \|w\|) \|\hat{x} - z^j\| \\ &\geq f(\hat{x}) - \sum_{j=1}^{n+1} \lambda_j 2\kappa\epsilon \\ &= f(\hat{x}) - 2\kappa\epsilon, \end{aligned}$$

where $w \in \partial f(\hat{x})$ and we have used the fact that κ is a bound on any subgradient at a point in $\mathcal{L}$. $\square$

Our next convergence result is in the spirit of the *sequential unconstrained minimization technique* (SUMT) employed in [FM68]. Here we consider the behavior of the set of cluster points of the GS algorithm in fixed ϵ mode as ϵ is decreased to zero.

COROLLARY 3.6. *Let $\{\epsilon_j\}$ be a sequence of positive scalars decreasing to zero. For each $j = 1, 2, \dots$ consider the GS algorithm with $\nu_0 = 0$ and $\epsilon = \epsilon_j$. Let C_j denote either the point of finite termination of the iterates or, alternatively, the set of all cluster points of the resulting infinite sequence of iterates. By Theorem 3.4, with probability 1 there exists a sequence $\{x^j\}$ with $x^j \in C_j$, $j = 1, 2, \dots$, satisfying $0 \in \bar{\partial}_{\epsilon_j} f(x^j)$ for each $j = 1, 2, \dots$. Then every cluster point of the sequence $\{x^j\}$ is a Clarke stationary point for f . Moreover, if the set $\mathcal{L}$ contains only one Clarke stationary point $\bar{x}$, in which case $\bar{x}$ is the strict global minimizer of f , then with probability 1 the entire sequence $\{x^j\}$ must converge to $\bar{x}$ and $\sup \{\|x - \bar{x}\| \mid x \in C_j\} \rightarrow 0$; that is, the sets C_j converge to the single point $\bar{x}$ in the Hausdorff sense.*

Proof. Let $\{\delta_j\}$ be any positive sequence of scalars decreasing to zero. Since $0 \in \bar{\partial}_{\epsilon_j} f(x^j)$ for each $j = 1, 2, \dots$, Carathéodory's theorem tells us that for each

$j = 1, 2, \dots$ there exists $\{x_1^j, \dots, x_{n+1}^j\} \subset x^j + \epsilon_j \mathbb{B}$, $\{z_1^j, \dots, z_{n+1}^j\} \subset \mathbf{R}^n$ with $z_s^j \in \bar{\partial}f(x_s^j)$, $s = 1, \dots, n+1$, and $\lambda_1^j, \dots, \lambda_{n+1}^j \in \mathbf{R}_+$ with $\sum_{s=1}^{n+1} \lambda_s^j = 1$ such that $\|\sum_{s=1}^{n+1} \lambda_s^j z_s^j\| \leq \delta_j$. Note that we need to choose the δ_j 's strictly positive due to the closure operation in the definition of $\bar{\partial}_\epsilon f$. Let $\hat{x}$ be any cluster point of the sequence $\{x^j\}$ with associated subsequence $J \subset \mathbb{N}$ such that $x^j \rightarrow_J \hat{x}$. By compactness and the upper semicontinuity of $\bar{\partial}f$, we may assume with no loss in generality that there exist $\lambda_1, \dots, \lambda_{n+1} \in \mathbf{R}_+$ with $\sum_{s=1}^{n+1} \lambda_s = 1$ and $z_s \in \bar{\partial}f(\hat{x})$ such that $\lambda_s^j \rightarrow_J \lambda_s$ and $z_s^j \rightarrow_J z_s$, $s = 1, \dots, n+1$. Then $\sum_{s=1}^{n+1} \lambda_s^j z_s^j \rightarrow_J \sum_{s=1}^{n+1} \lambda_s z_s \in \bar{\partial}f(\hat{x})$. But by construction $\|\sum_{s=1}^{n+1} \lambda_s^j z_s^j\| \leq \delta_j$ with $\delta_j \downarrow 0$. Hence it must be the case that $\sum_{s=1}^{n+1} \lambda_s z_s = 0$, which shows that $0 \in \bar{\partial}f(\hat{x})$.

Next assume that the set $\mathcal{L}$ contains only one Clarke stationary point $\bar{x}$. If the sequence $\{x^j\}$ does not converge to $\bar{x}$, then there is a subsequence that remains bounded away from $\bar{x}$. Since the GS algorithm only generates iterates in the compact set $\mathcal{L}$, this subsequence must have a cluster point in $\mathcal{L}$ that differs from $\bar{x}$. Since we have just shown that this cluster point must be a Clarke stationary point of f , we obtain the contradiction that establishes the result.

Finally suppose that $\bar{x}$ is a local minimizer of f . Since any local minimizer is a Clarke stationary point, we have that $\bar{x}$ is the unique local minimizer of f in the level set $\mathcal{L}$ and hence is the unique global minimizer of f in the level set $\mathcal{L}$ as well. Observe that since the GS algorithm is a descent algorithm, we have that $f(x) = f(x^j)$ for all $x \in C_j$ and all $j = 1, 2, \dots$. We have just shown that $x^j \rightarrow \bar{x}$; hence $f(x^j) \rightarrow f(\bar{x})$. Now if the sequence of values $\sup\{\|x - \bar{x}\| \mid x \in C_j\}$ does not converge to zero, then there must be a sequence $\{\hat{x}^j\}$ with $\hat{x}^j \in C_j$ for each $j = 1, 2, \dots$ such that the sequence is bounded away from $\bar{x}$. Due to compactness there is a subsequence $J \subset \mathbb{N}$ such that $\{\hat{x}^j\}$ converges to some $\hat{x} \in \mathcal{L}$. But then $f(\bar{x}) = \lim_J f(x^j) = \lim_J f(\hat{x}^j) = f(\hat{x})$. Therefore, $\hat{x}$ must also be a global minimizer of f on $\mathcal{L}$. Since this global minimizer is unique, we arrive at the contradiction $\hat{x} = \bar{x}$, which proves the result. $\square$

COROLLARY 3.7. *Suppose that the set $\mathcal{L}$ contains a unique Clarke stationary point $\bar{x}$ of f , in which case $\bar{x}$ is the unique global minimizer of f . Then for every $\delta > 0$ there exists an $\bar{\epsilon} > 0$ such that if the GS algorithm is initiated with $\epsilon \in (0, \bar{\epsilon})$, then with probability 1 either the algorithm terminates finitely at a point within a distance δ of $\bar{x}$ or, alternatively, every cluster point of the resulting infinite sequence of iterates is within a distance δ of $\bar{x}$.*

Proof. Suppose the result is false. Then with positive probability there exists $\delta > 0$ and a sequence $\epsilon_j \downarrow 0$ such that the set of cluster points C_j (or the finite termination point) of the sequence of iterates generated by the GS algorithm with $\epsilon = \epsilon_j$ satisfy $\sup\{\|x - \bar{x}\| \mid x \in C_j\} > \delta$. But this contradicts the final statement of Corollary 3.6 whereby the result is established. $\square$

The convergence results stated above describe the behavior of the GS algorithm in fixed ϵ mode. We now give a final convergence result for the algorithm in the case where ϵ_k and ν_k are allowed to decrease.

THEOREM 3.8. *Let $\{x^k\}$ be a sequence generated by the GS algorithm with $\nu_0 > 0$, $\epsilon_0 > 0$, $\mu \in (0, 1)$, and $\theta \in (0, 1)$. With probability 1 the sequence $\{x^k\}$ is infinite. Moreover, if the sequence $\{x^k\}$ converges to some point $\bar{x}$, then, with probability 1, $\nu_k \downarrow 0$ and $\bar{x}$ is a Clarke stationary point for f .*

Proof. We restrict our discussion to instances of the GS algorithm in $\mathcal{E}$. The algorithm terminates finitely only if it terminates in either Step 1 or 2. As previously observed, finite termination in Step 1 has zero probability. Finite termination occurs

in Step 2 if $\nu_k = \|g^k\| = 0$. But by construction, $0 < \nu_k$ for all k if $0 < \nu_0$. Hence the algorithm cannot terminate in Step 2. Therefore, with probability 1 the sequence $\{x^k\}$ is infinite.

Let us first observe that if $\nu_k \downarrow 0$, then every cluster point is a Clarke stationary point due to the upper semicontinuity of $\bar{\partial}f$ and the relation

$$\text{dist}(0 \mid \bar{\partial}_{\epsilon_k} f(x^k)) \leq \text{dist}(0 \mid G_{\epsilon_k}(x^k)) \leq \text{dist}(0 \mid G_k) = \|g^k\| \leq \nu_k.$$

In this case it follows that the entire sequence must converge to $\bar{x}$ since $\bar{x}$ is the unique Clarke stationary point.

Hence, we may as well assume that

$$\bar{\nu} = \inf_k \nu_k > 0.$$

If $\bar{\nu} > 0$, then it must be the case that ν_k and ϵ_k were updated only finitely many times. That is, there is an index k_0 and an $\bar{\epsilon} > 0$ such that

$$\nu_k = \bar{\nu} \quad \text{and} \quad \epsilon_k = \bar{\epsilon} \quad \forall k \geq k_0.$$

This places us in the context of Theorem 3.4 where finite termination almost surely does not occur since $\epsilon_k = \bar{\epsilon}$ and $\|g^k\| > \nu_k = \bar{\nu}$ for all k sufficiently large. Hence the cluster point $\bar{x}$ must be a Clarke $\bar{\epsilon}$ -stationary point of f . By part (i) of Lemma 3.2, there is a $\tau > 0$ and a nonempty open set $\bar{V}$ such that the set $V_{\bar{\epsilon}}(\bar{x}, \bar{\nu}/2)$ contains $\bar{V}$ whenever $\|x - \bar{x}\| \leq \tau$. Now since we assume event $\mathcal{E}$, the sequence $\{u^{k1}, \dots, u^{km}\}$ hits every open subset of the unit ball $\mathbb{B}^m$ infinitely often, or, equivalently, the sequence $\{(\bar{\epsilon}u^{k1}, \dots, \bar{\epsilon}u^{km})\}$ hits every open subset of $\bar{\epsilon}\mathbb{B}^m$ infinitely often. As $x^k \rightarrow \bar{x}$, we have that the sequence $\{x^{k1}, \dots, x^{km}\}$ hits $\bar{V}$ infinitely often. But then it must be the case that $\|g^k\| \leq \bar{\nu}/2$ infinitely often since $\rho_{\bar{\epsilon}}(\bar{x}) = 0$. This is the contradiction that proves the result. $\square$

We end this section with a summary of some open questions.

1. Theorem 3.4 indicates that the algorithm may not terminate. Under what conditions can one guarantee that the GS algorithm terminates finitely?
2. In Theorem 3.4, we show that if the GS algorithm does not terminate finitely, then there exists a subsequence $J \subset \mathbb{N}$ such that $\rho_{\epsilon}(x^k) \rightarrow_J 0$. But we cannot show that the corresponding subsequence g^k converges to 0. Can one show that $\|g^k\| \rightarrow_J 0$? Or, is there a counterexample? We believe that a counterexample should exist.
3. We have successfully applied the GS algorithm in many cases where the function f is not Lipschitz continuous. Is there an analogue of Theorem 3.4 in the non-Lipschitzian case?
4. Is it possible to remove the uniqueness hypothesis in Corollary 3.7?
5. In Theorem 3.8, can one show that all cluster points of the sequence are Clarke stationary points?

4. Numerical results. We have had substantial experience using the GS algorithm to solve a wide variety of nonsmooth, nonconvex minimization problems that arise in practice. In this section we describe some of these problems and present some of the numerical results. As far as we are aware, none of the problems we describe here has been solved previously by any method.

We begin by describing our choices for the parameters defining the GS algorithm as well as changes that must be made to implement it in finite precision. We have

attempted to minimize the discrepancies between the theoretical and implemented versions of the method, but some are unavoidable. The algorithm was implemented in MATLAB, which uses IEEE double precision arithmetic (thus accuracy is limited to about 16 decimal digits). For the solution of the convex quadratic program required in Step 2, we tried several QP solvers; we found MOSEK [Mos03] to be the best of these in terms of reliability, efficiency, and accuracy.

Sample size. Bearing in mind the requirement that the sample size m must be greater than n , the number of variables, we always set $m = 2n$.

Sampling parameters. We used $\epsilon_0 = \mu = 0.1$, thus initializing the sampling radius to 0.1 and reducing it by factors of 0.1. Naturally, appropriate choices depend on problem scaling; these values worked well for our problems.

Optimality tolerance parameters. We used $\nu_0 = 10^{-6}$ and $\theta = 1$, thus fixing the optimality tolerance to 10^{-6} throughout. (We experimented with $\theta < 1$, so that a coarser sampling radius is associated with a coarser optimality tolerance; this reduced the iteration counts for easier problems but led to difficulty on harder ones.)

Line search parameters. We set the backtracking reduction factor γ to the standard choice of 0.5, but we set the Armijo parameter β to the decidedly nonstandard choice of 0. The theoretical analysis requires $\beta > 0$, but in practice, on difficult problems, even a modest “sufficient decrease” test can cause the algorithm to fail prematurely, and we never encountered any convergence difficulty that could be attributed to the choice $\beta = 0$. Setting β to a very small number such as 10^{-16} is, for all practical purposes, equivalent to setting it to zero.

Maximum number of iterations and line search failure. We limited the number of iterations for each sampling radius to 100; once the limit of 100 is reached, the sampling radius is reduced just as if the condition $\|g^k\| \leq \nu_k = 10^{-6}$ in Step 2 were satisfied, so the line search is skipped, and sampling continues with the smaller sampling radius. The smallest sampling radius allowed was 10^{-6} ; instead of reducing it to 10^{-7} , the algorithm terminates. Thus, the total number of iterations is at most 600. Also, the line search may fail to find a lower function value (either because a limit on the number of backtracking steps, namely, 50, is exceeded, or because the computed direction of search is actually not a descent direction). Line search failure is quite common for the more difficult problems and generally indicates that, roughly speaking, the maximum accuracy has been achieved for the current sampling radius. When line search failure occurs, we reduce the sampling radius and continue, just as if the condition $\|g^k\| \leq \nu_k = 10^{-6}$ were satisfied or the limit of 100 iterations were reached.

Skipping the differentiability check. We do not attempt to check whether the iterates lie in the set D where f is differentiable, in either Step 1 or Step 4. This is simply impossible in finite precision and in any case would make life very difficult for the user who provides function and gradient values. The user need not be concerned about returning special values if the gradient is not defined at a point; typically, this happens because a “tie” takes place in the evaluation of the function, and the user may simply break the tie arbitrarily. The justification for this is that, for all practical purposes, in finite precision the set D is never encountered except in contrived, trivial cases.

Ensuring boundedness. In practice it is advisable to terminate the algorithm if an a priori bound on the norm of the iterates x^k is exceeded; we set this bound to 1000, but it was not activated in the runs described here.

All parameters and limits described above are easily changed by users of our

MATLAB implementation.

We now present a selection of numerical results for the GS algorithm, concentrating on problems that have not, to our knowledge, been solved previously. In the tables below, each line corresponds to running the GS algorithm on one problem (one instance of f). Because of the stochastic nature of the algorithm, we ran it 10 times, either from one given starting point, when specified, or from 10 different random starting points, when so indicated; the results shown are for the run achieving the lowest value of f . For each problem class, we display the results for a range of problems, ranging from easiest to hardest, and collected in a single table. Each line of every table displays the final value of f , the total number of iterations for all six sampling radii (the total number of times Step 2 was executed), and an approximate “optimality certificate.” The last deserves a detailed explanation. An approximate optimality certificate consists of two numbers; the first is an “optimality residual norm” $\|g^k\|$, and the second is the value of the sampling radius ϵ_k for which the first value $\|g^k\|$ was achieved. These quantities together provide an estimate of nearness to Clarke stationarity. Instead of simply displaying the final optimality certificate, we show the certificate for the smallest sampling radius ϵ_k for which the test $\|g^k\| \leq \nu_k = 10^{-6}$ was satisfied, or, if it was satisfied for no ϵ_k , simply the final values.

We note that for the problems described in sections 4.1 through 4.4 we think that the local minimizers approximated by the GS algorithm are in fact global minimizers, based on the failure to find other locally minimal optimal values when initializing the algorithm at other starting points.² However, we discuss the difficulty of finding global minimizers in section 4.5.

We also remark that, for comparison purposes, we have attempted to solve the same problems by other methods, particularly the Bundle Trust (BT) FORTRAN code of [SZ92]. It is faster than our code but, in our experience, generally provides less accurate results and is unable to solve any of the harder problems described below. We also experimented with a variety of “direct search” methods which are not intended for nonsmooth problems but are so robust that they are worth trying anyway. Of these, the most successful was the well-known Nelder–Mead method, but it was only able to solve the easier problems with very small size n . An important observation is that a user of the BT code or Nelder–Mead method generally has no way of knowing how good a computed approximation might be in the absence of any kind of local optimality certificate.

The data matrices for the problems discussed in sections 4.2 and 4.5 are available on the Web, as are the computed solutions that we obtained for all the problems.³

4.1. Chebyshev approximation by exponential sums. Our first example is a classical one: Chebyshev approximation. The function to be minimized is

$$f(x) = \sup_{s \in [\ell, u]} |h(s, x)|,$$

where $[\ell, u]$ is any real interval and $h : \mathbf{R} \times \mathbf{R}^n \rightarrow \mathbf{R}$ is any smooth function. To evaluate $f(x)$ for a given $x \in \mathbf{R}^n$, we evaluate $h(\cdot, x)$ on a one-dimensional grid, find the maximum (in absolute value), and use this to initialize a one-dimensional local maximization method, based on successive cubic interpolation using the derivative of h with respect to s , to accurately locate a maximizer, say $\bar{s}$. The finer the grid is, the

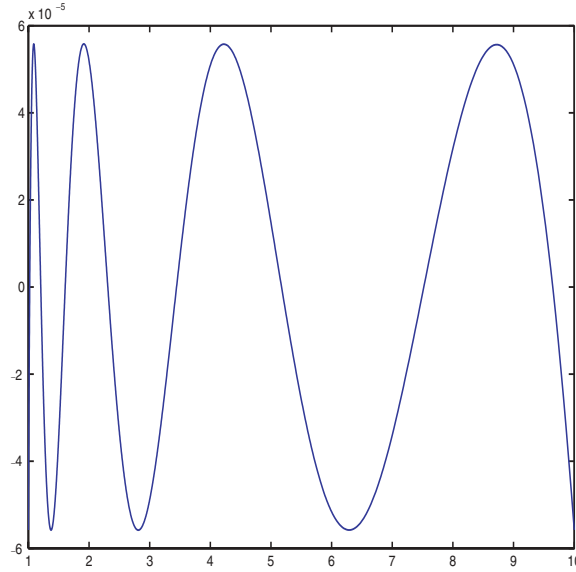
²A possible exception is the problem defined by $N = 8$ in section 4.2.

³<http://www.cs.nyu.edu/overtton/papers/gradsamp/probs/>.

TABLE 1

Results for exponential Chebyshev approximation, starting from $x = 0$.

n	f	Opt cert	Iters
2	8.55641e-002	(9.0e-011, 1.0e-004)	42
4	8.75226e-003	(8.9e-009, 1.0e-006)	63
6	7.14507e-004	(6.5e-007, 1.0e-004)	166
8	5.58100e-005	(2.2e-005, 1.0e-006)	282

FIG. 1. *The error function for the sum of exponentials approximation to $1/s$ on $[1, 10]$ for $n = 8$, with nine alternation points.*

more likely one is to obtain the global maximizer. The function f is differentiable if the maximizer is unique, with gradient

$$\nabla f(x) = \text{sign}(h(\bar{s}, x)) \nabla h_x(\bar{s}, x).$$

We use

$$h(s, x) = \frac{1}{s} - \sum_{j=1}^{n/2} x_{2j-1} \exp(-x_{2j}s),$$

where n is even and $\ell > 0$. Thus the problem is to approximate the function $1/s$ on a positive interval by a sum of decaying exponentials. We chose $[\ell, u] = [1, 10]$ with 2000 grid points, equally spaced in the target function value $1/s$.

Table 1 shows the results obtained by the GS algorithm (see above for interpretation of the optimality certificate “opt cert”). We may safely conjecture on the basis of these results that the optimal value decays exponentially with n . The accuracy achieved is of course limited by the conditioning of the problem and the finite precision being used: accurate solutions were not obtainable for $n > 8$.

Figure 1 shows the error function $h(s, x)$ as a function of s for the optimal parameter vector x found for $n = 8$. Notice the alternation in the error function; the

TABLE 2
Results for minimizing eigenvalue product, using random starting points.

n	N	k	f	Mult	Opt cert	Iters
1	2	1	1.00000e+000	1	(1.3e-008, 1.0e-006)	10
6	4	2	7.46286e-001	2	(7.3e-006, 1.0e-006)	68
15	6	3	6.33477e-001	2	(5.8e-006, 1.0e-006)	150
28	8	4	5.58820e-001	4	(1.0e-001, 1.0e-006)	600
45	10	5	2.17193e-001	3	(1.7e-005, 1.0e-006)	277
66	12	6	1.22226e-001	4	(9.7e-003, 1.0e-006)	432
91	14	7	8.01010e-002	5	(4.5e-006, 1.0e-006)	309
120	16	8	5.57912e-002	6	(2.7e-003, 1.0e-006)	595

maximum error is achieved at nine places on the interval (the leftmost one being essentially invisible). Work of Rice in the 1950s [Mei67] showed that if an optimal approximation exists, such alternation must take place, with the number of alternation points equal to one plus the number of parameters, but computation of such an error function was impossible at the time. A picture like Figure 1 may well have appeared in the more recent literature (perhaps computed by semi-infinite programming), but we have not seen one.

We remark that the optimal approximations seem to be unique up to the obvious permutation of pairs of parameters with one another; the ordering of pairs (x_{2j-1}, x_{2j}) is arbitrary.

4.2. Minimization of eigenvalue products. Our second example is the following problem: minimize the product of the largest k eigenvalues of a Hadamard (componentwise) matrix product $A \circ X$, where A is a fixed positive semidefinite symmetric matrix and X is a variable symmetric matrix constrained to have ones on its diagonal and to be positive semidefinite. Since the latter constraint is convex, we could impose it via projection, but for simplicity we handle it by an exact penalty function. Thus the function to be minimized is

$$f(x) = \prod_{j=1}^k \lambda_j(A \circ X) - \rho \min(0, \lambda_N(X)),$$

where λ_j means j th largest eigenvalue and the $N \times N$ symmetric matrix X has ones on its diagonal and $n = N(N-1)/2$ variables from the vector x in its off-diagonal positions. We set $\rho = 100$. The function f is differentiable at a vector x corresponding to a matrix X if X is positive definite and $\lambda_k(A \circ X) > \lambda_{k+1}(A \circ X)$. The gradient of f at such points is easily computed using the chain rule and the fact that the derivative of a simple eigenvalue λ_j in matrix space is the outer product qq^T defined by its corresponding normalized eigenvector q . As explained earlier, the user coding the gradient need not be concerned about ties, whether these are ties for the choice of k th eigenvalue of $A \circ X$, ties for the ordering of its eigenvalues λ_j for $j < k$, or for the boundary case $\lambda_N(X) = 0$.

Table 2 shows results for various instances of this problem: The matrices A are the leading $N \times N$ submatrices of a specific 63×63 covariance data matrix arising in an environmental application [AL04]. In each case we set k , the number of eigenvalues in the product, to $N/2$. For each minimizer approximated for $N > 2$, the matrix $A \circ X$ has a multiple interior eigenvalue including λ_k ; its multiplicity is shown in the table. In addition, for $N > 4$, the minimizer X has a multiple zero eigenvalue.

The results in Table 2 demonstrate that use of the GS algorithm is by no means

restricted to very small n . Each iteration of the algorithm requires the solution of a quadratic program in m variables, a cost that is a small degree polynomial in n since $m = 2n$. However, solving this problem for $N > 20$ ($n > 200$) would take an unreasonable amount of computer time at present.

4.3. Spectral and pseudospectral minimization. We now enter the realm of nonsymmetric real matrices, whose eigenvalues may be complex and are non-Lipschitz at some points in matrix space. We are interested in a function known as the pseudospectral abscissa of a matrix, $\alpha_\delta(X)$, defined, for any given $\delta \geq 0$, as the maximum of the real parts of the δ -pseudospectrum of X , that is, the set of all z in the complex plane such that z is an eigenvalue of some complex matrix within a distance δ of X [Tre97]. Here, distance is measured in the operator 2-norm. Pseudospectra, and more specifically the pseudospectral abscissa, arise naturally in the study of robust stability of dynamical systems. When $\delta = 0$ the pseudospectral abscissa reduces to the spectral abscissa (the maximum of the real parts of the eigenvalues of the given matrix X). An algorithm for computing α_δ was given by the authors in [BLO03b]. As is the case with so many of the applications we have encountered, computing the function value is quite complicated, but once it is computed, the gradient is easy to obtain where defined, in this case requiring only the computation of singular vectors corresponding to a certain least singular value. As usual, the user coding the gradient need not be concerned with ties for the minimum value.

We consider a simple parameterized matrix,

$$(15) \quad X(x) = \begin{bmatrix} -x_1 & 1 & 0 & \cdot & \cdot & 0 \\ x_1 & 0 & 1 & 0 & \cdot & 0 \\ x_2 & 0 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 \\ x_n & 0 & \cdot & \cdot & \cdot & 0 \end{bmatrix},$$

where n , the number of parameters, is one less than the order of the matrix, say N . Our optimization problem is to minimize

$$(16) \quad f(x) = \alpha_\delta(X(x))$$

over the parameter vector x . The authors showed in [BLO01] that, in the case $\delta = 0$, the global minimizer of f is 0. It is easy to verify in this case that f is not Lipschitz at 0; in fact, f grows proportionally with $|x_n|^{1/N}$. The authors have also shown [BLO03a] that, for fixed small positive δ , f is Lipschitz near 0, but it is not known whether this is true on the whole parameter space.

Figure 2 shows the optimal values of (16) found by the GS algorithm for various values of δ and N . We used $x = 0$ as the starting point since this is the optimal solution for $\delta = 0$. The figure suggests a conjecture: that the optimal value is proportional to $\delta^{2/(N+1)}$ for small δ . The irregularity at the top right of the plot is not numerical error, but a reminder that the phenomenon we are studying is nonlinear. We verified that the function (16) is indeed nonsmooth at all the minimizers approximated by the GS algorithm.

Table 3 shows more detailed results for $N = 5$. Note particularly the final line in the table, which shows the case $\delta = 0$ (minimizing the pure spectral abscissa). Since the solution is $x = 0$, we initialized the runs randomly in this case. Because the exact optimal value of f is 0, the computed value of f necessarily has no correct

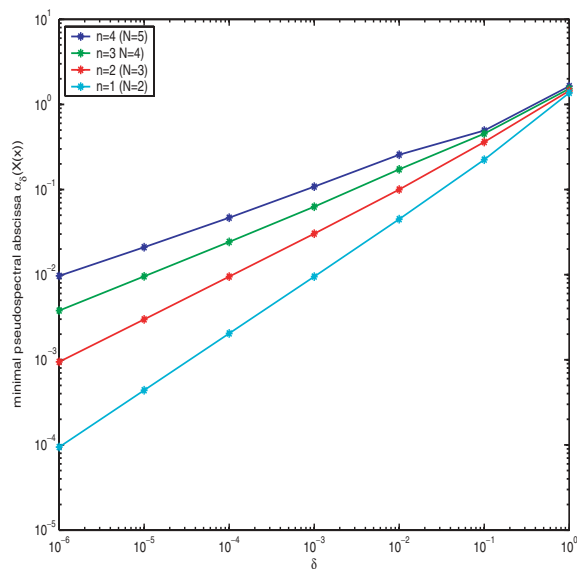


FIG. 2. The minimal pseudospectral abscissa $\alpha_\delta(X(x))$ plotted as a function of δ , for various n .

TABLE 3

Results for minimizing pseudospectral abscissa $\alpha_\delta(X(x))$ for $n = 4$ ($N = 5$), starting from $x = 0$ (except pure spectral abscissa case $\delta = 0$, started randomly).

δ	f	Opt cert	Iters
1	1.63547e+000	(8.4e-007, 1.0e-006)	81
1.0e-001	4.92831e-001	(4.5e-006, 1.0e-006)	105
1.0e-002	2.56467e-001	(7.4e-009, 1.0e-002)	112
1.0e-003	1.08221e-001	(7.6e-008, 1.0e-003)	163
1.0e-004	4.66477e-002	(3.2e-010, 1.0e-005)	236
1.0e-005	2.10125e-002	(3.0e-007, 1.0e-006)	322
1.0e-006	9.68237e-003	(6.3e-007, 1.0e-006)	403
0	4.03358e-003	(3.0e-007, 1.0e-006)	157

digits. However, its order of magnitude is about as good as can be expected using a precision of 16 decimal digits, because the exact spectral abscissa of $X(x)$ has order of magnitude $(10^{-15})^{1/5} = 10^{-3}$ for $\|x\| = 10^{-15}$, the approximate rounding level. This experiment indicates that the GS algorithm has no inherent difficulty with minimizing functions that are non-Lipschitz at their minimizers.

4.4. Maximization of distance to instability. A stable matrix is one with all its eigenvalues in the open left half-plane. The matrix $X(x)$ defined in (15) is not stable for any x , but the shifted matrix $X(x) - sI$ is stable for all $s > 0$ and sufficiently small $\|x\|$. Given a matrix X , its distance to instability, denoted $d_{\text{inst}}(X)$, is the least value δ such that some complex matrix Y within a distance δ of X is not stable. The distance to instability is a well-studied function [Bye88], especially in robust control, where it is known as the complex stability radius (or, more generally, as the inverse of the H_∞ -norm of a transfer function) [BB90]. The relationship between α_δ and d_{inst} is summarized by

$$\alpha_\delta(X) = 0 \quad \text{for} \quad \delta = d_{\text{inst}}(X)$$

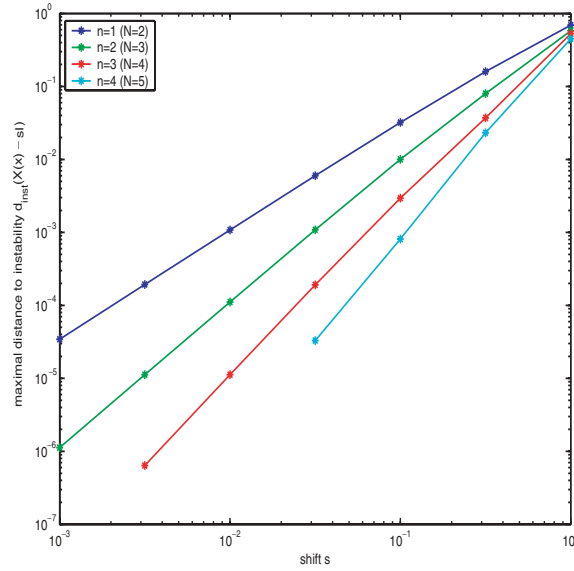


FIG. 3. The maximal distance to instability $d_{\text{inst}}(X(x) - sI)$ plotted as a function of the shift s , for various n .

TABLE 4

Results for maximizing distance to instability $-f(x) = d_{\text{inst}}(X(x) - sI)$ for $n = 4$ ($N = 5$), starting from $x = 0$.

s	f	Opt cert	Iters
1	-4.49450e-001	(1.4e-006, 1.0e-006)	55
3.16228e-001	-2.31760e-002	(1.5e-005, 1.0e-006)	71
1.00000e-001	-8.12170e-004	(8.5e-007, 1.0e-002)	110
3.16228e-002	-3.28692e-005	(1.8e-006, 1.0e-006)	141

for all stable X . By definition, $d_{\text{inst}}(X) = 0$ if X is not stable.

We now consider the problem of maximizing $d_{\text{inst}}(X(x) - sI)$ (equivalently, minimizing $f(x) = -d_{\text{inst}}(X(x) - sI)$) over the parameter vector x , given $s > 0$. This is a difficult problem for small s because the set of x for which $f(x) < 0$ shrinks to 0 as $s \rightarrow 0$. We use the starting point $x = 0$ since $f(0) < 0$ for all $s > 0$. Figure 3 shows the optimal values found by the GS algorithm for various s and N . The missing data points in the table were suppressed because the computed values were too inaccurate to be meaningful. The figure suggests another conjecture, related to the one in the previous subsection: that the optimal value is proportional to $s^{(N+1)/2}$.

Table 4 gives details for $N = 5$.

4.5. Static output feedback and low-order controller design. Suppose the following are given: an $N \times N$ matrix A associated with a dynamical system $\dot{\xi} = A\xi$, together with an $N \times m$ matrix B (defining controllability of the system) and a $p \times N$ matrix C (defining observability of the system), with $m < N$ and $p < N$. Then, given an integer $k < N$, the *order k controller design problem* is to find X_1 , X_2 , X_3 , and X_4 , respectively with dimensions $m \times p$, $m \times k$, $k \times p$, and $k \times k$, such

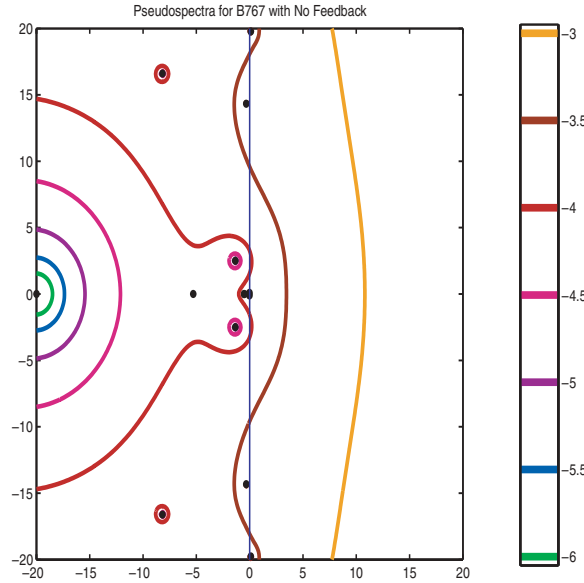


FIG. 4. *Eigenvalues and pseudospectra of B767 model at flutter condition with no controller.*

that the matrix describing the controlled system, namely,

$$(17) \quad \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} B & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} X_1 & X_2 \\ X_3 & X_4 \end{bmatrix} \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix},$$

satisfies desired objectives. We confine our attention to optimizing the following functions of this matrix: asymptotic stability, as measured by α_0 , and robust stability, as measured by d_{inst} , respectively defined in the previous two subsections. When $k = 0$, the controlled system reduces to static (or memoryless) output feedback (SOF), the most basic control model possible, with just mp free variables. Clearly, one may think of the order k controller design problem as an SOF problem with $(m+k)(p+k)$ variables instead of mp , redefining A , B , and C as the larger block matrices in (17).

When k , m , and p are sufficiently large, it is known that stabilization is generically possible and there are various well-known techniques for finding such solutions [Won85, Wil97]. However, for $k, m, p \ll N$, how to efficiently find stabilizing X_1, X_2, X_3, X_4 (or show that this is not possible) is a long-standing open problem in control [BGL95]. The title of this subsection reflects the fact that we are interested only in small k .

We focus on a specific, difficult example, which arises from a model of a Boeing 767 at a flutter condition [Dav90]. The matrix A in this case describes a linearized model when flutter has started to occur; in other words, the plane is flying so fast that the aerodynamic and structural forces are interacting to generate an instability in the system. The matrix A has size $N = 55$, but the controllability and observability matrices B and C have only $m = 2$ columns and $p = 2$ rows, respectively. Figure 4 shows the eigenvalues and δ -pseudospectra boundaries of A as points (solid dots) and curves in the complex plane. (Recall from section 4.3 that the δ -pseudospectrum of A is the set of complex numbers z such that z is an eigenvalue of a complex matrix within a distance δ of A .) The legend on the right shows the values of δ using a log 10 scale. Note the complex conjugate pair of unstable eigenvalues near the top

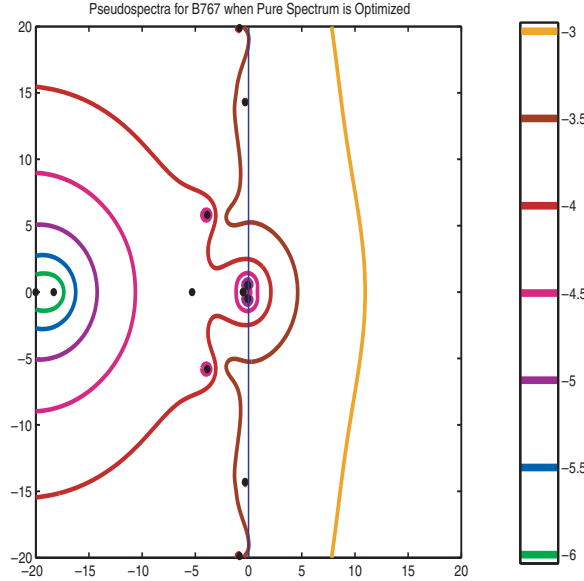


FIG. 5. *Eigenvalues and pseudospectra of B767 model when spectral abscissa is minimized for SOF (order 0 controller).*

and bottom of the figure. This plot and subsequent ones were drawn using Wright's software EigTool [Wri02].

We first investigated the case $k = 0$ (SOF), applying the GS algorithm to minimize the spectral abscissa α_0 of the matrix (17) over the four free parameters (the entries in X_1). Hundreds of runs from randomly chosen starting points repeatedly found the same unstable local minimizers, with $\alpha_0 > 0$. Eventually, however, the GS algorithm found a stable local minimizer, with $\alpha_0 = -7.79 \times 10^{-2}$ and an optimality certificate $(2.8 \times 10^{-15}, 10^{-5})$. The reason it was so difficult to find this minimizer is that the data is very badly scaled. Once it became evident what scaling to use for the starting point, the GS algorithm had no difficulty repeatedly finding this minimizer. Figure 5 shows the eigenvalues and pseudospectra of the stabilized matrix. Although all the eigenvalues are now (barely) to the left of the imaginary axis, even the 10^{-6} -pseudospectrum extends into the right half-plane. Thus, the matrix is not robustly stable: Tiny perturbations to it can generate instability.

We then used this stabilizing minimizer, as well as randomly generated small relative perturbations of it, as starting points for maximizing d_{inst} over the same four variables. The GS algorithm found a local optimizer with $d_{\text{inst}} = 7.91 \times 10^{-5}$ and optimality certificate $(9.2 \times 10^{-7}, 10^{-6})$, whose eigenvalues and pseudospectra are shown in Figure 6. Notice that the 10^{-5} -pseudospectrum now lies in the left half-plane, but that the 10^{-4} -pseudospectrum still extends into the right half-plane.

We now turn to order 1 and order 2 controllers ($k = 1$ and $k = 2$, respectively). We used the local optimizer for $k = 0$ as a starting point, as well as randomly generated small relative perturbations, to maximize the same d_{inst} objective over the 9-variable parametrization for an order 1 controller and the 16-variable parametrization for an order 2 controller. For $k = 1$, the GS algorithm found a local optimizer with $d_{\text{inst}} = 9.98 \times 10^{-5}$ and with optimality certificate $(7.9 \times 10^{-7}, 10^{-4})$, whose eigenvalues and pseudospectra are shown in Figure 7. For $k = 2$, the GS algorithm found a local

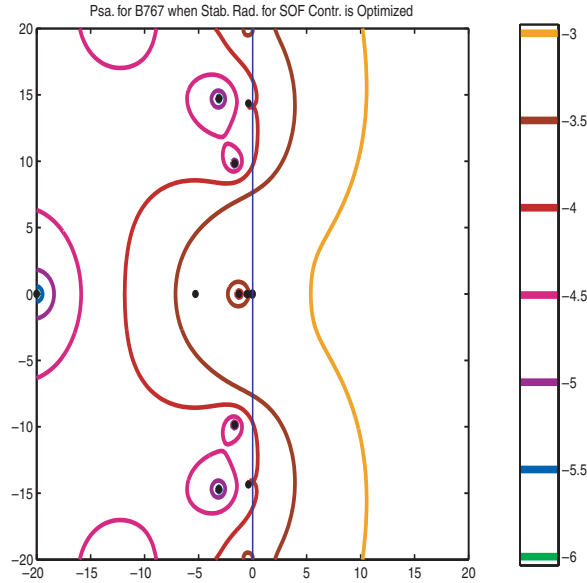


FIG. 6. *Eigenvalues and pseudospectra of B767 model when distance to instability is maximized for SOF model (order 0 controller).*

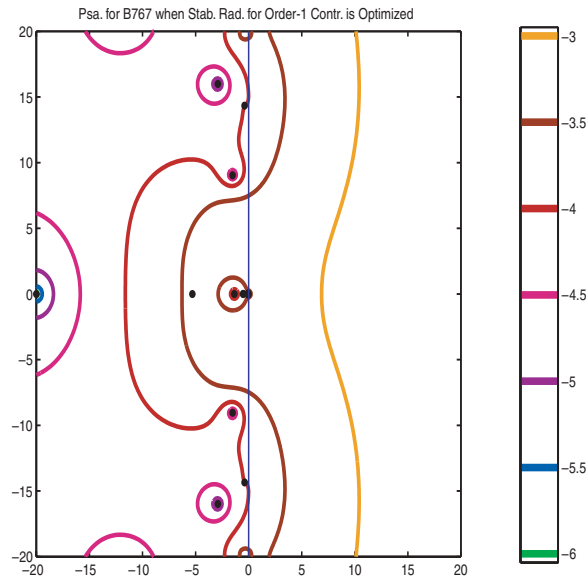


FIG. 7. *Eigenvalues and pseudospectra of B767 model when distance to instability is maximized for order 1 controller.*

optimizer with $d_{\text{inst}} = 1.02 \times 10^{-4}$ and with optimality certificate $(7.3 \times 10^{-6}, 10^{-6})$, whose eigenvalues and pseudospectra are shown in Figure 8. For $k = 1$, the 10^{-4} -pseudospectrum extends just slightly into the right half-plane, while for $k = 2$, it is barely to the left of the imaginary axis, indicating that perturbations of magnitude 10^{-4} or less cannot destabilize the matrix.

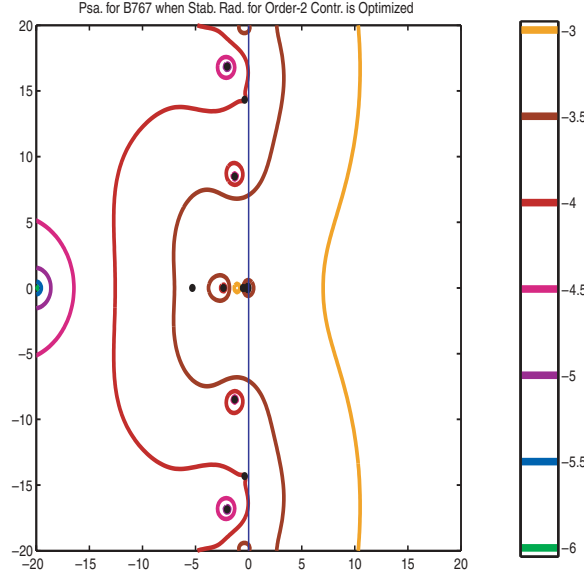


FIG. 8. *Eigenvalues and pseudospectra of B767 model when distance to instability is maximized for order 2 controller.*

As far as we are aware, no such low-order stabilizing controllers were known for the Boeing 767 model before we conducted this work. The optimality certificates that we obtained give us confidence that the optimizers we approximated are indeed local optimizers. However, our initial difficulty in finding even one stabilizing local minimizer of α_0 in the case $k = 0$ illustrates how difficult it is to find global minimizers, and we certainly cannot conclude that the local optimizers we found are global optimizers.

5. Concluding remarks. We have presented a new algorithm for nonsmooth, nonconvex optimization, proved its convergence to Clarke stationary points under strong assumptions, raised questions about other possible convergence results under weaker assumptions, extensively tested the algorithm, presented solutions of quite a number of interesting optimization problems that have not been solved previously, and shown how approximate first-order optimality certificates may be used to give some confidence that the solutions found are meaningful. We make a few final remarks.

All of the functions that we have minimized by the GS algorithm are subdifferentially regular (in the sense of Clarke; see section 1) at the minimizers that we found. We view regularity as a fundamental property that is crucial for the understanding of an optimization problem when smoothness and convexity, both of which are essentially special cases, are lacking. It is regularity that combines with Clarke stationarity to give a genuine first-order optimality condition: that the ordinary directional derivative is nonnegative in all directions. We have been able to show, although we do not give details here, that Chebyshev approximation error, eigenvalue products for symmetric matrices, and minus the distance to instability (on the stable domain) are globally regular functions using results like [RW98, Theorem 10.31] and [Lew99, Corollary 4]. The case of the pseudospectral abscissa, including the pure spectral abscissa, is much more challenging. The authors' theoretical results on regularity of this function may be found in [BO01, BLO03a, Lew02].

Our MATLAB implementation of the GS algorithm is freely available. Furthermore, it is our intention to make publicly available a nonsmooth, nonconvex optimization test set that will include all the problems described here, so that others may use them in the future. For all its power, the GS algorithm is nothing more than a generalized steepest descent method, and will hopefully provide a benchmark against which other algorithms, perhaps more rapidly convergent and efficient, may be compared in the future.

Acknowledgments. We thank F. Leibfritz for the Boeing 767 stabilization problem, K. Anstreicher and J. Lee for the eigenvalue product problem and associated environmental covariance data, J. Kocvara for providing us with the Bundle Trust FORTRAN code, and E. Mengi for writing a MATLAB interface for it. We also thank the referees for carefully reading the manuscript and providing several useful suggestions.

REFERENCES

- [AL04] K. ANSTREICHER AND J. LEE, *A masked spectral bound for maximum-entropy sampling*, in MODA 7—Advances in Model-Oriented Design and Analysis, A. di Bucchianico, H. Läuter, and H. P. Wynn, eds., Springer-Verlag, Berlin, 2004, pp. 1–10.
- [BB90] S. BOYD AND V. BALAKRISHNAN, *A regularity result for the singular values of a transfer matrix and a quadratically convergent algorithm for computing its L_∞ -norm*, Systems Control Lett., 15 (1990), pp. 1–7.
- [BGL95] V. BLONDEL, M. GEVERS, AND A. LINDQUIST, *Survey on the state of systems and control*, Eur. J. Control, 1 (1995), pp. 5–23.
- [BLO01] J. V. BURKE, A. S. LEWIS, AND M. L. OVERTON, *Optimizing matrix stability*, Proc. Amer. Math. Soc., 129 (2001), pp. 1635–1642.
- [BLO02a] J. V. BURKE, A. S. LEWIS, AND M. L. OVERTON, *Approximating subdifferentials by random sampling of gradients*, Math. Oper. Res., 27 (2002), pp. 567–584.
- [BLO02b] J. V. BURKE, A. S. LEWIS, AND M. L. OVERTON, *Two numerical methods for optimizing matrix stability*, Linear Algebra Appl., 351–352 (2002), pp. 117–145.
- [BLO03a] J. V. BURKE, A. S. LEWIS, AND M. L. OVERTON, *Optimization and pseudospectra, with applications to robust stability*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 80–104.
- [BLO03b] J. V. BURKE, A. S. LEWIS, AND M. L. OVERTON, *Robust stability and a criss-cross algorithm for pseudospectra*, IMA J. Numer. Anal., 23 (2003), pp. 359–375.
- [BO01] J. V. BURKE AND M. L. OVERTON, *Variational analysis of non-Lipschitz spectral functions*, Math. Program., 90 (2001), pp. 317–352.
- [Bur85] J. V. BURKE, *Descent methods for composite nondifferentiable optimization problems*, Math. Program., 33 (1985), pp. 260–279.
- [Bye88] R. BYERS, *A bisection method for measuring the distance of a stable matrix to the unstable matrices*, SIAM J. Sci. Stat. Comput., 9 (1988), pp. 875–881.
- [CG78] R. CHANEY AND A. GOLDSTEIN, *An extension of the method of subgradients*, in Nonsmooth Optimization, C. Lemaréchal and R. Mifflin, eds., Pergamon Press, Oxford, UK, New York, 1978, pp. 51–70.
- [Cla73] F. H. CLARKE, *Necessary Conditions for Nonsmooth Problems in Optimal Control and the Calculus of Variations*, Ph.D. thesis, University of Washington, Seattle, WA, 1973.
- [Cla83] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Classics Appl. Math. 5, SIAM, Philadelphia, 1990.
- [CLSW98] F. H. CLARKE, YU. S. LEDYAEV, R. J. STERN, AND P. R. WOLENSKI, *Nonsmooth Analysis and Control Theory*, Grad. Texts in Math. 178, Springer-Verlag, New York, 1998.
- [Dav90] E. J. DAVISON, ED., *Benchmark Problems for Control System Design: Report of the IFAC Theory Committee*, Pergamon Press, Oxford, UK, 1990.
- [DR95] V. F. DEMYANOV AND A. RUBINOV, *Constructive Nonsmooth Analysis*, Approx. Optim. 7, Peter Lang, Frankfurt am Main, Germany, 1995.
- [FG02] A. FUDULI, M. GAUDIOSO, AND G. GIALLOMBARDO, *A DC piecewise affine model and a bundling technique in nonconvex nonsmooth minimization*, Optim. Methods Softw., 19 (2004), pp. 89–102.

- [Fle87] R. FLETCHER, *Practical Methods of Optimization*, 2nd ed., John Wiley, Chichester, UK, New York, 1987.
- [FM68] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, Classics Appl. Math. 4, SIAM, Philadelphia, 1990.
- [Gol77] A. A. GOLDSTEIN, *Optimization of Lipschitz continuous functions*, Math. Program., 13 (1977), pp. 14–22.
- [Gro02] A. GROTHEY, *A Second Order Trust Region Bundle Method for Nonconvex Nonsmooth Optimization*, Technical report MS02-001, University of Edinburgh, Edinburgh, UK, 2002.
- [HUL93] J. B. HIRIART-URRUTY AND C. LEMARÉCHAL, *Convex Analysis and Minimization Algorithms II: Advanced Theory and Bundle Methods*, Springer-Verlag, Berlin, 1993.
- [Kiw85] K. C. KIWIEL, *Methods of Descent for Nondifferentiable Optimization*, Lecture Notes in Math. 1133, Springer-Verlag, Berlin, New York, 1985.
- [Lem75] C. LEMARÉCHAL, *An extension of Davidon methods to non-differentiable problems*, in Nondifferentiable Optimization, M. L. Balinski and P. Wolfe, eds., Math. Programming Stud., 3 (1975), pp. 95–109.
- [Lew99] A. S. LEWIS, *Nonsmooth analysis of eigenvalues*, Math. Program., 84 (1999), pp. 1–24.
- [Lew02] A. S. LEWIS, *Robust regularization*, Math. Program., submitted; available at <http://www.orie.cornell.edu/~aslewis/publications/Robust.ps>.
- [LSB91] C. LEMARÉCHAL, J.-J. STRODIOT, AND A. BIHAIN, *On a bundle algorithm for nonsmooth optimization*, in Nonlinear Programming 4, O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds., Academic Press, New York, 1991, pp. 245–282.
- [LV98] L. LUKŠAN AND J. VLČEK, *A bundle Newton method for nonsmooth unconstrained minimization*, Math. Program., 83 (1998), pp. 373–391.
- [Mei67] G. MEINARDUS, *Approximation of Functions: Theory and Numerical Methods*, Springer Tracts Nat. Philos. 13, Springer-Verlag, New York, 1967.
- [MN92] M. M. MÄKELÄ AND P. NEITTAANMÄKI, *Nonsmooth Optimization*, World Scientific, River Edge, NJ, 1992.
- [Mos03] MOSEK *Optimization Software*, <http://www.mosek.com/>, 2003.
- [OKZ98] J. OUTRATA, M. KOCVARA, AND J. ZOWE, *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints*, Nonconvex Optim. Appl. 28, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.
- [Osb85] M. R. OSBORNE, *Finite Algorithms in Optimization and Data Analysis*, John Wiley, Chichester, UK, New York, 1985.
- [Roc70] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [RW98] R. T. ROCKAFELLAR AND R. J. B. WETS, *Variational Analysis*, Springer-Verlag, New York, 1998.
- [SZ92] H. SCHRAMM AND J. ZOWE, *A version of the bundle idea for minimizing a nonsmooth function: Conceptual idea, convergence analysis, numerical results*, SIAM J. Optim., 2 (1992), pp. 121–152.
- [Tre97] L. N. TREFETHEN, *Pseudospectra of linear operators*, SIAM Rev., 39 (1997), pp. 383–406.
- [Wil97] J. C. WILLEMS, *Generic eigenvalue assignability by real memoryless output feedback made simple*, in Communications, Computation, Control and Signal Processing, A. Paulraj, V. Roychowdhury, and C. D. Schaper, eds., Kluwer Academic Publishers, Norwell, MA, 1997, pp. 343–354.
- [Wol75] P. WOLFE, *A method of conjugate subgradients for minimizing nondifferentiable functions*, in Nondifferentiable Optimization, M. L. Balinski and P. Wolfe, eds., Math. Programming Stud., 3 (1975), pp. 145–173.
- [Won85] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, 3rd ed., Appl. Math. 10, Springer-Verlag, New York, 1985.
- [Wri02] T. G. WRIGHT, *EigTool: A graphical tool for nonsymmetric eigenproblems*, Oxford University Computer Laboratory, Oxford, UK, <http://web.comlab.ox.ac.uk/projects/pseudospectra/eigtool/> (2002).

OPTIMAL INEQUALITIES IN PROBABILITY THEORY: A CONVEX OPTIMIZATION APPROACH*

DIMITRIS BERTSIMAS[†] AND IOANA POPESCU[‡]

Abstract. We propose a semidefinite optimization approach to the problem of deriving tight moment inequalities for $P(X \in S)$, for a set S defined by polynomial inequalities and a random vector X defined on $\Omega \subseteq \mathcal{R}^n$ that has a given collection of up to k th-order moments. In the univariate case, we provide optimal bounds on $P(X \in S)$, when the first k moments of X are given, as the solution of a semidefinite optimization problem in $k + 1$ dimensions. In the multivariate case, if the sets S and Ω are given by polynomial inequalities, we obtain an improving sequence of bounds by solving semidefinite optimization problems of polynomial size in n , for fixed k .

We characterize the complexity of the problem of deriving tight moment inequalities. We show that it is NP-hard to find tight bounds for $k \geq 4$ and $\Omega = \mathcal{R}^n$ and for $k \geq 2$ and $\Omega = \mathcal{R}_+^n$, when the data in the problem is rational. For $k = 1$ and $\Omega = \mathcal{R}_+^n$ we show that we can find tight upper bounds by solving n convex optimization problems when the set S is convex, and we provide a polynomial time algorithm when S and Ω are unions of convex sets, over which linear functions can be optimized efficiently. For the case $k = 2$ and $\Omega = \mathcal{R}^n$, we present an efficient algorithm for finding tight bounds when S is a union of convex sets, over which convex quadratic functions can be optimized efficiently.

Key words. probability bounds, Chebyshev inequalities, semidefinite optimization, convex optimization

AMS subject classifications. 60E15, 90C22, 90C25

DOI. 10.1137/S1052623401399903

1. Introduction. The problem of deriving bounds on the probability that a certain random variable belongs in a given set, given information on some of its moments, has a very rich and interesting history, which is very much connected with the development of probability theory in the twentieth century. The inequalities due to Markov, Chebyshev, and Chernoff are some of the classical and widely used results of modern probability theory. Natural questions, however, that arise are the following:

1. *Are such bounds “best possible”; i.e., do there exist distributions that match them?*
2. *Can such bounds be generalized in multivariate settings, and in what circumstances can they be explicitly and/or algorithmically computed?*
3. *Is there a general theory based on optimization methods to address in a unified manner moment-inequality problems in probability theory?*

In this paper, we formulate the problem of obtaining best possible bounds as an optimization problem and use modern optimization theory, in particular convex and semidefinite programming, to give concrete answers to the questions above. We first introduce some notation. Let $\kappa = (k_1, \dots, k_n)'$ with $k_j \in \mathcal{Z}_+$ nonnegative integers. We use the notation $\sigma_\kappa = \sigma_{k_1, \dots, k_n}$ and $J_k = \{\kappa = (k_1, \dots, k_n)' \mid k_1 + \dots + k_n \leq k, k_j \in \mathcal{Z}_+, j = 1, \dots, n\}$. We next introduce the notion of a *feasible moment sequence*.

*Received by the editors December 18, 2001; accepted for publication (in revised form) June 29, 2004; published electronically April 8, 2005.

<http://www.siam.org/journals/siopt/15-3/39990.html>

[†]Sloan School of Management, Rm. E53-363, Massachusetts Institute of Technology, Cambridge, MA 02139 (dbertsim@mit.edu). The research of this author was partially supported by NSF grant DMI-9610486 and the MIT-Singapore Alliance.

[‡]INSEAD, Fontainebleau Cedex 77305, France (ioana.popescu@insead.edu).

DEFINITION 1.1. A sequence $\bar{\sigma} = (\sigma_\kappa, \kappa \in J_k)$ is a feasible (n, k, Ω) -moment vector (or sequence), if there is a random vector $X = (X_1, \dots, X_n)'$ defined on $\Omega \subseteq \mathcal{R}^n$ endowed with its Borel sigma algebra of events, whose moments are given by $\bar{\sigma}$, that is, $\sigma_\kappa = \sigma_{k_1, \dots, k_n} = E[X_1^{k_1} \dots X_n^{k_n}]$ for all $\kappa \in J_k$. We say that any such random variable X has a $\bar{\sigma}$ -feasible distribution and denote this as $X \sim \bar{\sigma}$.

Throughout the paper, the underlying probability space is implicitly assumed to be $\Omega \subseteq \mathcal{R}^n$ endowed with its Borel sigma algebra of events. We denote by $\mathcal{M} = \mathcal{M}(n, k, \Omega)$ the set of feasible (n, k, Ω) -moment vectors. For the univariate case ($n = 1$), the problem of deciding if $\bar{\sigma} = (M_1, M_2, \dots, M_k)'$ is a feasible $(1, k, \Omega)$ -moment vector is the classical moment problem. This problem has been completely characterized by necessary and sufficient conditions by Stieltjes [43], [44] in 1894-95, who adopts the “moment” terminology from mechanics (see also Karlin and Shapley [18], Akhiezer [1], Siu, Sengupta, and Lind [38], and Kemperman [21]). For univariate, nonnegative random variables ($\Omega = \mathcal{R}_+$), these conditions can be written as $R_k \succeq 0$ and $R_{k-1} \succeq 0$, where for any integer $l \geq 0$ we define

$$R_{2l} = \begin{pmatrix} 1 & M_1 & \dots & M_l \\ M_1 & M_2 & \dots & M_{l+1} \\ \vdots & \vdots & \ddots & \vdots \\ M_l & M_{l+1} & \dots & M_{2l} \end{pmatrix}, \quad R_{2l+1} = \begin{pmatrix} M_1 & M_2 & \dots & M_{l+1} \\ M_2 & M_3 & \dots & M_{l+2} \\ \vdots & \vdots & \ddots & \vdots \\ M_{l+1} & M_{l+2} & \dots & M_{2l+1} \end{pmatrix}.$$

For univariate random variables with $\Omega = \mathcal{R}$, the necessary and sufficient condition (given by Hamburger [12], [13] in 1920-21) for a vector $\bar{\sigma} = (M_1, M_2, \dots, M_k)'$ to be a feasible $(1, k, \mathcal{R})$ -moment sequence is that $R_{2\lfloor \frac{k}{2} \rfloor} \succeq 0$. In the multivariate case, the formulation of the problem can be traced back to Haviland [14], [15] in 1935-36 (see also Godwin [10]). To date, the sufficiency part of the moment problem has not been completely resolved in the multivariate case, although substantial progress has been made in the last decade by Schmudgen [37] and Putinar [34].

Suppose that $\bar{\sigma}$ is a feasible moment sequence and X has a $\bar{\sigma}$ -feasible distribution. We now define the central problem that this paper addresses.

The (n, k, Ω) -bound problem. Given a sequence $\sigma_\kappa, \kappa = (k_1, k_2, \dots, k_n)' \in J_k$ of up to k th-order moments

$$\sigma_\kappa = E[X_1^{k_1} X_2^{k_2} \dots X_n^{k_n}], \quad \kappa \in J_k$$

of a random vector $X = (X_1, X_2, \dots, X_n)'$ on $\Omega \subseteq \mathcal{R}^n$ endowed with its Borel sigma algebra, find the “best possible” or “tight” upper and lower bounds on $P(X \in S)$, for measurable events $S \subseteq \Omega$.

The term “best possible” or “tight” upper (and by analogy lower) bound above is defined as follows.

DEFINITION 1.2. We say that γ is a tight upper bound on $P(X \in S)$ if $\gamma = \sup_{X \sim \bar{\sigma}} P(X \in S)$.

Note that a bound can be tight without necessarily being *exactly achievable* (i.e., there is a random variable $X^{(0)} \sim \bar{\sigma}$ for which $P(X^{(0)} \in S) = \gamma$), but only *asymptotically* (i.e., there exists a sequence of random variables $\{X^{(i)}\}_{i \geq 1}$ such that $X^{(i)} \sim \bar{\sigma}$ and $\lim_{i \rightarrow \infty} P(X^{(i)} \in S) = \gamma$).

The well-known inequalities due to Markov, Chebyshev, and Chernoff, which are widely used if we know the first moment, the first two moments, and all moments (i.e., the generating function) of a random variable, respectively, are feasible but not necessarily *optimal* solutions to the (n, k, Ω) -bound problem; i.e., they are not necessarily tight bounds.

Literature and historical perspective. The history of the developments in the area of (n, k, Ω) -bound problems, sometimes referred to as generalized Chebyshev inequalities, can be traced back to the work of Gauss, Cauchy, Chebyshev, Markov, etc., and has witnessed an unexpected evolution. The problem of finding bounds on univariate distributions under moment constraints was proposed and formulated without proof initially by Chebyshev [6] in 1874 and resolved ten years later by his student Markov [25] in his Ph.D. thesis, using continued fractions techniques. In the 1950s and 1960s there was a revival of the interest in this area, that resulted in a large literature on Chebyshev systems and inequalities. Surveys of early literature can be found in Shohat and Tamarkin [40] and Godwin [9], [10]. A detailed, unified account of the evolution of Chebyshev systems is given by Karlin and Studden [19] in their 1966 monograph (see in particular Chapters 12 and 13 that deal with (n, k, Ω) -type bounds). In particular, they summarize the results of Marshall and Olkin [27], [28] who computed tight, explicit bounds on probabilities given first- and second-order moments (the $(n, 2, \Omega)$ problem in our context), thus generalizing Chebyshev's inequality to a multivariate setting.

Not for the first time in its history, and in the words of Shohat and Tamarkin in 1943 [40, p. 10], “the problem of moments lay dormant for more than 20 years.” It revived briefly in the 1980s, with the book on probability inequalities and multivariate distributions of Tong [45] in 1980, who also published a monograph on probability inequalities in 1984. The latter notably contains, among others, a generalization of Markov's inequality for multivariate tails, due to Marshall [26]. A volume in *Moments in Mathematics* edited by Landau in 1987 includes a background survey by the same author [23], as well as relevant papers of Kemperman [22] and Diaconis [7].

The idea that optimization methods and duality theory can be used to address moment-type inequalities in probability first appeared in 1960, and is due independently and simultaneously to Isii [16] and Karlin (lecture notes at Stanford, see [19, p. 472]), who showed via strong duality results that certain types of Chebyshev inequalities for univariate random variables are sharp. Isii [17] extended these duality results for random vectors on complete regular spaces. Thirty two years after Isii's [17] original multivariate proof, Smith [42] rederived the duality result and proposed new interesting applications in decision analysis, dynamic programming, statistics, and finance. Shapiro [39] provided a rigorous discussion of necessary topological conditions for strong duality to hold, in that sense relaxing the compactness assumptions underlying Isii's proof. For an in-depth account of strong duality and sensitivity analysis for a general class of semi-infinite programming problems see Bonnans and Shapiro [3, Section 5.4].

For a broader investigation of the optimization framework underlying this type of problem, we refer the interested reader to Borwein and Lewis [4], [5] who provide an in-depth analysis of partially finite convex programming.

Despite its long and scattered history, the common belief among researchers is still that “the theory [of moment problems] is not up to the demands of applications” (Diaconis [7, p. 129]). The same author suggested that one of the reasons could be the high complexity of the problem: “numerical determination . . . is feasible for a small number of moments, but appears to be quite difficult in general cases.” Another reason, as identified by Kemperman (see [22, p. 20]), is the lack of a general algorithmic approach:

“... a deep study of algorithms has been rare so far in the theory of moments, except for certain very specific practical applications, for

instance, to crystallography, chemistry and tomography. No doubt, there is a considerable need for developing reasonably good numerical procedures for handling the great variety of moment problems which do arise in pure and applied mathematics and in the sciences in general”

In an attempt to address Kemperman’s criticism, Smith [42] actually introduced a computational procedure for the $(n, k, \mathcal{R}^n)$ -bound problem, although he does not refer to it in this way. Unfortunately, the procedure is far from a formal algorithm, as there is no proof of convergence, and no investigation (theoretical or experimental) of its efficiency. It is fair to say that a thorough understanding of the algorithmic aspects and the complexity of the (n, k, Ω) -bound problem is still lacking.

Yet another criticism brought by Smith is the lack of simple, closed-form solutions for the $(n, k, \mathcal{R}^n)$ -bound problem: “the bounds given by Chebyshev’s inequalities . . . are quite loose. The more general versions are rarely used because of the lack of simple closed-form expressions for the bounds” (see [42, p. 808]).

Goals and contributions. The previous discussion motivates our desire in the present paper to understand the complexity of deriving tight moment inequalities, search for efficient algorithms in a general framework, and, when possible, derive simple closed-form bounds. In particular, we characterize which classes of (n, k, Ω) -bound problems are efficiently solvable and which are NP-hard.

Let us remark that the theory of polynomial solvability and NP-hardness assume a rational computational model. For this reason, as far as complexity results are concerned, we work under the assumption that the problem data is rational; i.e., moments are rational numbers and the sets Ω and S are specified in terms of rational numbers. Specifically, they are defined by semialgebraic sets (i.e., sets that are defined in terms of polynomial inequalities), whose parameters are rational numbers. In addition, throughout the paper we refer to an efficient, or polynomial time algorithm when it takes polynomial time in the problem data and $\log(1/\epsilon)$, and it computes a bound within ϵ of the tight bound for all $\epsilon > 0$.

More concretely, the contributions and structure of the paper are as follows.

1. In section 3, we investigate in detail the univariate case, i.e., the $(1, k, \Omega)$ -bound problem for $\Omega = \mathcal{R}, \mathcal{R}_+$. We show that tight bounds can be computed efficiently by solving a single semidefinite optimization problem. We also derive tight bounds for tail probability events in closed form, when up to three moments are given. For $k = 1$, we recover the Markov inequality, which also shows that the Markov bound is tight. For $k = 2$, we recover a strict improvement of the Chebyshev inequality that retains the simplicity of the bound. This inequality dates back at least to Uspensky’s book (see [46, p. 198]) from 1937, who proposed it as an exercise. Despite its simplicity, the bound has been strangely ignored in the recent literature and textbooks. For $k = 3$, we present closed-form tight bounds that appear to be new.

2. In section 4, we generalize the results of the previous section to the multivariate (n, k, Ω) -bound problem. For a fairly general class of semialgebraic sets S and Ω (not necessarily convex), we propose a sequence of increasingly stronger, asymptotically exact upper bounds by solving semidefinite optimization problems of polynomial size in n . This includes the case when Ω is a bounded polyhedron, or a semialgebraic set such that $\Omega \subseteq \{x \in \mathcal{R}^n \mid x'x \leq M^2\}$ and M is known a priori. The proposed approach gives rise to a family of semidefinite relaxations whose size is polynomial in n . We also show that the $(n, k, \mathcal{R}_+^n)$, $(n, k, \mathcal{R}^n)$ -bound problems for $k \geq 2$, respectively $k \geq 4$, are NP-hard.

TABLE 1
The landscape of the (n, k, Ω) -problem (BP refers to the current paper).

	$n = 1$	n	
		S convex set	S union of convex sets
$k = 1$	Theorem 3.3 Markov [25]	Theorem 5.1 Marshall [26], BP	Algorithm A, section 5.2 BP
$k = 2$	Theorem 3.3 Chebyshev [6]	Theorem 6.1 Marshall and Olkin [27]	Algorithm B, section 6.2 BP
k	Theorem 3.2: SDP BP	Theorem 4.3: SDP; Propositions 4.5, 4.6: NP-hardness BP	

3. In section 5, we address the $(n, 1, \Omega)$ problem. We show that if the set S in the definition of the $(n, 1, \mathcal{R}^n)$ -bound problem is convex, we find best possible bounds for $P(X \in S)$ explicitly as a solution of n convex optimization problems. We also provide a polynomial time algorithm to solve the $(n, 1, \Omega)$ problem, when the sets S and Ω are the union of disjoint convex sets, over which a linear function can be efficiently optimized. These bounds represent natural extensions of the inequalities due to Markov in the univariate case, and Marshall [26] for multivariate tails.

4. In section 6, we first review the work of Marshall and Olkin [27] who showed that the $(n, 2, \mathcal{R}^n)$ -bound problem can be solved as a single convex optimization problem. We provide an efficient algorithm for the case when S is a union of disjoint convex sets, over which a convex quadratic function can be optimized efficiently.

Table 1 summarizes the contributions of the paper to the (n, k, Ω) -bound problem.

2. Primal and dual formulations of the (n, k, Ω) -bound problem. In this section, we formulate the (n, k, Ω) -upper bound problem as an optimization problem, where $\Omega \in \mathcal{R}^n$ is the domain of the random variables under consideration. We examine the corresponding dual problem and present weak and strong duality results that permit us to develop algorithms for the problem. The same approach and results apply to the (n, k, Ω) -lower bound problem.

We use the notation $\bar{z}^\kappa = z_1^{k_1} \cdots z_n^{k_n}$, where $\bar{z} = (z_1, \dots, z_n)'$ and $\kappa = (k_1, \dots, k_n)'$, $k_j \in \mathcal{Z}_+$. Recall that $J_k = \{\kappa = (k_1, \dots, k_n)' \mid k_1 + \cdots + k_n \leq k, k_j \in \mathcal{Z}_+, j = 1, \dots, n\}$. The (n, k, Ω) -upper bound problem can be formulated as the following optimization problem:

$$(2.1) \quad \begin{aligned} Z_P &= \max_{\mu} \int_S \mathbf{1} d\mu \\ \text{subject to} \quad &\int_{\Omega} \bar{z}^\kappa d\mu = \sigma_\kappa \quad \forall \kappa \in J_k. \end{aligned}$$

The variable of the infinite-dimensional optimization problem (2.1) is the probability measure μ (so implicitly $d\mu(\bar{z}) \geq 0$ a.e.). We assume that $\sigma_{0, \dots, 0} = 1$, corresponding to the probability mass constraint. If problem (2.1) is feasible, then $\bar{\sigma}$ is a feasible moment sequence, and any feasible distribution μ is a $\bar{\sigma}$ -feasible distribution. The feasibility problem is exactly the classical multidimensional moment problem.

In the spirit of linear programming duality theory, we associate a dual variable $y_\kappa = y_{k_1, \dots, k_n}$ with each equality constraint of the primal and we obtain

$$(2.2) \quad \begin{aligned} Z_D &= \min_{y \in \mathcal{R}^{|J_k|}} \sum_{\kappa \in J_k} y_\kappa \sigma_\kappa \\ \text{subject to} \quad &g(\bar{z}) = \sum_{\kappa \in J_k} y_\kappa \bar{z}^\kappa \geq 1 \quad \forall \bar{z} \in S, \end{aligned}$$

$$g(\bar{z}) = \sum_{\kappa \in J_k} y_{\kappa} \bar{z}^{\kappa} \geq 0 \forall \bar{z} \in \Omega.$$

While in principle the dual constraints may only need to hold a.e., this is equivalent to solving (2.2) since polynomial positivity holds a.e. if and only if it holds everywhere.

In general, the optimum in problem (2.1) may not be achievable. Whenever the primal optimum is achieved, we call the corresponding distribution *an extremal distribution*. In the case when only (say upper) bounds on moments σ_{κ} are known, then in the dual problem we add the inequalities $y_{\kappa} \geq 0$. Thus, our results (specifically Theorems 3.2 and 4.3 below) regarding efficient solvability of the underlying problem generalize to the case when only bounds on moments are known. We next establish weak duality.

THEOREM 2.1 (weak duality). $Z_P \leq Z_D$.

Proof. Let μ be a primal feasible solution and let y_{κ} , $\kappa \in J_k$, be a dual feasible solution. From dual feasibility we have that for all $\bar{z} \in \Omega$, $g(\bar{z}) = \sum_{\kappa \in J_k} y_{\kappa} \bar{z}^{\kappa} \geq \chi_S(\bar{z})$, where $\chi_S(\bar{z}) = 1$ if $\bar{z} \in S$, and 0, otherwise. Then

$$Z_P = \int_S \mathbf{1} d\mu = \int_{\Omega} \chi_S(\bar{z}) d\mu \leq \int_{\Omega} g(\bar{z}) d\mu = \sum_{\kappa \in J_k} y_{\kappa} \int_{\Omega} \bar{z}^{\kappa} d\mu = \sum_{\kappa \in J_k} y_{\kappa} \sigma_{\kappa} = Z_D. \quad \square$$

Theorem 2.1 indicates that by solving the dual problem (2.2) we obtain an upper bound on the primal objective, and hence on the probability we are trying to estimate. If the moment vector $\bar{\sigma}$ is an interior point of the set of feasible moment vectors, the dual bound turns out to be tight. This strong duality result follows from a univariate result due to Karlin and Isii in 1960 (see Karlin and Studden [19, p. 472]), and generalized by Isii in 1963 [17] for random vectors defined on complete regular spaces. Shapiro [39, Proposition 3.4] provides a modern proof based on conic linear duality, under a general topological structure. The following theorem is a consequence of their work, and holds for general distributions of X (discrete, continuous, with or without atoms, etc.) over a metric space Ω endowed with its Borel sigma algebra.

THEOREM 2.2 (strong duality). *If the moment vector $\bar{\sigma}$ is an interior point of the set $\mathcal{M}$ of feasible moment vectors, then $Z_P = Z_D$.*

If the dual is unbounded, it is immediate from weak duality that the multidimensional moment problem is infeasible. On the other hand, if the common optimal value is finite, then the set of dual optimal solutions is nonempty and bounded (see Isii [17], Kemperman [20], or Shapiro [39]). Furthermore, if $\bar{\sigma}$ is a boundary point of $\mathcal{M}$, then it can be shown that the $\bar{\sigma}$ -feasible distributions are concentrated on a subset Ω_0 of Ω , and strong duality holds provided we relax the dual to Ω_0 (see Isii [17, p. 190] or Smith [42, p. 824]). These authors also prove that it is equivalent to optimize only over discrete distributions that are concentrated on $m+2$ points, where m is the number of moment constraints.

In the univariate case, Isii [16] proves that if $\bar{\sigma}$ is a boundary point of $\mathcal{M}$, then exactly one $\bar{\sigma}$ -feasible distribution exists. Kemperman [20] proves that for almost every $\bar{\sigma}$ (with respect to the Lebesgue measure) such that the dual is finite, the dual solution is unique (see also Shapiro [39, Proposition 3.5]). In Chapter 5.4 of their recent book, Bonnans and Shapiro [3] provide necessary and sufficient conditions for the uniqueness of the optimal distribution, as well as sensitivity results.

If strong duality holds, then by optimizing over problem (2.2) we obtain a *tight* bound on $P(X \in S)$. On the other hand, it is worthwhile to remark that under

certain technical conditions (see Grötschel, Lovász, and Schrijver [11]) solving problem (2.2) is equivalent to solving the corresponding separation problem, which amounts to verifying polynomial positivity conditions over S and Ω .

3. The $(1, k, \Omega)$ -bound problem. In this section, we restrict our attention to univariate random variables. Given the first k moments $M_1, \dots, M_k$ (we let $M_0 = 1$) of a real random variable X with domain $\Omega \subseteq \mathcal{R}$, we are interested in deriving tight bounds on $P(X \in S)$. This is the $(1, k, \Omega)$ -bound problem. Our main result in this section is that tight bounds can be derived as a solution to a single semidefinite optimization problem. We also derive closed-form tight bounds when up to the first three moments are given.

3.1. Tight bounds as semidefinite optimization problems. In the one-dimensional case, problem (2.2) becomes

$$(3.1) \quad \begin{aligned} & \text{minimize} && \sum_{r=0}^k y_r M_r \\ & \text{subject to} && \sum_{r=0}^k y_r x^r \geq 1 \quad \forall x \in S, \\ & && \sum_{r=0}^k y_r x^r \geq 0 \quad \forall x \in \Omega. \end{aligned}$$

Problem (3.1) naturally leads us to investigate conditions for polynomials to be non-negative. When S and Ω are intervals on the real line, we show in the next proposition that the feasible region of problem (3.1) can be expressed using semidefinite constraints. Semidefinite optimization problems are efficiently solvable using interior point methods. For a review of semidefinite optimization see Wolkowicz, Saigal, and Vandenberghe [47]. The results and the proofs in the following proposition are inspired by Nesterov [31] (see also Ben-Tal and Nemirovski [2, pp. 140–142]).

PROPOSITION 3.1. (a) *The polynomial $g(x) = \sum_{r=0}^{2k} y_r x^r$ satisfies $g(x) \geq 0$ for all $x \in \mathcal{R}$ if and only if there exists a positive semidefinite matrix $X = [x_{ij}]_{i,j=0,\dots,k}$, such that*

$$(3.2) \quad y_r = \sum_{i,j:i+j=r} x_{ij}, \quad r = 0, \dots, 2k, \quad X \succeq 0.$$

(b) *The polynomial $g(x) = \sum_{r=0}^k y_r x^r$ satisfies $g(x) \geq 0$ for all $x \geq 0$ if and only if there exists a positive semidefinite matrix $X = [x_{ij}]_{i,j=0,\dots,k}$, such that*

$$(3.3) \quad \begin{aligned} 0 &= \sum_{i,j:i+j=2l-1} x_{ij}, \quad l = 1, \dots, k, \\ y_l &= \sum_{i,j:i+j=2l} x_{ij}, \quad l = 0, \dots, k, \\ X &\succeq 0. \end{aligned}$$

(c) The polynomial $g(x) = \sum_{r=0}^k y_r x^r$ satisfies $g(x) \geq 0$ for all $x \in [0, a]$ if and only if there exists a positive semidefinite matrix $X = [x_{ij}]_{i,j=0,\dots,k}$, such that

$$(3.4) \quad \begin{aligned} 0 &= \sum_{i,j:i+j=2l-1} x_{ij}, \quad l = 1, \dots, k, \\ \sum_{r=0}^l y_r \binom{k-r}{l-r} a^r &= \sum_{i,j:i+j=2l} x_{ij}, \quad l = 0, \dots, k, \\ X &\succeq 0. \end{aligned}$$

(d) The polynomial $g(x) = \sum_{r=0}^k y_r x^r$ satisfies $g(x) \geq 0$ for all $x \in [a, \infty)$ if and only if there exists a positive semidefinite matrix $X = [x_{ij}]_{i,j=0,\dots,k}$, such that

$$(3.5) \quad \begin{aligned} 0 &= \sum_{i,j:i+j=2l-1} x_{ij}, \quad l = 1, \dots, k, \\ \sum_{r=l}^k y_r \binom{r}{l} a^{r-l} &= \sum_{i,j:i+j=2l} x_{ij}, \quad l = 0, \dots, k, \\ X &\succeq 0. \end{aligned}$$

(e) The polynomial $g(x) = \sum_{r=0}^k y_r x^r$ satisfies $g(x) \geq 0$ for all $x \in (-\infty, a]$ if and only if there exists a positive semidefinite matrix $X = [x_{ij}]_{i,j=0,\dots,k}$, such that

$$(3.6) \quad \begin{aligned} 0 &= \sum_{i,j:i+j=2l-1} x_{ij}, \quad l = 1, \dots, k, \\ (-1)^l \sum_{r=l}^k y_r \binom{r}{l} a^{r-l} &= \sum_{i,j:i+j=2l} x_{ij}, \quad l = 0, \dots, k, \\ X &\succeq 0. \end{aligned}$$

(f) The polynomial $g(x) = \sum_{r=0}^k y_r x^r$ satisfies $g(x) \geq 0$ for all $x \in [a, b]$ if and only if there exists a positive semidefinite matrix $X = [x_{ij}]_{i,j=0,\dots,k}$, such that

$$(3.7) \quad \begin{aligned} 0 &= \sum_{i,j:i+j=2l-1} x_{ij}, \quad l = 1, \dots, k, \\ \sum_{m=0}^l \sum_{r=m}^{k+m-l} y_r \binom{r}{m} \binom{k-r}{l-m} a^{r-m} b^m &= \sum_{i,j:i+j=2l} x_{ij}, \quad l = 0, \dots, k, \\ X &\succeq 0. \end{aligned}$$

Proof.

(a) Suppose (3.2) holds. Let $x_{(k)} = (1, x, x^2, \dots, x^k)'$. Then

$$\begin{aligned} g(x) &= \sum_{r=0}^{2k} \sum_{i+j=r} x_{ij} x^r \\ &= \sum_{i=0}^k \sum_{j=0}^k x_{ij} x^i x^j \\ &= x'_{(k)} X x_{(k)} \\ &\geq 0, \end{aligned}$$

since $X \succeq 0$.

Conversely, suppose that the polynomial $g(x)$ of degree $2k$ is nonnegative for all x . Then, the real roots of $g(x)$ should have even multiplicity, otherwise $g(x)$ would alter its sign in a neighborhood of a root. Let λ_i , $i = 1, \dots, r$, be its real roots with corresponding multiplicity $2m_i$. Its complex roots can be arranged in conjugate pairs, $a_j + ib_j$, $a_j - ib_j$, $j = 1 \dots, h$. Then,

$$g(x) = y_{2k} \prod_{i=1}^r (x - \lambda_i)^{2m_i} \prod_{j=1}^h ((x - a_j)^2 + b_j^2).$$

The leading coefficient y_{2k} needs to be positive. Thus, by expanding the terms in the products, we see (cf. Proposition 4.2 below) that $g(x)$ can be written as a sum of squares of polynomials of the form

$$\begin{aligned} g(x) &= \sum_{i=1}^p \left(\sum_{j=0}^k q_{ij} x^j \right)^2 \\ &= x'_{(k)} Q' Q x_{(k)} = x'_{(k)} X x_{(k)}, \end{aligned}$$

with $X = Q'Q$ positive semidefinite, from which (3.2) follows.

(b) We observe that $g(x) \geq 0$ for $x \geq 0$ if and only if $g(t^2) \geq 0$ for all t . Since

$$g(t^2) = y_0 + 0 \cdot t + y_1 t^2 + 0 \cdot t^3 + y_2 t^4 + \dots + y_k t^{2k},$$

we obtain (3.3) by applying part (a).

(c) We observe that $g(x) \geq 0$ for $x \in [0, a]$ if and only if

$$(1 + t^2)^k g\left(\frac{at^2}{1 + t^2}\right) \geq 0 \quad \forall t.$$

Since

$$\begin{aligned} (1 + t^2)^k g\left(\frac{at^2}{1 + t^2}\right) &= \sum_{r=0}^k y_r a^r t^{2r} (1 + t^2)^{k-r} \\ &= \sum_{r=0}^k y_r a^r \sum_{l=0}^{k-r} \binom{k-r}{l} t^{2(l+r)} \\ &= \sum_{j=0}^k t^{2j} \left(\sum_{r=0}^j y_r \binom{k-r}{j-r} a^r \right), \end{aligned}$$

by applying part (a) we obtain (3.4).

(d) We observe that $g(x) \geq 0$ for $x \in [a, \infty)$ if and only if

$$g(a + t^2) \geq 0 \quad \forall t.$$

Since

$$\begin{aligned} g(a + t^2) &= \sum_{r=0}^k y_r (a + t^2)^r \\ &= \sum_{r=0}^k y_r \sum_{l=0}^r \binom{r}{l} a^{r-l} t^{2l} \\ &= \sum_{l=0}^k t^{2l} \left(\sum_{r=l}^k y_r \binom{r}{l} a^{r-l} \right), \end{aligned}$$

by applying part (a) we obtain (3.5).

(e) We observe that $g(x) \geq 0$ for $x \in (-\infty, a]$ if and only if

$$g(a - t^2) \geq 0 \quad \forall t.$$

Since

$$\begin{aligned} g(a - t^2) &= \sum_{r=0}^k y_r (a - t^2)^r \\ &= \sum_{r=0}^k y_r \sum_{l=0}^r \binom{r}{l} a^{r-l} (-t^2)^l \\ &= \sum_{l=0}^k t^{2l} \left((-1)^l \sum_{r=l}^k y_r \binom{r}{l} a^{r-l} \right), \end{aligned}$$

by applying part (a) we obtain (3.6).

(f) We observe that $g(x) \geq 0$ for $x \in [a, b]$ if and only if

$$(1 + t^2)^k g \left(a + (b - a) \frac{t^2}{1 + t^2} \right) \geq 0 \quad \forall t.$$

Since

$$\begin{aligned} (1 + t^2)^k g \left(a + (b - a) \frac{t^2}{1 + t^2} \right) &= \sum_{r=0}^k y_r (a + bt^2)^r (1 + t^2)^{k-r} \\ &= \sum_{r=0}^k y_r \sum_{m=0}^r \binom{r}{m} a^{r-m} b^m t^{2m} \sum_{j=0}^{k-r} \binom{k-r}{j} t^{2j} \\ &= \sum_{l=0}^k t^{2l} \left(\sum_{m=0}^l \sum_{r=m}^{k+m-l} y_r \binom{r}{m} \binom{k-r}{l-m} a^{r-m} b^m \right), \end{aligned}$$

by applying part (a) we obtain (3.7). $\square$

We next show that problem (3.1) can be written as a semidefinite optimization problem. Corresponding formulations, which we omit here for the sake of conciseness, can be obtained similarly for the problem of establishing lower bounds, i.e., problem (2.1) in which we replace maximization by minimization.

THEOREM 3.2. *Given the first k moments $(M_1, \dots, M_k)$ (we let $M_0 = 1$) of a random variable X defined on Ω we obtain the following tight upper bounds.*

(a) *If $\Omega = \mathcal{R}^+$, the tight upper bound on $P(X \geq a)$ is given as the solution of the*

semidefinite optimization problem

$$\begin{aligned}
 (3.8) \quad & \text{minimize} \quad \sum_{r=0}^k y_r M_r \\
 & \text{subject to} \quad 0 = \sum_{i,j:i+j=2l-1} x_{ij}, \quad l = 1, \dots, k, \\
 & \quad (y_0 - 1) + \sum_{r=1}^k y_r a^r = x_{00}, \\
 & \quad \sum_{r=l}^k y_r \binom{r}{l} a^{r-l} = \sum_{i,j:i+j=2l} x_{ij}, \quad l = 1, \dots, k, \\
 & \quad 0 = \sum_{i,j:i+j=2l-1} z_{ij}, \quad l = 1, \dots, k, \\
 & \quad \sum_{r=0}^l y_r \binom{k-r}{l-r} a^r = \sum_{i,j:i+j=2l} z_{ij}, \quad l = 0, \dots, k, \\
 & \quad X, Z \succeq 0.
 \end{aligned}$$

If $\Omega = \mathcal{R}$, then the last equation in (3.8) should be replaced by

$$(-1)^l \sum_{r=l}^k y_r \binom{r}{l} a^{r-l} = \sum_{i,j:i+j=2l} z_{ij}, \quad l = 0, \dots, k.$$

(b) If $\Omega = \mathcal{R}^+$, the tight upper bound on $P(a \leq X \leq b)$ is given as the solution of the semidefinite optimization problem

$$\begin{aligned}
 (3.9) \quad & \text{minimize} \quad \sum_{r=0}^k y_r M_r \\
 & \text{subject to} \quad 0 = \sum_{i,j:i+j=2l-1} x_{ij}, \quad l = 1, \dots, k, \\
 & \quad \sum_{m=0}^l \sum_{r=m}^{k+m-l} y_r \binom{r}{m} \binom{k-r}{l-m} a^{r-m} b^m = \binom{k}{l} + \sum_{i,j:i+j=2l} x_{ij}, \quad l = 0, \dots, k, \\
 & \quad 0 = \sum_{i,j:i+j=2l-1} z_{ij}, \quad l = 1, \dots, k, \\
 & \quad y_l = \sum_{i,j:i+j=2l} z_{ij}, \quad l = 0, \dots, k, \\
 & \quad X, Z \succeq 0.
 \end{aligned}$$

If $\Omega = \mathcal{R}$, then the last equation in (3.9) should be replaced by

$$(-1)^l \sum_{r=l}^k y_r \binom{r}{l} a^{r-l} = \sum_{i,j:i+j=2l} z_{ij}, \quad l = 0, \dots, k,$$

and the following equations added:

$$\begin{aligned} 0 &= \sum_{i,j:i+j=2l-1} u_{ij}, \quad l = 1, \dots, k, \\ \sum_{r=l}^k y_r \binom{r}{l} b^{r-l} &= \sum_{i,j:i+j=2l} u_{ij}, \quad l = 0, \dots, k, \\ U &\succeq 0. \end{aligned}$$

Proof.

(a) The feasible region of problem (3.1) for $S = [a, \infty)$ and $\Omega = \mathcal{R}_+$ becomes

$$g(x) = \sum_{r=0}^k y_r x^r \geq 1 \quad \forall x \in [a, \infty) \quad \text{and} \quad g(x) \geq 0 \quad \forall x \in [0, a).$$

By applying Proposition 3.1(c), (d) we obtain (3.8). If $\Omega = \mathcal{R}$, we apply Proposition 3.1(d), (e).

(b) The feasible region of problem (3.1) for $S = [a, b]$ and $\Omega = \mathcal{R}_+$ becomes

$$g(x) = \sum_{r=0}^k y_r x^r \geq 1 \quad \forall x \in [a, b] \quad \text{and} \quad g(x) \geq 0 \quad \forall x \in [0, \infty).$$

By applying Proposition 3.1(b), (f) we obtain (3.9). If $\Omega = \mathcal{R}$, we apply Proposition 3.1(d), (e), (f). $\square$

3.2. Closed-form bounds. In this section, we present closed-form bounds when moments up to third order are given. We define the squared coefficient of variation $C_M^2 = \frac{M_2 - M_1^2}{M_1^2}$, and the third-order coefficient of variation $D_M^2 = \frac{M_1 M_3 - M_2^2}{M_1^4}$. Let $\delta > 0$.

THEOREM 3.3. *The following bounds in Table 2 are tight for $k = 1, 2, 3$.*

The bounds marked with an asterisk (*) assume that $\delta < 1$ (the other case is trivial). The following definitions are used:

$$\begin{aligned} f_1(C_M^2, D_M^2, \delta) &= \begin{cases} \min \left(\frac{C_M^2}{C_M^2 + \delta^2}, \frac{1}{1 + \delta} \cdot \frac{D_M^2}{D_M^2 + (C_M^2 - \delta)^2} \right), & \text{if } \delta > C_M^2, \\ \frac{1}{1 + \delta} \cdot \frac{D_M^2 + (1 + \delta)(C_M^2 - \delta)}{D_M^2 + (1 + C_M^2)(C_M^2 - \delta)}, & \text{if } \delta \leq C_M^2, \end{cases} \\ f_2(C_M^2, D_M^2, \delta) &= 1 - \frac{(C_M^2 + \delta)^3}{(D_M^2 + (C_M^2 + 1)(C_M^2 + \delta))(D_M^2 + (C_M^2 + \delta)^2)}, \\ f_3(C_M^2, D_M^2, \delta) &= \min \left(1, 1 + 3^3 \frac{D_M^2 + C_M^4 - \delta^2}{4 + 3(1 + 3\delta^2) + 2(1 + 3\delta^2)^{\frac{3}{2}}} \right). \end{aligned}$$

The proof of the theorem is given in Popescu [33].

4. The (n, k, Ω) -bound problem: Semidefinite formulations and complexity. In this section, we generalize the results of the previous section to the multivariate case by investigating the (n, k, Ω) -bound problem. For a fairly general class

TABLE 2
Tight bounds for the $(1, k, \Omega)$ -problem for $k \leq 3$.

(k, Ω)	$P(X > (1 + \delta)M_1)$	$P(X < (1 - \delta)M_1)$	$P(X - M_1 > \delta M_1)$
$(1, \mathcal{R}_+)$	$\frac{1}{1+\delta}$	1^*	1^*
$(2, \mathcal{R})$	$\frac{C_M^2}{C_M^2 + \delta^2}$	$\frac{C_M^2}{C_M^2 + \delta^2}$	$\min\left(1, \frac{C_M^2}{\delta^2}\right)$
$(3, \mathcal{R}_+)$	$f_1(C_M^2, D_M^2, \delta)$	$f_2(C_M^2, D_M^2, \delta)^*$	$f_3(C_M^2, D_M^2, \delta)^*$

of semialgebraic sets S and Ω , i.e., sets that are given as intersections of inequalities involving polynomials, we propose a sequence of increasingly stronger, asymptotically exact upper bounds by solving semidefinite optimization problems. This includes the case when Ω is a bounded polyhedron, or a semialgebraic set such that $\Omega \subseteq \{x \in \mathcal{R}^n \mid x'x \leq M^2\}$ and M is known a priori. Note that it is not necessary for the sets S and Ω to be convex. We expect that in general an exact semidefinite formulation, if it exists, may be exponential in n even for fixed k . This fact should not be surprising, given that in section 4.2 we prove that it is NP-hard to find best possible bounds for the $(n, k, \mathcal{R}^n)$ -bound problem for $k \geq 4$. We also show that it is NP-hard to find best possible bounds for the $(n, k, \mathcal{R}_+^n)$ -bound problem for $k \geq 2$.

4.1. Semidefinite programming formulations. The approach we follow in this section has its origin in the work of Shor [41]. Recently it has been used by Lasserre [24] and Parrilo [32] to provide semidefinite relaxations for discrete optimization and nonconvex optimization problems.

In dimension $n = 1$, we have seen in the proof of Proposition 3.1 that a polynomial in one dimension is nonnegative if and only if it can be written as a sum of squares of polynomials. Clearly, in multiple dimensions if a polynomial can be written as a sum of squares of other polynomials, then it is nonnegative. Hilbert, in a nonconstructive way, has shown that it is possible for a polynomial in higher dimensions to be nonnegative without being a sum of squares. Motzkin (see Reznick [35]) has shown that the polynomial

$$M(x, y, z) = x^4y^2 + x^2y^4 + z^6 - 3x^2y^2z^2$$

is nonnegative without being a sum of squares of polynomials. The connection between nonnegative polynomials and sum of squares representations has a long history that is nicely outlined in Reznick [35]. In this paper we will rely on the following theorem due to Putinar [34].

THEOREM 4.1 (Putinar [34]). *Suppose that the set*

$$K = \{x \in \mathcal{R}^n \mid g_i(x) \geq 0, i \in I\}$$

is compact and there exists a polynomial $h(x)$ of the form

$$h(x) = h_0(x) + \sum_{i \in I} h_i(x)g_i(x),$$

such that $\{x \in \mathcal{R}^n \mid h(x) \geq 0\}$ is compact and $h_i(x)$, $i \in I \cup \{0\}$ are polynomials that have a sum of squares representation. Then, for any polynomial $g(x)$ that is strictly

positive for all $x \in K$, there exist $p_i(x)$, $i \in I \cup \{0\}$, that are sums of squares such that

$$(4.1) \quad g(x) = p_0(x) + \sum_{i \in I} p_i(x) g_i(x).$$

Putinar [34] shows that the conditions for Theorem 4.1 to hold are relatively mild. In particular, examples of sets K that satisfy the conditions of Theorem 4.1 include the following.

(a) One of the defining inequalities of K represents a compact set $\{x \in \mathcal{R}^n \mid g_l(x) \geq 0\}$. This is the case, for example, when $g_l(x) = c - (x - x_0)'Q(x - x_0)$, and $Q \succ 0$.

(b) K is a bounded polyhedron.

(c) K is compact, and there is a bound M a priori known such that $K \subseteq \{x \in \mathcal{R}^n \mid x'x \leq M^2\}$.

While Theorem 4.1 guarantees the existence of the representation in (4.1), it does not give any information on the degree of the polynomials $p_i(x)$. From Theorem 4.1 it is natural to investigate when a polynomial is a sum of squares. Let $x_{(d)}$ be the vector of all monomials in the variables $x_1, \dots, x_n$ of degree d and below. For example, for $n = 2$, $d = 2$,

$$x_{(2)} = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2)'$$

There are $a = \binom{n+d}{d}$ such monomials.

PROPOSITION 4.2. *The polynomial $f(x)$ of degree $2d$ has a sum of squares decomposition if and only if there exists a positive semidefinite matrix Q for which $f(x) = x'_{(d)} Q x_{(d)}$.*

Proof. To simplify notation, let $x = x_{(d)}$ throughout the proof. An arbitrary polynomial of degree $2d$ can be written as $f(x) = x'Qx$ for some $a \times a$ matrix Q (the matrix Q is not unique, however). If $Q \succeq 0$, then $Q = HH'$ for some H , and thus,

$$f(x) = x'HH'x = \sum_{i=1}^a (H'x)_i^2.$$

Since $(H'x)_i$ is a polynomial, we have expressed $f(x)$ as a sum of squares of the polynomials $(H'x)_i$.

Conversely, suppose that $f(x)$ has a sum of squares decomposition

$$f(x) = \sum_{i=1}^l h_i(x)^2 = \sum_{i=1}^l (h'_i x)^2 = \sum_{i=1}^l x' (h_i h'_i) x = x' \left(\sum_{i=1}^l (h_i h'_i) \right) x = x' Q x,$$

where h_i is the vector of coefficients of the polynomial $h_i(x)$. Since $Q = \sum_{i=1}^l (h_i h'_i) \succeq 0$, the proposition follows. $\square$

Therefore, if the sets Ω and S satisfy the conditions of Theorem 4.1, by applying Proposition 4.2, we can express problem (2.2) as a semidefinite optimization problem. Suppose that the sets Ω and S are semialgebraic sets, given by

$$\Omega = \left\{ x \in \mathcal{R}^n \mid \omega_i(x) = \sum_{\kappa \in J_i} \omega_{\kappa}^i x^{\kappa} \geq 0, \ i = 1, \dots, r \right\},$$

$$S = \left\{ x \in \mathcal{R}^n \mid s_i(x) = \sum_{\kappa \in J_i} s_{\kappa}^i x^{\kappa} \geq 0, \ i = 1, \dots, m \right\},$$

where $\omega_\kappa^i, s_\kappa^i \in \mathcal{R}$, that is Ω and S are defined by polynomial inequalities. We use the notation $\delta_{\kappa,0} = 1$ if $\kappa = 0$, and zero, otherwise.

THEOREM 4.3. *If the sets Ω and S satisfy the conditions of Theorem 4.1, then for every $\epsilon > 0$ there exists a nonnegative integer $d \in \mathbb{Z}_+$ (representing the degree of the polynomials in (4.1)), such that the objective function value Z_D in problem (2.2) satisfies $|Z_D - Z_D(d)| \leq \epsilon$, where $Z_D(d)$ is the value of the following semidefinite program:*

$$\begin{aligned}
 (4.2) \quad Z_D(d) = \min \quad & \sum_{\kappa \in J_k} y_\kappa \sigma_\kappa \\
 \text{subject to} \quad & y_\kappa - \delta_{\kappa,0} = q_\kappa^0 + \sum_{i=1}^m \sum_{\substack{\eta \in J_d, \theta \in J_t \\ \eta + \theta = \kappa}} q_\eta^i s_\theta^i \quad \forall \kappa \in J_k, \\
 & 0 = q_\kappa^0 + \sum_{i=1}^m \sum_{\substack{\eta \in J_d, \theta \in J_t \\ \eta + \theta = \kappa}} q_\eta^i s_\theta^i \quad \forall \kappa \in J_{s+d} \setminus J_k, \\
 & y_\kappa = p_\kappa^0 + \sum_{i=1}^r \sum_{\substack{\eta \in J_d, \theta \in J_l \\ \eta + \theta = \kappa}} p_\eta^i \omega_\theta^i \quad \forall \kappa \in J_k, \\
 & 0 = p_\kappa^0 + \sum_{i=1}^r \sum_{\substack{\eta \in J_d, \theta \in J_l \\ \eta + \theta = \kappa}} p_\eta^i \omega_\theta^i \quad \forall \kappa \in J_{l+d} \setminus J_k, \\
 & q_\kappa^i = \sum_{\eta, \theta \in J_d, \eta + \theta = \kappa} q_\eta^i \theta \quad \forall \kappa \in J_d, \quad i = 0, 1, \dots, m, \\
 & p_\kappa^i = \sum_{\eta, \theta \in J_d, \eta + \theta = \kappa} p_\eta^i \theta \quad \forall \kappa \in J_d, \quad i = 0, 1, \dots, m, \\
 & Q^i = [q_{\eta, \theta}^i]_{\eta, \theta \in J_d} \succeq 0, \quad i = 0, 1, \dots, r, \\
 & P^i = [p_{\eta, \theta}^i]_{\eta, \theta \in J_d} \succeq 0, \quad i = 0, 1, \dots, m.
 \end{aligned}$$

Proof. We first remark that the value of the dual problem (2.2) equals that of the following strict inequality formulation:

$$\begin{aligned}
 (4.3) \quad Z_D = \inf_{y \in \mathcal{R}^{|J_k|}} \quad & \sum_{\kappa \in J_k} y_\kappa \sigma_\kappa \\
 \text{subject to} \quad & g(x) = \sum_{\kappa \in J_k} y_\kappa x^\kappa > 1 \quad \forall x \in S, \\
 & g(x) = \sum_{\kappa \in J_k} y_\kappa x^\kappa > 0 \quad \forall x \in \Omega.
 \end{aligned}$$

This problem may not admit an optimal solution. However, for any $\epsilon > 0$ there exists a feasible polynomial g_ϵ resulting in an objective value that is less than $Z_D + \epsilon$. Fix $\epsilon > 0$ and let $g = g_\epsilon$. For this particular g , from Theorem 4.1 and Proposition 4.2, the above feasibility constraint $g(x) > 0$ for all $x \in \Omega$ is equivalent to

$$g(x) = p_0(x) + \sum_{i=1}^m p_i(x) \omega_i(x),$$

where $p_i(x) = x'P^i x$, $P^i \succeq 0$, and x includes monomials up to a certain degree d_0 . Writing

$$p_i(x) = \sum_{\eta \in J_d} p_{\eta}^i x^{\eta}, \quad \omega_i(x) = \sum_{\theta \in J_l} \omega_{\theta}^i x^{\theta},$$

we have that $g(x) > 0$ for $x \in \Omega$ if and only if

$$g(x) := \sum_{\kappa \in I} y_{\kappa} x^{\kappa} = \sum_{\eta \in J_d} p_{\eta}^0 x^{\eta} + \sum_{i=1}^r \left(\sum_{\eta \in J_d} p_{\eta}^i x^{\eta} \right) \left(\sum_{\theta \in J_l} \omega_{\theta}^i x^{\theta} \right).$$

Equating terms, we obtain the third and fourth sets of linear constraints in problem (4.2), corresponding to the degree d_0 .

Similarly, one can translate the feasibility constraint $g(x) - 1 > 0$ for all $x \in S$ into the first two sets of constraints in problem (4.2), corresponding to a certain degree d_1 . It follows that the vector of coefficients y of the polynomial g is feasible for problem (4.2) with degree $d = \max(d_0, d_1)$, so $Z_D(d) \leq Z_D + \epsilon$.

On the other hand, for any d , problem (4.2) is a restriction of problem (4.3) for feasible polynomials of degree d , so $Z_D(d) \geq Z_D$, and the desired result follows. $\square$

Theorem 4.3 gives an asymptotically exact sequence of semidefinite formulations (4.2) of problem (2.2). Unfortunately, the sizes of these formulations are not bounded by a polynomial in n , even for fixed k , as they depend on the degree d of the polynomials appearing in (4.1). This is not surprising, given our NP-hardness results in the next section. Nevertheless, we can obtain increasingly better upper bounds on Z_P by solving semidefinite problems of size polynomially bounded in n for fixed k . Indeed, since any g that can be represented using sums of squares of polynomials of degree d can clearly also be represented using sums of squares of polynomials of degree $d + 1$, we obtain a family of semidefinite relaxations as follows.

COROLLARY 4.4. $Z_P = Z_D \leq Z_D(d) \leq \dots \leq Z_D(2) \leq Z_D(1)$.

4.2. On the complexity of the (n, k, Ω) -bound problem. In this section, we show that the separation problem associated with problem (2.2) for the cases $(n, k, \mathcal{R}_+^n)$, $(n, 2k, \mathcal{R}^n)$ are NP-hard for $k \geq 2$. By the equivalence of optimization and separation (see Grötschel, Lovász, and Schrijver [11]), solving problem (2.2) is NP-hard as well in these cases. Finally, since strong duality (Theorem 2.2) holds in these instances, solving the $(n, k, \mathcal{R}_+^n)$ -bound problems for $k \geq 2$ and solving the $(n, k, \mathcal{R}^n)$ -bound problems with $k \geq 4$ is NP-hard.

The $(n, 2k + 1, \mathcal{R}^n)$ -bound problem does not make a case on its own, since an odd degree polynomial cannot be nonnegative over all of $\mathcal{R}^n$. This means that in the corresponding dual problem, the variables y corresponding to $(2k + 1)$ -degree coefficients must be zero for the problem to be feasible. Thus the $(n, 2k + 1, \mathcal{R}^n)$ -bound problem reduces to the $(n, 2k, \mathcal{R}^n)$ -bound problem. In other words, if the highest-order moments that are known are of odd order, they can be disregarded as they will not improve the bound.

4.2.1. The complexity of the $(n, k, \mathcal{R}_+^n)$ -bound problem. The separation problem is equivalent to the following problem.

Problem k -SEP⁺: Given a multivariate polynomial $g(x)$ with rational coefficients and degree k , and a (nonempty) set $S \subseteq \mathcal{R}_+^n$, does there exist $x \in \mathcal{R}_+^n$ such that $g(x) < \chi_S(x)$?

PROPOSITION 4.5. *Problem $k\text{-SEP}^+$ is NP-hard for $k \geq 2$, even for polyhedral sets S .*

Proof. Consider the following problem of checking if a given matrix is not copositive.

Problem COPOS: Given a matrix H with rational entries, does there exist $x \in \mathcal{R}_+^n$ such that $x'Hx < 0$?

This problem is NP-hard (see Murty and Kabadi [29]). We will prove that the problem COPOS reduces to Problem $k\text{-SEP}^+$. Since COPOS is NP-hard, so is $k\text{-SEP}^+$.

Consider first the case $k = 2$, $g(x) = x'Hx + 1$, where H has rational entries, and $S = \mathcal{R}_+^n$. Problem 2-SEP^+ is equivalent to the question of whether the matrix H is not copositive. Now, for arbitrary $k \geq 2$, consider the k -degree polynomial $g(x) = x_n^k + (x_1, \dots, x_{n-1})H(x_1, \dots, x_{n-1})' + 1$, where H has rational entries, and $S = \mathcal{R}_+^n$. The separation problem $k\text{-SEP}^+$ amounts in this case to checking $g(x) < 1$ for some nonnegative x , which is equivalent to checking if matrix H is not copositive, therefore it is NP-hard. $\square$

4.2.2. The complexity of the $(n, 2k, \mathcal{R}^n)$ -bound problem for $k \geq 2$. For $k \geq 2$, the separation problem can be formulated as follows.

Problem $2k\text{-SEP}$: Given a multivariate polynomial $g(\cdot)$ with rational coefficients and degree $2k \geq 4$, and a (nonempty) set $S \subseteq \mathcal{R}^n$, does there exist $x \in \mathcal{R}^n$ such that $g(x) < \chi_S(x)$?

PROPOSITION 4.6. *Problem $2k\text{-SEP}$ is NP-hard for $k \geq 2$ even for polyhedral sets S .*

Proof. As in the previous proof, we show that the problem COPOS polynomial time reduces to Problem $2k\text{-SEP}$ for $k \geq 2$. Since COPOS is NP-hard (see Murty and Kabadi [29]), it follows that $2k\text{-SEP}$ is also NP-hard.

Consider first the case $k = 2$, $g(x) = (x_1^2, \dots, x_n^2)'H(x_1^2, \dots, x_n^2) + 1$, where H has rational entries, and $S = \mathcal{R}^n$. Problem 4-SEP is equivalent to the question of whether the matrix H is not copositive. Now, for arbitrary $k \geq 2$, consider the $2k$ -degree polynomial $g(x) = x_n^{2k} + (x_1^2, \dots, x_{n-1}^2)H(x_1^2, \dots, x_{n-1}^2)' + 1$, where H has rational entries, and $S = \mathcal{R}^n$. The separation problem $2k\text{-SEP}$ amounts in this case to checking $g(x) < 1$ for some $x \in \mathcal{R}^n$, which is equivalent to checking if matrix H is not copositive; therefore it is NP-hard. $\square$

5. The $(n, 1, \Omega)$ -bound problem. In this section, we address the $(n, 1, \Omega)$ -bound problem. For convex sets S , the tight bound for the $(n, 1, \mathcal{R}_+^n)$ -bound problem is computed in (5.2) as the solution of n convex optimization problems. We present a polynomial time algorithm for more general sets.

Given a vector $M > 0$ representing the means of a random vector X defined in $\mathcal{R}_+^n$, we would like to find tight bounds on $P(X \in S)$ for a convex set S . Marshall [26] derived a tight bound for the case that $S = \{x_i \geq (1 + \delta_i)M_i, i = 1, \dots, n\}$ (see (5.3) below).

THEOREM 5.1. *The tight $(n, 1, \mathcal{R}_+^n)$ -upper bound for a convex event S is given by*

$$(5.1) \quad \sup_{X \sim M} P(X \in S) = \min \left(1, \frac{1}{\inf_{x \in S} \max_{i=1, \dots, n} \frac{x_i}{M_i}} \right).$$

Proof. Since only first moments are given, the condition that the vector of moments is in the interior is trivially satisfied, and thus Theorem 2.2 applies. The

corresponding dual problem (2.2) is

$$\begin{aligned} Z_D &= \min \quad a'M + b \\ \text{subject to} \quad & a'x + b \geq 1 \quad \forall x \in S, \\ & a'x + b \geq 0 \quad \forall x \in \mathcal{R}_+^n. \end{aligned}$$

If the optimal solution (a_0, b_0) satisfies $\min_{x \in S} a'_0 x + b_0 = \kappa > 1$, then the solution $(\frac{a_0}{\kappa}, \frac{b_0}{\kappa})$ has value $Z_D/\kappa < Z_D$. Therefore, $\inf_{x \in S} a'_0 x + b_0 = 1$. By a similar argument we have that $b_0 \leq 1$. Moreover, since $a'x + b \geq 0$ for all $x \in \mathcal{R}_+^n$, $a \geq 0$, and $b \geq 0$. We thus obtain

$$\begin{aligned} Z_D &= \min \quad a'M + b \\ \text{subject to} \quad & \inf_{x \in S} a'x = 1 - b, \\ & a \geq 0, \quad 0 \leq b \leq 1. \end{aligned}$$

Let $a = \lambda v$, where $\lambda \geq 0$ is a scalar, and $v \geq 0$ is a vector with $\|v\| = 1$. We obtain

$$\begin{aligned} Z_D &= \min \quad (1 - b) \frac{v'M}{\inf_{x \in S} v'x} + b \\ \text{subject to} \quad & v \geq 0, \quad \|v\| = 1, \quad 0 \leq b \leq 1. \end{aligned}$$

Thus,

$$\begin{aligned} Z_D &= \min \left(1, \min_{\|v\|=1, v \geq 0} \frac{v'M}{\inf_{x \in S} v'x} \right) \\ &= \min \left(1, \frac{1}{\max_{\|v\|=1, v \geq 0} \inf_{x \in S} \frac{v'x}{v'M}} \right) \\ &= \min \left(1, \frac{1}{\inf_{x \in S} \max_{\|v\|=1, v \geq 0} \frac{v'x}{v'M}} \right) \\ &= \min \left(1, \frac{1}{\inf_{x \in S} \max_{i=1, \dots, n} \frac{x_i}{M_i}} \right). \end{aligned}$$

We exchanged the order of max and inf (see [36, p. 382]); then we relied on the fact that $\max_{\|v\|=1, v \geq 0} \frac{v'x}{v'M}$ is attained at $v = e_j$, where $\frac{x_j}{M_j} = \max_{i=1, \dots, n} \frac{x_i}{M_i}$. $\square$

Denote $\phi(x) = \max_{i=1, \dots, n} \frac{x_i}{M_i}$, so $\phi(x) = \frac{x_i}{M_i}$ whenever $x \in S_i$. The last term in (5.1) can be written as

$$\frac{1}{\inf_{x \in S} \phi(x)} = \max_{i=1, \dots, n} \sup_{x \in S_i} \frac{1}{\phi(x)} = \max_{i=1, \dots, n} \sup_{x \in S_i} \frac{M_i}{x_i} = \max_{i=1, \dots, n} \frac{M_i}{\inf_{x \in S_i} x_i}.$$

This shows that the bound in Theorem 5.1 can be obtained by solving n convex optimization problems:

$$(5.2) \quad \sup_{X \sim M} P(X \in S) = \min \left(1, \max_{i=1, \dots, n} \frac{M_i}{\inf_{x \in S_i} x_i} \right),$$

where $S_i = \{x \in S \mid \frac{x_i}{M_i} \geq \frac{x_j}{M_j} \forall j \neq i\}$ is the convex subset of S for which the mean-rescaled i th coordinate is largest. Remark that $\inf_{x \in S_i} x_i = 0$ for some i if and only if 0 is a boundary point of S , in which case the bound is trivially 1.

When we specialize bound (5.2) for the set $S = \{x \mid x_i \geq (1 + \delta_i)M_i, \forall i = 1, \dots, n\}$, we obtain

$$(5.3) \quad \sup_{X \sim M} P(X_1 \geq (1 + \delta_1)M_1, \dots, X_n \geq (1 + \delta_n)M_n) = \min_{i=1, \dots, n} \frac{1}{1 + \delta_i},$$

which represents a multidimensional generalization of Markov's inequality due to Marshall [26]. In particular, for a univariate random variable, in the case that $S = [(1 + \delta)M, \infty)$, this is exactly Markov's inequality:

$$\sup_{X \sim M} P(X \geq (1 + \delta)M) = \frac{1}{1 + \delta}.$$

5.1. Extremal distributions for the $(n, 1, \mathcal{R}_+^n)$ -bound problem. In this section, we construct a distribution that achieves bound (5.1). We will say that the bound (5.1) is achievable, when there exists an $x^* \in S$ such that

$$\min \left(1, \frac{1}{\inf_{x \in S} \max_{i=1, \dots, n} \frac{x_i}{M_i}} \right) = \frac{M_i}{x_i^*} < 1.$$

In particular, the bound is achievable when the set S is closed and $M \notin S$.

THEOREM 5.2. (a) *If $M \in S$ or if the bound (5.1) is achievable, then there is an extremal distribution that exactly achieves it.*

(b) *Otherwise, there is a sequence of distributions defined on $\mathcal{R}_+^n$ with mean M , that asymptotically achieves it.*

Proof. (a) If $M \in S$, then the extremal distribution is simply $P(X = M) = 1$. Now suppose that $M \notin S$ and the bound (5.1) is achievable. We assume without loss of generality that the bound equals $\frac{M_1}{x_1^*} < 1$, and it is achieved at $x^* \in S$ with $\frac{x_1^*}{M_1} = \max_{i=1, \dots, n} \frac{x_i^*}{M_i}$, i.e., $x^* \in S_1$. The random variable

$$X = \begin{cases} x^*, & \text{with probability } p = \frac{M_1}{x_1^*}, \\ v = \frac{x_1^*M - M_1x^*}{x_1^* - M_1}, & \text{with probability } 1 - p = 1 - \frac{M_1}{x_1^*}, \end{cases}$$

has mean $E[X] = M$ and is nonnegative: $v_i = \frac{M_i x_1^* - M_1 x_i^*}{x_1^* - M_1} \geq 0$ for all i since $\frac{x_1^*}{M_1} = \max_{i=1, \dots, n} \frac{x_i^*}{M_i}$. Moreover, $v \notin S$, or else convexity of S implies $M = px^* + (1-p)v \in S$, a contradiction. Therefore,

$$P(X \in S) = P(X = x^*) = \frac{M_1}{x_1^*}.$$

(b) If $M \notin S$ and the bound (5.1) is not achievable, then we construct a sequence of nonnegative distributions with mean M that approach it. Suppose without loss of generality that $\inf_{x \in S} \max_{i=1, \dots, n} \frac{x_i}{M_i} = \inf_{x \in S} \frac{x_1}{M_1} = \frac{x_1^*}{M_1}$. Thus, bound (5.1) equals $\min(1, \frac{M_1}{x_1^*})$. Consider a sequence $x^k \in S_1$ with $x_1^k \rightarrow x_1^*$, so $\lim_{k \rightarrow \infty} \max_{i=1, \dots, n} \frac{x_i^k}{M_i} = \lim_{k \rightarrow \infty} \frac{x_1^k}{M_1} = \frac{x_1^*}{M_1}$, and a sequence p_k , $0 < p_k < \min(1, \frac{M_1}{x_1^k})$ so that $p_k \rightarrow \min(1, \frac{M_1}{x_1^*})$. Therefore the random variables

$$X_k = \begin{cases} x^k, & \text{with probability } p_k, \\ v^k = \frac{M - p_k x^k}{1 - p_k}, & \text{with probability } 1 - p_k, \end{cases}$$

are nonnegative with mean $E[X_k] = M$. Also $v^k \notin S$ or else $M \in S$, so $P(X_k \in S) = P(X_k = x^k) = p_k \rightarrow \min(1, \frac{M_1}{x_1^*})$. This shows that the sequence of nonnegative distributions X_k with mean M asymptotically achieves the bound (5.1). $\square$

5.2. A polynomial time algorithm for unions of convex sets. In this section, we present a polynomial time algorithm that computes a tight $(n, 1, \Omega)$ -bound for events S and Ω that can be decomposed as a disjoint union of a polynomial (in n) number of convex sets such that we can solve linear optimization problems over them in polynomial time. Examples include polyhedra and sets defined by semidefinite constraints. Our overall strategy is to formulate the problem as an optimization problem, consider its dual, and exhibit an algorithm that solves the corresponding separation problem in polynomial time.

In this case we are given the mean vector $M = (M_1, \dots, M_n)$ of an n -dimensional random variable X with domain Ω that can be decomposed in a polynomial (in n) number of convex sets, and we want to derive tight bounds on $P(X \in S)$. Problem (2.2) can be written as follows:

$$(5.4) \quad \begin{aligned} Z_D = \min \quad & y'M + y_0 \\ \text{subject to} \quad & g(x) = y'x + y_0 \geq \chi_S(x) \quad \forall x \in \Omega. \end{aligned}$$

The separation problem associated with problem (5.4) is defined as follows: given a vector a and a scalar b we want to check whether $g(x) = a'x + b \geq \chi_S(x)$, for all $x \in \Omega$, and if not, we want to exhibit a violated inequality. The following algorithm achieves this goal.

ALGORITHM A.

1. Solve the problem $\inf_{x \in \Omega} g(x)$ (note that the problem involves a polynomial number of convex optimization problems; in particular if Ω is polyhedral, this is a linear optimization problem). Let z_0 be the optimal solution value and let $x_0 \in \Omega$ be an optimal solution.
2. If $z_0 < 0$, then we have $g(x_0) = z_0 < 0$; this constitutes a violated inequality.
3. Otherwise, we solve $\inf_{x \in S} g(x)$ (again, the problem involves a polynomial number of convex optimization problems, while if S is polyhedral, this is a linear optimization problem). Let z_1 be the optimal solution value and let $x_1 \in S$ be an optimal solution.
 - (a) If $z_1 < 1$, then for $x_1 \in S$ we have $g(x_1) = z_1 < 1$; this constitutes a violated inequality.
 - (b) If $z_1 \geq 1$, then a, b are feasible.

The above algorithm solves the separation problem in polynomial time, and thus the $(n, 1, \Omega)$ -upper bound problem is polynomially solvable (see Nesterov and Nemirovskii [30] and Grötschel, Lovász, and Schrijver [11]).

6. The $(n, 2, \mathcal{R}^n)$ -bound problem. In this section, we address the $(n, 2, \mathcal{R}^n)$ -bound problem. Rather than assuming that $E[X]$ and $E[XX']$ are known, we assume equivalently that the vector $M = E[X]$ and the covariance matrix $\Gamma = E[(X - M)(X - M)']$ are known and that Γ is invertible. Given a set $S \subseteq \mathcal{R}^n$, we find tight upper bounds, denoted by $\sup_{X \sim (M, \Gamma)} P(X \in S)$, on the probability $P(X \in S)$ for all random vectors X defined on $\mathcal{R}^n$ with mean $M = E[X]$ and covariance matrix $\Gamma = E[(X - M)(X - M)']$. If the set S is convex, we review the work of Marshall and Olkin [27], who solved the problem as a convex optimization problem. If the set S is a union of convex sets, over which we can optimize a convex quadratic function efficiently, we provide a polynomial time algorithm.

6.1. The $(n, 2, \mathcal{R}^n)$ -bound problem for a convex set. The following result is due to Marshall and Olkin [27], who give a constructive proof. An alternative, optimization based proof is provided in Popescu [33].

THEOREM 6.1 (Marshall and Olkin [27]). *The tight $(n, 2, \mathcal{R}^n)$ -upper bound for a convex event S is given by*

$$(6.1) \quad \sup_{X \sim (M, \Gamma)} P(X \in S) = \frac{1}{1 + d^2},$$

where $d^2 = \inf_{x \in S} (x - M)' \Gamma^{-1} (x - M)$ is the squared distance from M to the set S , under the norm induced by the matrix Γ^{-1} .

The actual formulation provided by Marshall and Olkin [27] is the following:

$$(6.2) \quad \sup_{X \sim (0, \Gamma)} P(X \in S) = \inf_{a \in S^\perp} \frac{1}{1 + (a' \Gamma a)^{-1}},$$

where $S^\perp = \{a \in \mathcal{R}^n \mid a'x \geq 1 \ \forall x \in S\}$, is the so-called antipolar of S (a.k.a “blocker”, or “upper-dual”). The above result is with zero mean, but can be easily extended for nonzero mean by a simple transformation (see [27, p. 1013, eqs. (7.8)–(7.9)], or the first part of the proof of Theorem 6.2). Given that $(a' \Gamma a)(x' \Gamma^{-1} x) \geq (a'x)^2 \geq 1$ for all $x \in S$, $a \in S^\perp$, one can easily see that the bound (6.1) is at least as tight as (6.2). Equality follows from nonlinear gauge duality principles (see Freund [8]).

Marshall and Olkin [27] construct an extremal distribution of a random variable $X \sim (M, \Gamma)$, so that $P(X \in S) = 1/(1 + d^2)$ with $d^2 = \inf_{x \in S} (x - M)' \Gamma^{-1} (x - M)$. We will say that the bound d is achievable, when there exists an $x^* \in S$ such that $d^2 = (x^* - M)' \Gamma^{-1} (x^* - M)$. In particular, d is achievable if the set S is closed.

THEOREM 6.2 (see [27]).

(a) *If $M \notin S$ and if $d^2 = \inf_{x \in S} (x - M)' \Gamma^{-1} (x - M)$ is achievable, then there is an extremal distribution that exactly achieves the bound (6.1).*

(b) *Otherwise, if $M \in S$ or if d^2 is not achievable, then there is a sequence of (M, Γ) -feasible distributions that asymptotically approach the bound (6.1).*

Bounds on tail probabilities. Given a vector $M = (M_1, \dots, M_n)'$, and an $n \times n$ positive definite, full rank matrix Γ , we derive next tight bounds on the following upper, lower, and two-sided tail probabilities of a random vector $X = (X_1, \dots, X_n)'$ with mean $M = E[X]$ and covariance matrix $\Gamma = E[(X - M)(X - M)']$:

$$P(X > M_{e+\delta}) = P(X_i > (1 + \delta_i)M_i \ \forall i = 1, \dots, n),$$

$$P(X < M_{e-\delta}) = P(X_i < (1 - \delta_i)M_i \ \forall i = 1, \dots, n),$$

$$P(X > M_{e+\delta} \text{ or } X < M_{e-\delta}) = P(X_i - M_i > \delta_i M_i \ \forall i, \text{ or } X_i - M_i < -\delta_i M_i \ \forall i),$$

where $\delta = (\delta_1, \dots, \delta_n)'$, and we denote $M_\delta = (\delta_1 M_1, \dots, \delta_n M_n)'$. In order to obtain nontrivial bounds we require that not all $\delta_i M_i \leq 0$, which expresses the fact that the tail event does not include the mean vector.

The one-sided Chebyshev inequality. In the following, we find a tight bound for $P(X > M_{e+\delta})$. The bound immediately extends to $P(X < M_{e-\delta})$.

PROPOSITION 6.3. (a) *The tight multivariate one-sided $(n, 2, \mathcal{R}^n)$ -Chebyshev bound is*

$$(6.3) \quad \sup_{X \sim (M, \Gamma)} P(X > M_{e+\delta}) = \frac{1}{1 + d^2},$$

where d^2 is given by

$$(6.4) \quad \begin{aligned} d^2 &= \min x' \Gamma^{-1} x \\ \text{subject to } x &\geq M_\delta, \end{aligned}$$

or alternatively d^2 is given by the gauge dual problem of (6.4)

$$(6.5) \quad \begin{aligned} \frac{1}{d^2} &= \min x' \Gamma x \\ \text{subject to } x' M_\delta &= 1, \\ x &\geq 0. \end{aligned}$$

(b) If $\Gamma^{-1} M_\delta \geq 0$, then the tight bound is expressible in closed form:

$$(6.6) \quad \sup_{X \sim (M, \Gamma)} P(X > M_{e+\delta}) = \frac{1}{1 + M'_\delta \Gamma^{-1} M_\delta}.$$

Proof. (a) Applying the bound (6.1) for $S = \{x \mid x_i > (1 + \delta_i) M_i \ \forall i = 1, \dots, n\}$, and changing variables we obtain (6.3). The alternative expression (6.5) for d^2 follows from elementary gauge duality theory (see Freund [8]).

(b) The Kuhn–Tucker conditions for problem (6.4) are as follows:

$$2\Gamma^{-1}x - \lambda = 0, \quad \lambda \geq 0, \quad x \geq M_\delta, \quad \lambda_i(x_i - (M_\delta)_i) = 0 \quad \forall i.$$

The choice $x = M_\delta$, $\lambda = 2\Gamma^{-1} M_\delta \geq 0$ (by assumption) satisfies the Kuhn–Tucker conditions, which are sufficient (this is a convex quadratic optimization problem). Thus, $d^2 = M'_\delta \Gamma^{-1} M_\delta$, and hence, (6.6) follows. $\square$

The two-sided Chebyshev inequality. The following result provides a tight bound for $P(X > M_{e+\delta} \text{ or } X < M_{e-\delta})$.

PROPOSITION 6.4 (see [27]).

(a) The tight multivariate two-sided $(n, 2, \mathcal{R}^n)$ -Chebyshev bound is

$$(6.7) \quad \sup_{X \sim (M, \Gamma)} P(X > M_{e+\delta} \text{ or } X < M_{e-\delta}) = \min(1, t^2),$$

where

$$(6.8) \quad \begin{aligned} t^2 &= \min x' \Gamma x \\ \text{subject to } x' M_\delta &= 1, \\ x &\geq 0. \end{aligned}$$

(b) If $\Gamma^{-1} M_\delta \geq 0$, then the tight bound is expressible in closed-form:

$$(6.9) \quad \sup_{X \sim (M, \Gamma)} P(X > M_{e+\delta} \text{ or } X < M_{e-\delta}) = \min \left(1, \frac{1}{M'_\delta \Gamma^{-1} M_\delta} \right).$$

This formulation follows via gauge duality principles from the original zero-mean result of Marshall and Olkin [28].

In the univariate case $M_\delta = \delta M$ and $\Gamma = \sigma^2$. Therefore, $\Gamma^{-1} M_\delta = \frac{\delta M}{\sigma^2} \geq 0$, and the closed-form bound applies, i.e.,

$$(6.10) \quad P(X > (1 + \delta)M) \leq \frac{C_M^2}{\delta^2 + C_M^2},$$

where $C_M^2 = \frac{\sigma^2}{M^2}$ is the coefficient of variation of the random variable X . The usual Chebyshev inequality is given by $P(X > (1 + \delta)M) \leq \frac{C_M^2}{\delta^2}$. Inequality (6.10) is always stronger. Moreover, there exist extremal distributions that satisfy it with equality (see Theorem 6.2). The original univariate result can be traced back to the 1937 book of Uspensky [46], and is mentioned later by Marshall and Olkin (1960) [27], [28], but has not received much attention in modern probability textbooks.

6.2. The tight $(n, 2, \mathcal{R}^n)$ -bound for unions of convex sets. We are given first- and second-order moment information (M, Γ) on the n -dimensional random variable X , and we would like to compute $\sup_{X \sim (M, \Gamma)} P(X \in S)$. Recall that the corresponding dual problem can be written as

$$(6.11) \quad \begin{aligned} Z_D = \min_{Y, y, y_0} \quad & Y \bullet \Gamma + y' M + y_0 \\ \text{subject to} \quad & g(x) = x' Y x + y' x + y_0 \geq \chi_S(x) \quad \forall x \in \mathcal{R}^n. \end{aligned}$$

We consider convex sets for which we can solve a convex quadratic optimization problem over them in polynomial time. Examples include polyhedra and sets defined by semidefinite constraints. The separation problem corresponding to problem (6.11) can be stated as follows: Given a matrix H , a vector c , and a scalar d , we need to check whether $g(x) = x' H x + c' x + d \geq \chi_S(x)$ for all $x \in \mathcal{R}^n$, and, if not, we need to find a violated inequality. Notice that we can assume without loss of generality that the matrix H is symmetric.

The following algorithm solves the separation problem in polynomial time.

ALGORITHM B.

1. If H is not positive semidefinite, then one can exhibit a polynomial size $x_0 \in \mathcal{Q}^n$ so that $x_0' H x_0 < -1$. Let $c_0 = c' x_0$, so $g(\lambda x_0) < -\lambda^2 + \lambda c_0 + d$ for any scalar λ . Choosing λ large enough (e.g., $\lambda > |c_0| + \sqrt{|d|}$) so that $g(\lambda x_0) < 0$, produces a violated inequality.
2. Otherwise, if H is positive semidefinite, then we do the following:
 - (a) We test if $g(x) \geq 0$ for all $x \in \mathcal{R}^n$ by solving the convex optimization problem

$$\inf_{x \in \mathcal{R}^n} g(x).$$

Let z_0 be the optimal value. If $z_0 < 0$, we find x_0 such that $g(x_0) < 0$, which represents a violated inequality.

- (b) Otherwise, we test if $g(x) \geq 1$ for all $x \in S$ by solving a polynomial collection of convex optimization problems

$$\inf_{x \in S} g(x).$$

Let z_1 be the optimal value. If $z_1 \geq 1$, then $g(x) \geq 1$ for all $x \in S$, and thus (H, c, d) is feasible. If not, we exhibit an x_1 such that $g(x_1) < 1$, and thus we identify a violated inequality.

Since we can solve the separation problem in polynomial time, we can also solve (within ϵ) the $(n, 2, \mathcal{R}^n)$ -bound problem in polynomial time (in the problem data and $\log \frac{1}{\epsilon}$).

7. Concluding remarks. In this paper we characterized the (n, k, Ω) -bound problem. We provided polynomial time algorithms via semidefinite programming algorithms for the $(1, k, \Omega)$ -bound problem, and via convex optimization methods, for the

$(n, 1, \mathcal{R}_+^n)$, $(n, 2, \mathcal{R}^n)$ -bound problems. We showed that the $(n, k, \mathcal{R}_+^n)$ and $(n, k, \mathcal{R}^n)$ -bound problems are NP-hard for $k \geq 2$, respectively $k \geq 4$. We also provided a family of semidefinite relaxations of polynomial size for the general (n, k, Ω) -bound problem that provide increasingly stronger bounds.

Acknowledgments. The authors thank Michael Todd and two reviewers for many valuable comments. We would like to thank Jim Dyer, Guillermo Gallego, James Primbs, Alexander Shapiro, and James Smith for pointers to the literature and for insightful comments on an earlier version of the paper.

REFERENCES

- [1] N. I. AKHIEZER, *The Classical Moment Problem*, Hafner, New York, 1965.
- [2] A. BEN-TAL AND A. NEMIROVSKI, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, MPS/SIAM Ser. Optim. 2, SIAM, Philadelphia, 2001.
- [3] F. BONNANS AND A. SHAPIRO, *Perturbation Analysis of Optimization Problems*, Springer, New York, 2000.
- [4] J. M. BORWEIN AND A. S. LEWIS, *Partially finite convex programming, part I: Quasi relative interiors and duality theory*, Math. Program., 57 (1992), pp. 15–48.
- [5] J. M. BORWEIN AND A. S. LEWIS, *Partially finite convex programming, part II: Explicit lattice models*, Math. Program., 57 (1992), pp. 49–83.
- [6] P. CHEBYSHEV, *Sur les valeurs limites des intégrales*, J. Math. Pure. Appl., 19 (1874), pp. 157–160.
- [7] P. DIACONIS, *Application of the method of moments in probability and statistics*, in Moments in Mathematics, Proc. Sympos. Appl. Math. 37, H. J. Landau, ed., AMS, Providence, RI, 1987, pp. 125–139.
- [8] R. FREUND, *Dual gauge programs, with applications to quadratic programming and the minimum-norm problem*, Math. Program., 38 (1987), pp. 47–67.
- [9] H. J. GODWIN, *On generalizations of Tchebycheff's inequality*, J. Amer. Statist. Assoc., 50 (1955), pp. 923–945.
- [10] H. J. GODWIN, *Inequalities on distribution functions*, in Griffin's Statistical Monographs and Courses 16, M. G. Kendall, ed., Charles Griffin and Co., London, 1964.
- [11] M. GRÖTSCHHEL, L. LOVÁSZ, AND A. SCHRIJVER, *Geometric Algorithms and Combinatorial Optimization*, Algorithms and Combinatorics, Springer, Berlin, Heidelberg, 1988.
- [12] H. HAMBURGER, *Über eine erweiterung der Stieltjes'schen momentenprobleme*, Math. Ann., 81 (1920), pp. 235–319.
- [13] H. HAMBURGER, *Über eine erweiterung der Stieltjes'schen momentenprobleme*, Math. Ann., 82 (1921), pp. 120–164, 168–187.
- [14] E. K. HAVILAND, *On the momentum problem for distributions in more than one dimension*, Amer. J. Math., 57 (1935), pp. 562–568.
- [15] E. K. HAVILAND, *On the momentum problem for distribution functions in more than one dimension*, Amer. J. Math., 58 (1936), pp. 164–168.
- [16] K. ISII, *The extrema of probability determined by generalized moments. I. Bounded random variables*, Ann. Inst. Statist. Math., 12 (1960), pp. 119–133.
- [17] K. ISII, *On the sharpness of Chebyshev-type inequalities*, Ann. Inst. Statist. Math., 14 (1963), pp. 185–197.
- [18] S. KARLIN AND L. S. SHAPLEY, *Geometry of moment spaces*, Mem. Amer. Math. Soc., 12 (1953).
- [19] S. KARLIN AND W. J. STUDDEN, *Tchebycheff Systems: With Applications in Analysis and Statistics*, Pure Appl. Math. 15, Interscience, John Wiley and Sons, New York, 1966.
- [20] J. H. B. KEMPERMAN, *The general moment problem, a geometric approach*, Ann. Math. Statist., 39 (1968), pp. 93–122.
- [21] J. H. B. KEMPERMAN, *On the role of duality in the theory of moments*, in Semi-infinite Programming and Applications, Lecture Notes in Econom. and Math. Systems 215, Springer, Berlin, New York, 1983, pp. 63–92.
- [22] J. H. B. KEMPERMAN, *Geometry of the moment problem*, in Moments in Mathematics, Proc. Sympos. Appl. Math. 37, H. J. Landau, ed., AMS, Providence, RI, 1987, pp. 16–53.
- [23] H. J. LANDAU, *Classical background on the moment problem*, in Moments in Mathematics, Proc. Sympos. Appl. Math. 37, H. J. Landau, ed., AMS, Providence, RI, 1987, pp. 1–15.

- [24] J. B. LASSERRE, *Global optimization with polynomials and the problem of moments*, SIAM J. Optim., 11 (2001), pp. 796–817.
- [25] A. MARKOV, *On Certain Applications of Algebraic Continued Fractions*, Ph.D. thesis, St. Petersburg, 1884 (in Russian).
- [26] A. MARSHALL, *Markov's inequality for random variables taking values in a linear topological space*, in *Inequalities in Statistics and Probability*, IMS Lecture Notes Monogr. Ser. 5, Y. L. Tong, ed., Institute of Mathematical Statistics, Hayward, CA, 1984, pp. 104–108.
- [27] A. MARSHALL AND I. OLKIN, *Multivariate Chebyshev inequalities*, Ann. Math. Statist., 31 (1960), pp. 1001–1014.
- [28] A. MARSHALL AND I. OLKIN, *A one-sided inequality of the Chebyshev type*, Ann. Math. Statist., 31 (1960), pp. 488–491.
- [29] K. G. MURTY AND S. N. KABADI, *Some NP-complete problems in quadratic and nonlinear programming*, Math. Program., 39 (1987), pp. 117–129.
- [30] Y. NESTEROV AND A. NEMIROVSKII, *Interior Point Polynomial Algorithms in Convex Programming*, SIAM Stud. Appl. Math. 13, SIAM, Philadelphia, 1994.
- [31] YU. NESTEROV, *Structure of Non-Negative Polynomial and Optimization Problems*, Preprint DP 9749, Louvain-la-Neuve, Belgium, 1997.
- [32] P. PARRILO, *Structured Semidefinite Programs and Semialgebraic Geometry Methods in Robustness and Optimization*, Ph.D. thesis, Caltech, 2000.
- [33] I. POPESCU, *Applications of Optimization in Probability, Finance and Revenue Management*, Ph.D. Dissertation, Applied Mathematics Department and Operations Research Center, MIT, Cambridge, MA, 1999.
- [34] M. PUTINAR, *Positive polynomials on compact semi-algebraic sets*, Indiana Univ. Math. J., 42 (1993), pp. 969–984.
- [35] B. REZNICK, *Some concrete aspects of Hilbert's 17th problem*, in *Real Algebraic Geometry and Ordered Structures*, Contemp. Math. 253, C. N. Delzell and J. J. Madden, eds., AMS, Providence, RI, 2000, pp. 251–272.
- [36] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [37] K. SCHMUDGEN, *The k -moment problem for compact semialgebraic sets*, Math. Ann., 289 (1991), pp. 203–206.
- [38] S. SENGUPTA, W. SIU, AND N. C. LIND, *A linear programming formulation of the problem of moments*, Z. Angew. Math. Mech., 10 (1979), pp. 533–537.
- [39] A. SHAPIRO, *On duality theory of conic linear problems*, in *Semi-Infinite Programming: Recent Advances*, Miguel A. Goberna and Marco A. López, eds., Kluwer, Massachusetts, 2001, pp. 135–165.
- [40] J. A. SHOHAT AND J. D. TAMARKIN, *The Problem of Moments*, AMS, New York, 1943.
- [41] N. Z. SHOR, *Class of global minimum bounds of polynomial functions*, Cybernetics, 23 (1987), pp. 731–734.
- [42] J. SMITH, *Generalized Chebyshev inequalities: Theory and applications in decision analysis*, Oper. Res., 43 (1995), pp. 807–825.
- [43] T. J. STIELTJES, *Recherches sur les fractions continues*, Ann. Fac. Sci. Toulouse Math. (6), 8 (1894), pp. 1–122.
- [44] T. J. STIELTJES, *Recherches sur les fractions continues*, Ann. Fac. Sci. Toulouse Math. (6), 9 (1895), pp. 5–47.
- [45] Y. L. TONG, *Probability Inequalities in Multivariate Distributions*, Academic Press, New York, 1980.
- [46] J. V. USPENSKY, *Introduction to Mathematical Probability*, McGraw-Hill, New York, 1937.
- [47] H. WOLKOWICZ, R. SAIGAL, AND L. VANDENBERGHE, EDS., *Handbook of Semidefinite Programming: Theory, Algorithms and Applications*, Kluwer, Massachusetts, 2000.

OPTIMIZATION OF POLYNOMIALS ON COMPACT SEMIALGEBRAIC SETS*

MARKUS SCHWEIGHOFER†

Abstract. A basic closed semialgebraic subset S of $\mathbb{R}^n$ is defined by simultaneous polynomial inequalities $g_1 \geq 0, \dots, g_m \geq 0$. We give a short introduction to Lasserre’s method for minimizing a polynomial f on a compact set S of this kind. It consists of successively solving tighter and tighter convex relaxations of this problem which can be formulated as semidefinite programs. We give a new short proof for the convergence of the optimal values of these relaxations to the infimum f^* of f on S which is constructive and elementary. In the case where f possesses a *unique* minimizer x^* , we prove that every sequence of “nearly” optimal solutions of the successive relaxations gives rise to a sequence of points in $\mathbb{R}^n$ converging to x^* .

Key words. nonconvex optimization, positive polynomial, sum of squares, moment problem, Positivstellensatz, semidefinite programming

AMS subject classifications. Primary 90C26, 13J30; Secondary 44A60, 14P10, 11E25

DOI. 10.1137/S1052623403431779

1. Introduction to Lasserre’s method. Throughout the paper, we suppose $1 \leq n \in \mathbb{N}$ and abbreviate $(X_1, \dots, X_n)$ by $\bar{X}$. We let $\mathbb{R}[\bar{X}]$ denote the ring of real polynomials in n indeterminates. Suppose we are given a so-called basic closed semialgebraic set, i.e., a set

$$S := \{x \in \mathbb{R}^n \mid g_1(x) \geq 0, \dots, g_m(x) \geq 0\}$$

defined by polynomials $g_1, \dots, g_m \in \mathbb{R}[\bar{X}]$. We consider the problem of minimizing a polynomial $f \in \mathbb{R}[\bar{X}]$ on S in the case where S is compact. So we are interested in computing numerically the optimal value

$$f^* := \inf\{f(x) \mid x \in S\} \in \mathbb{R} \cup \{\infty\}$$

and, if possible, an optimal point, i.e., an element of the set

$$S^* := \{x^* \in S \mid \forall x \in S : f(x^*) \leq f(x)\}.$$

Note that $S^* \neq \emptyset$ and $f^* = \min\{f(x) \mid x \in S\}$ unless $S = \emptyset$ since S is assumed to be compact.

We will now outline Lasserre’s approach [Las] to solve this problem. It will be convenient to set $g_0 := 1 \in \mathbb{R}[\bar{X}]$. Furthermore, we denote by $\mathbb{R}[\bar{X}]^2$ the set of all squares p^2 of polynomials $p \in \mathbb{R}[\bar{X}]$, by $\mathbb{R}[\bar{X}]^2 g_i$ the set of all $p^2 g_i$, by $\sum \mathbb{R}[\bar{X}]^2 g_i$ the

*Received by the editors July 19, 2003; accepted for publication (in revised form) August 23, 2004; published electronically April 8, 2005. This work was partially done during the author’s one-year stay at Université de Rennes 1 supported by the European network RAAG (EC contract HPRN-CT-2001-00271). Part of the work was also supported by the project “Darstellung positiver Polynome” (Deutsche Forschungsgemeinschaft, Kennung 214371).

<http://www.siam.org/journals/siopt/15-3/43177.html>

†Universität Konstanz, Fachbereich Mathematik und Statistik, 78457 Konstanz, Allemagne (Markus.Schweighofer@uni-konstanz.de).

set of all finite sums of such elements, and so on. In real algebra, the set

$$(1.1) \quad \begin{aligned} M &:= \sum \mathbb{R}[\bar{X}]^2 + \sum \mathbb{R}[\bar{X}]^2 g_1 + \cdots + \sum \mathbb{R}[\bar{X}]^2 g_m \\ &= \left\{ \sum_{i=0}^m \sigma_i g_i \mid \sigma_i \in \sum \mathbb{R}[\bar{X}]^2 \right\} \subseteq \mathbb{R}[\bar{X}] \end{aligned}$$

is called the quadratic module generated by $g_1, \dots, g_m$ [PD, 5.1.7]. It is a set of polynomials that are nonnegative on S and possess a very nice certificate for this property. Lasserre's method works only if a certain assumption is satisfied. It can be expressed by the following theorem in several different ways that are equivalent.

THEOREM 1.1 (Schmüdgen). *The following are equivalent:*

- (i) *There exist finitely many $p_1, \dots, p_s \in M$ such that the set*

$$\{x \in \mathbb{R}^n \mid p_1(x) \geq 0, \dots, p_s(x) \geq 0\}$$

(which contains S) is compact and $\prod_{i \in I} p_i \in M$ for all $I \subseteq \{1, \dots, s\}$.

- (ii) *There exists some $p \in M$ such that $\{x \in \mathbb{R}^n \mid p(x) \geq 0\}$ is compact.*

- (iii) *There exists an $N \in \mathbb{N}$ such that $N - \sum_{i=1}^n X_i^2 \in M$.*

- (iv) *For all $p \in \mathbb{R}[\bar{X}]$, there is some $N \in \mathbb{N}$ such that $N \pm p \in M$.*

It is clear that (iv) $\implies$ (iii) $\implies$ (ii) $\implies$ (i). We don't include the easy proof of (iii) $\implies$ (iv) [PD, 5.1.13], [S1, 2.1] since this implication comes out as a byproduct in this article: Soon we will impose (iii) as a general assumption under which we will be able to prove Theorem 1.3 below, which together with the compactness of S implies immediately (iv).

The essential part of Theorem 1.1 is the implication (i) $\implies$ (iii). This is a weak form of Schmüdgen's theorem [Sch, Corollary 3] because $N - \sum_{i=1}^n X_i^2 > 0$ holds on the compact set defined by the p_i when $N \in \mathbb{N}$ is big enough. The reader acquainted with real algebra should not be confused by the fact that in Schmüdgen's work there is no quadratic module appearing: When proving (i) $\implies$ (iii) one can assume without loss of generality that M is even a so-called *preorder* (the one generated by the p_i). We will not include a proof of this weak form of Schmüdgen's theorem and instead state our assumption in the seemingly stronger form (iii). We do this for the following reasons:

- Once assuming (iii), we have a new short proof of Theorem 1.3 (and thus the convergence of Lasserre's procedure to f^*) that uses only elementary analysis (see section 2). It completely avoids any arguments from functional analysis used in Putinar's original proof [Put] or from real algebra used in the other proof of Jacobi [Jac], [PD].

- In many practical applications, an $N \in \mathbb{N}$ is known such that S is contained in the closed ball around 0 with radius $\sqrt{N}$. Then the condition (iii) can be assured by adding the redundant inequality $N - \sum_{i=1}^n X_i^2 \geq 0$ in the definition of S . This increases the number m of g_i 's by one.

- Whether it is decidable that the equivalent conditions of Theorem 1.1 are satisfied is not known and is the subject of current research. But (iii) seems to be the condition most suitable for algorithmic exploitation among the given ones.

- Condition (i) is interesting if it is known that S is compact, but no ball containing S is known explicitly. It can be satisfied by adding the redundant inequalities $\prod_{i \in I} g_i \geq 0$ for all $I \subseteq \{1, \dots, m\}$ of cardinality $\#I \geq 2$ to the definition of S . However, this gives rise to an exponential growth of the number of inequalities, which is fatal from the complexity point of view unless the number m of g_i 's is very small.

We refer readers interested in the proof of the implication (i) $\implies$ (iii) to [PD, 5.1.17] or [S1, 2.2]. These are expositions of Wörmann's nice algebraic proof [BW, Theorem 4] that don't use a sledgehammer to crack a nut (the sledgehammer is [BW, Corollary 1]). Theorem 1.1 was first proved by Schmüdgen in his pioneering work [Sch], where he proved even Putinar's Theorems 1.2 and 1.3 in the special case where $g_1, \dots, g_m$ are the $m = 2^s$ products $\prod_{i \in I} p_i$ of some polynomials $p_1, \dots, p_s$. In this context, we should also mention that deep work of Jacobi and Prestel gives criteria for the conditions of Theorem 1.1 to be satisfied that are even more applicable than condition (i) [JP], [PD, 6.3]. We end the discussion around Theorem 1.1 once and for all by assuming from now on the following condition, which is really stronger than the compactness of S [PD, 6.3.1].

General assumption.

$$\exists N \in \mathbb{N} : N - \sum_{i=1}^n X_i^2 \in M$$

A fundamental obstacle when one tries to find f^* is that the set S and the function $f|_S : S \rightarrow \mathbb{R}$ are in general far from being convex. Of course, S need not even be connected. The first step towards Lasserre's method is to convexify the problem by brute force. There are two ways of doing this.

The first one is to exchange the points of S by probability measures on S . Every point $x \in S$ can be identified with the Dirac measure δ_x at x , and a convex combination of probability measures is again a probability measure. Obviously, we have

$$(1.2) \quad f^* = \inf \left\{ \int f d\mu \mid \mu \in \mathcal{M}^1(S) \right\}.$$

Here, for any set $A \subseteq \mathbb{R}^n$, $\mathcal{M}^1(A)$ denotes the set of all probability measures on A . To be precise, A is understood as a measurable space. More exactly, it is equipped with the σ -algebra $\mathcal{B}(A)$ of its Borel sets, i.e., the σ -algebra generated by all sets that are open in the topological space A . Consequently, a probability measure μ on A is a σ -additive map $\mu : \mathcal{B}(A) \rightarrow [0, \infty)$ with $\mu(\emptyset) = 0$ and $\mu(A) = 1$.

The second method of convexification is to take a dual standpoint, namely, to write

$$(1.3) \quad f^* = \sup \{a \in \mathbb{R} \mid f - a \geq 0 \text{ on } S\} = \sup \{a \in \mathbb{R} \mid f - a > 0 \text{ on } S\}.$$

In (1.2) and (1.3), we got rid of the usually nonconvex set S by introducing a convex object which is, however, very hard to describe and not suitable for algorithmic purposes—namely, the set $\mathcal{M}^1(S)$ of probability measures on S and the set of polynomials nonnegative (or positive) on S , respectively. The following key theorems of Putinar [Put] will help us to overcome this problem.

THEOREM 1.2 (Putinar). *For any map $L : \mathbb{R}[\bar{X}] \rightarrow \mathbb{R}$, the following are equivalent:*

- (i) *L is linear, $L(1) = 1$, and $L(M) \subseteq [0, \infty)$.*
- (ii) *L is integration with respect to a probability measure μ on S , i.e.,*

$$\exists \mu \in \mathcal{M}^1(S) : \forall p \in \mathbb{R}[\bar{X}] : L(p) = \int p d\mu.$$

THEOREM 1.3 (Putinar). *If $p \in \mathbb{R}[\bar{X}]$ satisfies $p > 0$ on S , then $p \in M$.*

Theorem 1.2 does not really describe $\mathcal{M}^1(S)$, but rather all real families $(a_\alpha)_{\alpha \in \mathbb{N}^n}$ that are the sequence of moments of a probability measure on S , that is,

$$a_\alpha = \int \bar{X}^\alpha d\mu \quad \text{for all } \alpha \in \mathbb{N}^n,$$

where $\bar{X}^\alpha := X_1^{\alpha_1} \cdots X_n^{\alpha_n}$. This is clear from the fact that every linear map $L : \mathbb{R}[\bar{X}] \rightarrow \mathbb{R}$ is given uniquely by its values $L(\bar{X}^\alpha)$ on the basis $(\bar{X}^\alpha)_{\alpha \in \mathbb{N}^n}$ of $\mathbb{R}[\bar{X}]$. One says therefore that Theorem 1.2 is a solution to the *moment problem* on S . For our purposes this is sufficient since the integral of a polynomial depends only on the moments, and we can rewrite (1.2) as

$$(1.4) \quad f^* = \inf\{L(f) \mid L : \mathbb{R}[\bar{X}] \rightarrow \mathbb{R} \text{ is linear, } L(1) = 1, L(M) \subseteq [0, \infty)\}.$$

Remark 1.4. Some articles (e.g., [Put]) prefer the viewpoint of a linear map $L : \mathbb{R}[\bar{X}] \rightarrow \mathbb{R}$ as a family $(L(\bar{X}^\alpha))_{\alpha \in \mathbb{N}^n}$. In other words, they use an infinite matrix representation of L (with respect to the standard basis of $\mathbb{R}[\bar{X}]$). Sometimes this is useful, as we will see (especially in the proof of Theorem 3.4).

A third way of seeing such an L is by looking at the bilinear forms

$$\mathbb{R}[\bar{X}] \times \mathbb{R}[\bar{X}] \rightarrow \mathbb{R} : (p, q) \mapsto L(pq_i)$$

(compare proof of Lemma 5.1) for $L(M) \subseteq [0, \infty)$ means nothing else than that these bilinear forms are all positive semidefinite. Again, an infinite matrix representation of these bilinear forms can be useful; see, for example, [Lau].

Theorem 1.3 is called Putinar's *Positivstellensatz* in analogy to Hilbert's Nullstellensatz. Of course, it does not really characterize the polynomials positive on S since the polynomials lying in M must only be nonnegative on S but not positive. Also, it does not fully describe the polynomials nonnegative on S since they are not always contained in M [PD, example before 8.2.3]. Nevertheless, we get together with (1.3) that

$$(1.5) \quad f^* = \sup\{a \in \mathbb{R} \mid f - a \in M\}.$$

Putinar proved first Theorem 1.2 using tools from functional analysis and then deduced Theorem 1.3 from it. Jacobi found an algebraic approach to these theorems. He proved an abstract generalization of Putinar's Positivstellensatz using tools of real algebra [Jac], [PD, 5.3.7]. (In fact, the author observed that it is not hard to get that abstract theorem back from Theorem 1.3 using compactness arguments.) Putinar's solution to the moment problem can then be derived from his Positivstellensatz using the well-known Riesz representation theorem (see section 3).

In section 2, we shall present a new proof of Putinar's Positivstellensatz that uses only elementary arguments. Like an earlier approach of the author to Schmüdgen's Positivstellensatz [S1], it is based on an old theorem of Pólya, Theorem 2.1 below. At first glance, this earlier approach seems to be inferior to the one we shall take here, since it works only for Schmüdgen's Positivstellensatz, which is a special case of Theorem 1.3. On the other hand, it led to complexity results [S2] that we were, despite all efforts, unable to reproduce in the more general situation that we encounter here.

The idea is now to relax (1.4) and (1.5) by introducing approximations $M_k \subseteq \mathbb{R}[\bar{X}]_k$ of $M \subseteq \mathbb{R}[\bar{X}]$. Here, for $k \in \mathbb{N}$, $\mathbb{R}[\bar{X}]_k$ denotes the $\mathbb{R}$ -vector space consisting of

the polynomials $p \in \mathbb{R}[\bar{X}]$ of degree $\deg p \leq k$ (we set $\deg 0 := -\infty$). So we define, in analogy to (1.1),

$$(1.6) \quad \begin{aligned} M_k &:= \sum \mathbb{R}[\bar{X}]_{d_0}^2 + \sum \mathbb{R}[\bar{X}]_{d_1}^2 g_1 + \cdots + \sum \mathbb{R}[\bar{X}]_{d_m}^2 g_m \\ &= \left\{ \sum_{i=0}^m \sigma_i g_i \mid \sigma_i \in \sum \mathbb{R}[\bar{X}]^2, \deg(\sigma_i g_i) \leq k \right\} \end{aligned}$$

for arbitrary

$$(1.7) \quad k \in \mathcal{N} := \{s \in \mathbb{N} \mid s \geq \max\{\deg g_1, \dots, \deg g_m, \deg f\}\}.$$

Here the d_i depend on k and $\deg g_i$, more exactly,

$$(1.8) \quad d_i := \max\{d \in \mathbb{N} \mid 2d + \deg g_i \leq k\}$$

if $g_i \neq 0$, and (for instance) $d_i := 0$ otherwise. The equality in (1.6) follows from the fact that the leading form (the highest homogeneous part, i.e., the sum of the monomials of highest degree) of a nonzero square polynomial is a square polynomial and a fortiori a globally nonnegative polynomial. Hence, when adding square polynomials, no cancellation of leading forms can occur because the sum of two or more nonzero globally nonnegative polynomials cannot vanish. In contrast to this, cancellation of leading forms of the terms in the sum $\sum_{i=0}^m \sigma_i g_i$ is a frequent phenomenon on which the validity of Theorem 1.3 relies. In the expression of a given polynomial p as such a sum, the degree of the terms of the sum sometimes necessarily has to exceed $\deg p$ enormously [Ste] (see also [S2], [PD, 8.3]). So M_k should never be confused with $M \cap \mathbb{R}[\bar{X}]_k \supseteq M_k$.

Replacing M by M_k in (1.4) and (1.5) motivates the consideration of the following pair of optimization problems for each $k \in \mathcal{N}$:

$$\begin{aligned} (P_k) \quad & \text{minimize} \quad L(f) \quad \text{subject to} \quad \begin{aligned} & L : \mathbb{R}[\bar{X}]_k \rightarrow \mathbb{R} \text{ is linear,} \\ & L(1) = 1 \text{ and} \\ & L(M_k) \subseteq [0, \infty), \end{aligned} \\ (D_k) \quad & \text{maximize} \quad a \quad \text{subject to} \quad \begin{aligned} & a \in \mathbb{R} \text{ and} \\ & f - a \in M_k. \end{aligned} \end{aligned}$$

We call (P_k) and (D_k) the *primal* and *dual* relaxation of *order* $k \in \mathcal{N}$, respectively. The *optimal value* of (P_k) , i.e., the infimum over all $L(f)$ where L ranges over all feasible solutions of (P_k) , is denoted by $P_k^* \in \mathbb{R} \cup \{\pm\infty\}$. Analogously, we write $D_k^* \in \mathbb{R} \cup \{\pm\infty\}$ for the *optimal value* of (D_k) , i.e., the supremum of the feasible set $\{a \in \mathbb{R} \mid f - a \in M_k\}$ of (D_k) .

For every $x \in S$, the evaluation at x ,

$$\varepsilon_x : \mathbb{R}[\bar{X}]_k \rightarrow \mathbb{R} : p \mapsto p(x),$$

is a feasible solution of (P_k) . This shows that $P_k^* \leq f^*$. Moreover, if L is feasible for (P_k) and a for (D_k) , then $L(f) \geq a$ since $f - a \in M_k$ implies $L(f) - a = L(f) - aL(1) = L(f - a) \geq 0$. We therefore also get $D_k^* \leq P_k^*$. It is easy to see that every feasible solution $a \in \mathbb{R}$ of (D_k) is also feasible for (D_{k+1}) , and every feasible solution L of (P_{k+1}) is feasible for (P_k) when restricted to the subspace $\mathbb{R}[\bar{X}]_k$ of $\mathbb{R}[\bar{X}]_{k+1}$ ($k \in \mathcal{N}$). This tells us that the sequences $(D_k)_{k \in \mathcal{N}}$ and $(P_k)_{k \in \mathcal{N}}$ are increasing. Furthermore,

$(D_k^*)_{k \in \mathcal{N}}$ (and, consequently, $(P_k^*)_{k \in \mathcal{N}}$) converges to f^* . Indeed, for any $\varepsilon > 0$, we have for sufficiently large $k \in \mathcal{N}$ that $f - f^* + \varepsilon \in M_k$ by Theorem 1.3; i.e., $f^* - \varepsilon$ is feasible for (D_k) , whence $f^* - \varepsilon \leq D_k^*$. Summarizing, we have seen that the following is true.

THEOREM 1.5 (Lasserre). *$(D_k^*)_{k \in \mathcal{N}}$ and $(P_k^*)_{k \in \mathcal{N}}$ are increasing sequences that converge to f^* and satisfy $D_k^* \leq P_k^* \leq f^*$ for all $k \in \mathcal{N}$.*

In special cases, something has been proved about the rate of convergence. For example, if $S = \{x \in \mathbb{R}^n \mid g(x) \geq 0\}$ is defined by only one polynomial $g \in \mathbb{R}[\bar{X}]$, then there is a constant $c \in \mathbb{N}$ depending on f and g and a constant $0 < d \in \mathbb{N}$ depending on g such that

$$f^* - D_k^* \leq \frac{c}{\sqrt[d]{k}} \quad \text{for big } k.$$

Much more precise information can be found in [S2, section 2]. It is based on the author's earlier constructive approach to representations of positive polynomials [S1]. The new approach taken here is simpler and more general as long as one is interested only in the mere existence of representations of positive polynomials. But on the other hand, it seems to be less suited to the purpose of keeping track of complexity. At least we did not succeed in proving a similar rate of convergence for the general case.

Lasserre observed that (P_k) and (D_k) (when modified insignificantly) can be easily translated into semidefinite programs and are as such dual to each other. We will recapitulate this in section 5. Semidefinite programs are generalizations of linear programs and can be solved efficiently. Currently, there exist two implementations of the method [HL], [PPP]. Practical experience shows that f^* is often actually reached by D_k^* and P_k^* already for small $k \in \mathcal{N}$.

If $S \neq \emptyset$, one is usually interested not only in finding the minimum value f^* of f on S , but also in obtaining a minimizer $x^* \in S^*$. Section 3 deals with this problem. Speaking very roughly, Theorem 3.4 shows that the method allows for getting finite families of real numbers that are arbitrarily close to being the family of moments up to some fixed order of some probability measure on S^* . In case $S^* = \{x^*\}$ is a singleton, i.e., f possesses a *unique* minimizer, there is only one probability measure on S^* . Its moments of order one are the coordinates of the unique minimizer x^* , and we can consequently approximate them arbitrarily closely (see Corollary 3.5). So the method is good for computing minimizers when they are unique. As soon as there are two or more minimizers, symmetries in the problem, which often occur in practice, frequently prevent the algorithm from steering toward one of the minimizers. In the proof of Theorem 3.4, we use Theorem 1.2. For the convenience of the reader, we include the brief well-known derivation of Theorem 1.2 from Theorem 1.3 at the beginning of section 3.

In section 4, we investigate the duality of the problems (P_k) and (D_k) . If S has nonempty interior, then there is no duality gap, i.e., $P_k^* = D_k^*$ for all $k \in \mathcal{N}$. This result is essentially due to Lasserre, who proved it by applying the duality theory of semidefinite programming. However, the duality of the problems (P_k) and (D_k) is much easier to understand than the duality theory of semidefinite programming in general. We try to give the reader a good visual image of this duality, and then take Marshall's approach [Mar] to it, which in our opinion gives a much better understanding than looking at the duality in general semidefinite programming. We liberate Marshall's exposition from arguments using algebraic geometry and present

it in a less technical framework. The only really new material in section 4 are some negative examples in the case where S has no interior points.

2. Convergence to the infimum: A new proof of Putinar's theorem.

In this section, we present our new proof of Theorem 1.3. Thinking in terms of optimization, this means that we prove Theorem 1.5, that is, the convergence of $(D_k^*)_{k \in \mathcal{N}}$ (and hence of $(P_k^*)_{k \in \mathcal{N}}$) to f^* . The key is the following theorem proved by Pólya in 1928 [Pól]. Like the proof of all the other statements in this section, its proof requires only very elementary analysis. We recommend [PR] as a reference where, in addition, a bound on the exponent k can be found.

THEOREM 2.1 (Pólya). *Suppose $F \in \mathbb{R}[\bar{X}]$ is homogeneous and satisfies*

$$F > 0 \quad \text{on } [0, \infty)^n \setminus \{0\}.$$

Then for all big enough k , the polynomial

$$(X_1 + \cdots + X_n)^k F$$

has only nonnegative coefficients.

LEMMA 2.2. *For all $k \in \mathbb{N}$ and $y \in [0, 1]$,*

$$(y - 1)^{2k} y \leq \frac{1}{2k + 1}.$$

Proof. In fact, by elementary calculus, we have even

$$(y - 1)^{2k} y \leq \left(\frac{1}{2k + 1} - 1 \right)^{2k} \frac{1}{2k + 1} \leq \frac{1}{2k + 1}. \quad \square$$

Suppose now that we are given some $p \in \mathbb{R}[\bar{X}]$ with $p > 0$ on S . We want to show that it lies in M . The idea is to subtract first from p a certain element of M in order to extend the positivity condition from S to a much larger set C . To understand the next lemma, we propose considering the pointwise limit of the left-hand side of (2.1) below separately at points of $C \cap S$ and at points of $C \setminus S$, using the hypothesis that $g_i \leq 1$ on C together with the preceding lemma.

LEMMA 2.3. *Suppose $C \subseteq \mathbb{R}^n$ is compact and $g_i \leq 1$ on C for all $i \in \{1, \dots, m\}$. Suppose $p \in \mathbb{R}[\bar{X}]$ satisfies $p > 0$ on S . Then there exists $s \in \mathbb{N}$ such that for all sufficiently large $k \in \mathbb{N}$,*

$$(2.1) \quad p - s \sum_{i=1}^m (1 - g_i)^{2k} g_i > 0 \quad \text{on } C.$$

Proof. Since S is compact, we can choose $\varepsilon > 0$ such that $p > \varepsilon$ on S . Set

$$A := \{x \in C \mid p(x) \leq \varepsilon\}.$$

Then A is compact and disjoint to S , whence we can choose $\delta > 0$ such that

$$(2.2) \quad \min\{g_1(x), \dots, g_m(x)\} \leq -\delta \quad \text{for } x \in A.$$

It suffices to show that for $s, k \in \mathbb{N}$ satisfying

$$(2.3) \quad p + \frac{s\delta}{2} > 0 \quad \text{on } C,$$

$$(2.4) \quad \frac{\delta}{2} \geq \frac{m-1}{2k+1},$$

and

$$(2.5) \quad \varepsilon \geq \frac{sm}{2k+1},$$

we have (2.1). If $x \in A$, then in the sum

$$(2.6) \quad \sum_{i=1}^m (1 - g_i(x))^{2k} g_i(x)$$

at most $m - 1$ terms are nonnegative. By Lemma 2.2, these nonnegative terms add up to at most $\frac{m-1}{2k+1}$. At least one term is negative, even $\leq -\delta$ by (2.2). All in all, if we evaluate the left-hand side of our claim (2.1) at a point $x \in A$, then it is

$$\geq p(x) - s \frac{m-1}{2k+1} + s\delta \geq \underbrace{p(x) + \frac{s\delta}{2}}_{>0 \text{ by (2.3)}} + s \underbrace{\left(\frac{\delta}{2} - \frac{m-1}{2k+1} \right)}_{\geq 0 \text{ by (2.4)}} > 0.$$

When we evaluate it at a point $x \in C \setminus A$, all terms of the sum (2.6) might happen to be nonnegative. Again by Lemma 2.2, they add up to at most $\frac{m}{2k+1}$. But at the same time, the definition of A gives us a good lower bound on $p(x)$, so that the result is

$$> \varepsilon - s \frac{m}{2k+1} \geq 0$$

by (2.5). $\square$

Now we are able to prove Putinar's Positivstellensatz.

Proof of Theorem 1.3. In this proof, N is chosen as in the general assumption we imposed in section 1. Consider the compact set

$$\Delta := \left\{ y \in [0, \infty)^{2n} \mid y_1 + \cdots + y_{2n} = 2n \left(N + \frac{1}{4} \right) \right\} \subseteq \mathbb{R}^{2n},$$

and let $C := l(\Delta) \subseteq \mathbb{R}^n$ be its image under the linear map

$$l : \mathbb{R}^{2n} \rightarrow \mathbb{R}^n : y \mapsto \left(\frac{y_1 - y_{n+1}}{2}, \dots, \frac{y_n - y_{2n}}{2} \right).$$

Since $l(\Delta)$ is compact, we can scale each g_i with a positive factor such that $g_i \leq 1$ on C . Thereby no generality is lost since this affects neither S nor M (noting that all positive real numbers are squares). So we can apply the preceding lemma and get $s, k \in \mathbb{N}$ such that

$$q := p - s \sum_{i=1}^m (g_i - 1)^{2k} g_i > 0 \quad \text{on } C.$$

It suffices to show that $q \in M$, and we shall even show that

$$q \in T := \sum \mathbb{R}[\bar{X}]^2 + \sum \mathbb{R}[\bar{X}]^2 \left(N - \sum_{i=1}^n X_i^2 \right) \subseteq M.$$

To start with, write

$$q = \sum_{i=0}^d Q_i,$$

where $d := \deg q$ and $Q_i \in \mathbb{R}[\bar{X}]$ is homogeneous of degree i , $i = 0, \dots, d$. Now define

$$F := \sum_{i=0}^d Q_i \left(\frac{Y_1 - Y_{n+1}}{2}, \dots, \frac{Y_n - Y_{2n}}{2} \right) \left(\frac{Y_1 + \dots + Y_{2n}}{2n(N + \frac{1}{4})} \right)^{d-i} \in \mathbb{R}[\bar{Y}],$$

where $\mathbb{R}[\bar{Y}]$ denotes the polynomial ring in $2n$ new indeterminates $Y_1, \dots, Y_{2n}$. For each $y \in \Delta$, we obtain

$$F(y) = \sum_{i=0}^d Q_i(l(y)) = q(l(y)) > 0,$$

since $l(y) \in l(\Delta) = C$. Since F is a homogeneous polynomial, it has constant sign on each ray emanating by the origin, whence

$$F > 0 \quad \text{on } [0, \infty)^{2n} \setminus \{0\}.$$

By Pólya's Theorem 2.1, there is some $e \in \mathbb{N}$ such that

$$G := \left(\frac{Y_1 + \dots + Y_{2n}}{2n(N + \frac{1}{4})} \right)^e F \in \mathbb{R}[\bar{Y}]$$

has only nonnegative coefficients. Now we apply on this polynomial the $\mathbb{R}$ -algebra homomorphism $\varphi : \mathbb{R}[\bar{Y}] \rightarrow \mathbb{R}[\bar{X}]$ defined by

$$Y_i \mapsto \left(N + \frac{1}{4} \right) + X_i, \quad Y_{n+i} \mapsto \left(N + \frac{1}{4} \right) - X_i \quad (i \in \{1, \dots, n\}).$$

Note that $\varphi(Y_i) \in T$ for each $i \in \{1, \dots, 2n\}$ since

$$\left(N + \frac{1}{4} \right) \pm X_i = \sum_{j \neq i} X_j^2 + \left(X_i \pm \frac{1}{2} \right)^2 + \left(N - \sum_{j=1}^n X_j^2 \right) \in T.$$

Noting that T is closed under addition and multiplication, we therefore obtain that

$$\varphi(G) = \varphi(F) = \sum_{i=0}^d Q_i = q$$

is contained in T . $\square$

3. Convergence to the unique minimizer. Recall that $\mathcal{N}$ is the set of k for which we have defined the relaxations (P_k) and (D_k) ; see (1.7) on page 809. If, for some $k \in \mathcal{N}$, (P_k) happens to possess an optimal solution L that comes from a probability measure μ on S , that is,

$$\forall p \in \mathbb{R}[\bar{X}]_k : L(p) = \int p d\mu,$$

then we can conclude that

$$f^* \leq \int f d\mu = L(f) \leq f^*.$$

Consequently, $L(f) = f^*$ and $\mu(S^*) = 1$. If $S^* = \{x^*\}$ is a singleton, then μ must be the Dirac measure at x^* and

$$(L(X_1), \dots, L(X_n)) = \left(\int X_1 d\mu, \dots, \int X_n d\mu \right) = x^*.$$

In this section, we shall prove that, for high $k \in \mathcal{N}$, we are never terribly far from this nice situation (see Theorem 3.4). The important Corollary 3.5 tells us that Lasserre's procedure converges not only to the infimum f^* , but also to the minimizer x^* in the case that it is unique. These results are based on Theorem 1.2, whose well-established proof we include.

Proof of Theorem 1.2. The implication (ii) $\implies$ (i) is trivial. To show the reverse, suppose that (i) holds. Consider the ring homomorphism

$$\varphi : \mathbb{R}[\bar{X}] \rightarrow \mathcal{C}(S, \mathbb{R}) : p \mapsto p|_S$$

from the polynomial ring into the ring $\mathcal{C}(S, \mathbb{R})$ of continuous real-valued functions on S . Suppose $p \in \mathbb{R}[\bar{X}]$ satisfies $p \geq 0$ on S . Then $p + \varepsilon \in M$ by Theorem 1.3 and a fortiori $L(p) + \varepsilon = L(p + \varepsilon) \geq 0$ for all $\varepsilon > 0$. This implies $L(p) \geq 0$. In particular, L vanishes on the kernel of φ and therefore induces a linear map $\bar{L} : \varphi(\mathbb{R}[\bar{X}]) \rightarrow \mathbb{R}$ well defined by $\bar{L}(\varphi(p)) := L(p)$ for all $p \in \mathbb{R}[\bar{X}]$. We equip $\mathcal{C}(S, \mathbb{R})$ with the supremum norm and thus turn it into a normed vector space, noting that $S = \emptyset$ would imply $-1 \in M$, whence $-1 = -L(1) = L(-1) \geq 0$. By the Stone–Weierstrass approximation theorem, $\varphi(\mathbb{R}[\bar{X}])$ lies dense in $\mathcal{C}(S, \mathbb{R})$. It is easy to see that $\bar{L}(\varphi(p)) = L(p) \leq \|p\|$ for all $p \in \mathbb{R}[\bar{X}]$. Hence the linear map $\bar{L}$ is (uniformly) continuous. But every map uniformly continuous on a subspace of a metric space extends uniquely to a continuous map on the closure of this subspace. Therefore we may consider $\bar{L}$ as a continuous map on the whole of $\mathcal{C}(S, \mathbb{R})$. Using again the Stone–Weierstrass theorem, it is easy to see that $\bar{L}$ maps $\mathcal{C}(S, [0, \infty))$ into $[0, \infty)$. Since S is compact, the Riesz representation theorem [Rud, 2.14] tells us that $\bar{L}$ is integration with respect to a measure on S . $\square$

DEFINITION 3.1. We say that L_k solves (P_k) nearly to optimality ($k \in \mathcal{N}$) if L_k is a feasible solution of (P_k) ($k \in \mathcal{N}$) such that $\lim_{k \rightarrow \infty} L_k(f) = \lim_{k \rightarrow \infty} P_k^*$.

If L_k is an optimal solution of (P_k) for every $k \in \mathcal{N}$, then, of course, L_k solves (P_k) nearly to optimality. The same applies, for example, if $|L_k(f) - P_k^*| \leq \frac{1}{k}$ for all big k . This is a useful notion because (P_k) might not possess an optimal solution (see Example 4.8 below) and even if it does, we might not be able to compute it exactly. In view of Theorem 1.5, we note for later reference the following.

Remark 3.2. Suppose L_k is a feasible solution of (P_k) ($k \in \mathcal{N}$). Then L_k solves (P_k) nearly to optimality ($k \in \mathcal{N}$) if and only if $\lim_{k \rightarrow \infty} L_k(f) = f^*$.

Notation 3.3. For $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{N}^n$, we set

$$|\alpha| := \alpha_1 + \dots + \alpha_n$$

and, for $d \in \mathbb{N}$, we introduce the notation

$$\Lambda(d) := \{\alpha \in \mathbb{N}^n \mid |\alpha| \leq d\}.$$

Note that $(\bar{X}^\alpha)_{\alpha \in \Lambda(d)}$ is a basis of $\mathbb{R}[\bar{X}]_d$ for each $d \in \mathbb{N}$. We now give the exact formulation of the main theorem in this section. The norm on $\mathbb{R}^{\Lambda(d)}$ appearing in the theorem can be any norm.

THEOREM 3.4. Suppose that $S \neq \emptyset$ and L_k solves (P_k) nearly to optimality ($k \in \mathcal{N}$). Then

$$\forall d \in \mathbb{N} : \forall \varepsilon > 0 : \exists k_0 \in \mathcal{N} \cap [d, \infty) : \forall k \geq k_0 : \exists \mu \in \mathcal{M}^1(S^*) : \\ \left\| \left(L_k(\bar{X}^\alpha) - \int \bar{X}^\alpha d\mu \right)_{\alpha \in \Lambda(d)} \right\| < \varepsilon.$$

Proof. Let $d \in \mathbb{N}$ and $\varepsilon > 0$ be given. For each $k \in \mathcal{N}$, there is some $N_k \in \mathbb{N}$ such that

$$N_k \pm \bar{X}^\alpha \in M_{N_k} \quad \forall \alpha \in \mathbb{N}^n \text{ with } |\alpha| = k$$

by (iv) of Theorem 1.1 (note that $M = \bigcup_{k \in \mathcal{N}} M_k$ and $N_0 \pm \bar{X}^\alpha \in M_k$ implies $N \pm \bar{X}^\alpha \in M_N$ for all $N \geq \max\{N_0, k\}$). The topological space

$$Z := \prod_{\alpha \in \mathbb{N}^n} [-N_{|\alpha|}, N_{|\alpha|}]$$

is compact by Tychonoff's theorem. Now we have that

$$(3.1) \quad \{(a_\alpha)_{\alpha \in \mathbb{N}^n} \in Z \mid a_0 = 1\} \\ \cap \bigcap_{p \in M} \left\{ (a_\alpha)_{\alpha \in \mathbb{N}^n} \in Z \mid p = \sum_{\alpha} b_\alpha \bar{X}^\alpha, b_\alpha \in \mathbb{R}, \sum_{\alpha} b_\alpha a_\alpha \geq 0 \right\} \\ \cap \bigcap_{\mu \in \mathcal{M}^1(S^*)} \left\{ (a_\alpha)_{\alpha \in \mathbb{N}^n} \in Z \mid \left\| \left(a_\alpha - \int \bar{X}^\alpha d\mu \right)_{\alpha \in \Lambda(d)} \right\| \geq \varepsilon \right\} \\ \cap \bigcap_{\delta > 0} \left\{ (a_\alpha)_{\alpha \in \mathbb{N}^n} \in Z \mid \left| \sum_{\alpha} c_\alpha a_\alpha - f^* \right| \leq \delta \right\} = \emptyset,$$

where $(c_\alpha)_{\alpha \in \mathbb{N}^n}$ is the family of coefficients of f , i.e., $f = \sum_{\alpha} c_\alpha \bar{X}^\alpha$.

Indeed, an element $(a_\alpha)_\alpha$ of this huge intersection would give rise to a linear map $L : \mathbb{R}[\bar{X}] \rightarrow \mathbb{R} : \bar{X}^\alpha \mapsto a_\alpha$, which by Theorem 1.2 comes from a probability measure $\mu \in \mathcal{M}^1(S)$, i.e.,

$$\forall p \in \mathbb{R}[\bar{X}] : L(p) = \int p d\mu,$$

in other words,

$$(3.2) \quad \forall \alpha \in \mathbb{N}^n : a_\alpha = \int \bar{X}^\alpha d\mu.$$

Moreover, we would have $|L(f) - f^*| \leq \delta$ for all $\delta > 0$, whence

$$\int (f - f^*) d\mu = L(f - f^*) = L(f) - f^* = 0$$

so that $\mu(S^*) = 1$, and μ can be looked at as a probability measure on S^* , that is, $\mu \in \mathcal{M}^1(S^*)$. Regarding (3.2), this means that the a_α are the moments of a probability measure on S^* , which is impossible since $(a_\alpha)_{\alpha \in \Lambda(d)}$ has distance $\geq \varepsilon > 0$

to the family of moments up to order d of each probability measure on S^* . This establishes (3.1).

In fact, we can even find finitely many sets contributing to the intersection in (3.1) that have no point in common since it is an empty intersection of closed sets in a compact space. In particular, we get a tight version of (3.1) where we take the intersection only over certain finitely many $p_1, \dots, p_s \in M$ and a sufficiently small $\delta > 0$ (instead of all $p \in M$ and all $\delta > 0$). Choose $t \in \mathcal{N} \cap [d, \infty)$ such that $p_1, \dots, p_s \in M_t \subseteq \mathbb{R}[\bar{X}]_t$. Choose

$$k_0 \geq \max\{t, N_0, \dots, N_t\}$$

so large that $|L_k(f) - f^*| \leq \delta$ for all $k \geq k_0$, which is possible by Remark 3.2. Now suppose that $k \geq k_0$ and define $(a_\alpha)_{\alpha \in \mathbb{N}^n}$ by

$$a_\alpha := \begin{cases} L_k(\bar{X}^\alpha) & \text{if } |\alpha| \leq t, \\ 0 & \text{otherwise.} \end{cases}$$

Then $(a_\alpha)_\alpha \in Z$ since, for $\alpha \in \mathbb{N}^n$ with $|\alpha| \leq t$,

$$N_{|\alpha|} \pm a_\alpha = N_{|\alpha|} \pm L_k(\bar{X}^\alpha) = L_k(N_{|\alpha|} \pm \bar{X}^\alpha) \geq 0.$$

Here the inequality is implied by $N_{|\alpha|} \pm \bar{X}^\alpha \in M_{N_{|\alpha|}} \subseteq M_k$, which follows in turn from $N_{|\alpha|} \leq k_0 \leq k$. Furthermore, $a_0 = L_k(1) = 1$ and, for $b_\alpha \in \mathbb{R}$ such that $\sum_\alpha b_\alpha \bar{X}^\alpha \in \{p_1, \dots, p_s\}$,

$$\begin{aligned} \sum_\alpha b_\alpha a_\alpha &= \sum_\alpha b_\alpha L_k(\bar{X}^\alpha) \\ &= L_k\left(\sum_\alpha b_\alpha \bar{X}^\alpha\right) \in \{L_k(p_1), \dots, L_k(p_s)\} \subseteq L_k(M_t) \subseteq L_k(M_k) \subseteq [0, \infty). \end{aligned}$$

Because of $\deg f \leq t$ (recall $t \in \mathcal{N}$), we have, moreover,

$$\left| \sum_\alpha c_\alpha a_\alpha - f^* \right| = |L_k(f) - f^*| \leq \delta,$$

and the tight version of (3.1) implies the existence of $\mu \in \mathcal{M}^1(S^*)$ such that

$$\left\| \left(a_\alpha - \int \bar{X}^\alpha d\mu \right)_{\alpha \in \Lambda(d)} \right\| < \varepsilon.$$

But here we can replace a_α by $L_k(\bar{X}^\alpha)$ since $d \leq t$. $\square$

COROLLARY 3.5. *Suppose that $S^* = \{x^*\}$ is a singleton and L_k solves (P_k) nearly to optimality ($k \in \mathcal{N}$). Then*

$$\lim_{k \rightarrow \infty} (L_k(X_1), \dots, L_k(X_n)) = x^*.$$

Proof. We set $d = 1$ in the preceding theorem and note that $\mathcal{M}^1(S^*)$ contains only the Dirac measure at the point x^* . $\square$

As a nice application, we can approximate unique minimizers of polynomials on polytopes not only from below but also from above. This allows one to eventually

detect with certitude that f^* has almost been reached when successively solving the relaxations (P_k) . Note that the general assumption from section 1 is automatically satisfied under the hypothesis that S is compact and $\deg g_i \leq 1$ for all $i \in \{1, \dots, m\}$ [PD, 6.5.3].

COROLLARY 3.6. *Suppose all the g_i have degree ≤ 1 , f has a unique minimizer on the polytope S , and L_k solves (P_k) nearly to optimality ($k \in \mathcal{N}$). Then for each $k \in \mathcal{N}$,*

$$[L_k(f), f(L_k(X_1), \dots, L_k(X_n))]$$

is an interval containing f^* , and the left and right endpoints of these intervals converge to f^* for $k \rightarrow \infty$, respectively.

Proof. Taking into account Theorem 1.5, it suffices to show that

$$(f(L_k(X_1), \dots, L_k(X_n)))_{k \in \mathcal{N}}$$

is a sequence, which is bounded from below by f^* and converges to f^* . The convergence follows immediately from Corollary 3.5. To see that f^* is a lower bound, observe that

$$g_i(L_k(X_1), \dots, L_k(X_n)) = L_k(g_i) \geq 0,$$

whence $(L_k(X_1), \dots, L_k(X_n)) \in S$ for all $k \in \mathcal{N}$. $\square$

4. Duality. Throughout this section, we fix some $k \in \mathcal{N}$ ($\mathcal{N}$ has been defined in (1.7) on page 809). The kernel of any feasible solution of (P_k) is a hyperplane of the finite-dimensional vector space $\mathbb{R}[\bar{X}]_k$ with the following properties:

- (i) The convex cone M_k is contained in one of the two closed half-spaces of $\mathbb{R}[\bar{X}]_k$ defined by the hyperplane.
- (ii) The vector $1 \in \mathbb{R}[\bar{X}]_k$ lies not on the hyperplane.

Conversely, all hyperplanes fulfilling (i) and (ii) are the kernel of a feasible solution of (P_k) .

Now the following image is helpful. Suppose at the point $f \in \mathbb{R}[\bar{X}]_k$ there is a source that produces subatomic particles and sends them on the way in the direction of the vector $-1 \in \mathbb{R}[\bar{X}]_k$. Every particle has constant speed and covers a distance of 1 per time unit (this will correspond to the constraint $L(1) = 1$ in (P_k)). We are interested in the age D_k^* of such a particle when it leaves the cone M_k (suppose that $D_k^* \in [0, \infty)$ in order to make this interpretation work). The only measuring instrument we have is a detector that catches the particles with a big metal plate which we view as a hyperplane. When the instrument detects a particle, it can determine its age. Since the cone M_k is a material body, the metal plate can touch it but not interpenetrate it; i.e., it satisfies constraint (i). If we brought the metal plate into a position where it contains the vector $1 \in \mathbb{R}[\bar{X}]_k$, then it would detect particles of all ages. So we will obey constraint (ii) while measuring. Thus P_k^* can be seen as the age we can measure by moving around the metal plate, and D_k^* as the actual age of the particles when they leave the cone.

This visualization nicely illustrates “weak duality” $P_k^* \geq D_k^*$, since a particle cannot be detected before it leaves the cone. It also suggests that “strong duality” $P_k^* = D_k^*$ should hold in many cases because it should often be possible to catch a particle just when it leaves the cone. However, one can imagine that there might be problems when the particle moves along the boundary of M_k and M_k is not closed. In this section, we will discuss these matters in detail.

Notation 4.1. For $s \in \mathbb{N}$, we denote by $\mathbb{R}^{s \times s}$ the vector space of real $s \times s$ matrices. Defining the scalar product $\langle A, B \rangle$ of two such matrices A and B by

$$(4.1) \quad \langle A, B \rangle := \sum_{i,j=1}^s A_{ij} B_{ij}$$

makes $\mathbb{R}^{s \times s}$ a Euclidean vector space. We write $S\mathbb{R}^{s \times s}$ for its subspace of *symmetric* matrices. By $S\mathbb{R}_+^{s \times s}$, we mean the set of symmetric *positive semidefinite* $s \times s$ matrices. It is a closed convex cone of $S\mathbb{R}^{s \times s}$.

Often, we will index rows and columns of matrices by elements of $\Lambda(d)$ for some $d \in \mathbb{N}$ (recall Notation 3.3). We denote the corresponding objects by $\mathbb{R}^{\Lambda(d) \times \Lambda(d)}$, $S\mathbb{R}^{\Lambda(d) \times \Lambda(d)}$, and $S\mathbb{R}_+^{\Lambda(d) \times \Lambda(d)}$. We write $A(\alpha, \beta)$ (instead of $A_{\alpha\beta}$) for the entry of a matrix $A \in \mathbb{R}^{\Lambda(d) \times \Lambda(d)}$ at line $\alpha \in \Lambda(d)$ and column $\beta \in \Lambda(d)$. Accordingly, we consider the vector space $\mathbb{R}^{\Lambda(d)}$ consisting of vectors v with entries $v(\alpha)$, $\alpha \in \Lambda(d)$.

Finally, we extend the definition of $\langle A, B \rangle$ by (4.1) to matrices A and B with polynomial entries.

LEMMA 4.2. *If $A \in S\mathbb{R}_+^{s \times s}$, then $A = \sum_{i=0}^s v_i v_i^T$ for some vectors $v_1, \dots, v_s \in \mathbb{R}^s$.*

Proof. As A is symmetric, there exists an (orthogonal) matrix $B \in \mathbb{R}^{s \times s}$ and a real diagonal matrix D such that $A = BDB^T$. The entries of D are the eigenvalues of A , which are all nonnegative by hypothesis. So there exists a real diagonal matrix $\sqrt{D}$ such that $(\sqrt{D})^2 = D$. Therefore we get $A = (B\sqrt{D})(B\sqrt{D})^T = CC^T$, where $C := B\sqrt{D}$. Choosing $v_1, \dots, v_s$ as the columns of C yields the desired expression of A . $\square$

The following lemma relates positive semidefinite matrices to sums of squares of polynomials. More about this well-known connection can be found in [CLR].

LEMMA 4.3. *For all $d \in \mathbb{N}$,*

$$\begin{aligned} \sum \mathbb{R}[\bar{X}]_d^2 &= \left\{ \langle (\bar{X}^{\beta+\gamma})_{(\beta,\gamma) \in \Lambda(d) \times \Lambda(d)}, G \rangle \mid G \in S\mathbb{R}_+^{\Lambda(d) \times \Lambda(d)} \right\} \\ &= \{q_1^2 + \dots + q_s^2 \mid q_1, \dots, q_s \in \mathbb{R}[\bar{X}]_d\}, \end{aligned}$$

where $s := \dim \mathbb{R}[\bar{X}]_d = \#\Lambda(d)$ denotes the dimension of the $\mathbb{R}$ -vector space $\mathbb{R}[\bar{X}]_d$.

Proof. To see that the first set is contained in the second one, note that the second one is closed under addition. So it suffices to show that it contains q^2 for each $q \in \mathbb{R}[\bar{X}]_d$. The polynomial q defines a vector $v \in \mathbb{R}^{\Lambda(d)}$ by setting $v(\alpha)$ to be the coefficient of $\bar{X}^\alpha$ in q for $\alpha \in \Lambda(d)$. Then $G := vv^T \in S\mathbb{R}^{\Lambda(d) \times \Lambda(d)}$ is positive semidefinite and

$$\begin{aligned} &\langle (\bar{X}^{\beta+\gamma})_{(\beta,\gamma) \in \Lambda(d) \times \Lambda(d)}, G \rangle \\ &= \sum_{\beta, \gamma \in \Lambda(d)} \bar{X}^{\beta+\gamma} G(\beta, \gamma) \\ &= \sum_{\beta, \gamma \in \Lambda(d)} \bar{X}^{\beta+\gamma} v(\beta) v(\gamma) = \left(\sum_{\beta \in \Lambda(d)} v(\beta) \bar{X}^\beta \right) \left(\sum_{\gamma \in \Lambda(d)} v(\gamma) \bar{X}^\gamma \right) = q^2. \end{aligned}$$

To show that the second set is contained in the last one, first consider the case that $G = vv^T$ for some vector $v \in \mathbb{R}^{\Lambda(k)}$. If we now set $q := \sum_{\alpha \in \Lambda(k)} v(\alpha) \bar{X}^\alpha$, then the above equation holds again. In the general case, use Lemma 4.2 and argue similarly. Finally, it is trivial that the last set is contained in the first. $\square$

Note that $\mathbb{R}[\bar{X}]_k$ carries a natural topology, which is induced by any norm on it. The proof of the next result is essentially contained in [PS, 2.6]; see also [Mar]. We repeat the proof without using results from algebraic geometry. The next remark should prevent misunderstandings concerning the notion of “interior” appearing in the statement.

Remark 4.4. The interior of S is understood in the usual topological sense, i.e., the set of all interior points of S where an interior point is the center of a nonempty open (full-dimensional) ball contained in S . Note that

$$S' := \{x \in \mathbb{R}^n \mid g_1(x) > 0, \dots, g_m(x) > 0\}$$

is contained in this interior. Examples where this inclusion is proper are trivial to find. Nevertheless, $S' = \emptyset$ implies that the interior of S is empty, too, provided that none of the g_i are the zero polynomial. Indeed, $S' = \emptyset$ would imply that the polynomial $g_1 \cdots g_m$ vanishes on S . If the interior of S were nonempty, then $g_1 \cdots g_m = 0$ would follow.

THEOREM 4.5 (Powers and Scheiderer). *If S has nonempty interior, then M_k is closed in $\mathbb{R}[\bar{X}]_k$.*

Proof. By (1.6) and Lemma 4.3, the convex cone M_k is the image under the continuous map

$$\varphi : \begin{cases} \mathbb{R}[\bar{X}]_{d_0}^{\Lambda(d_0)} \times \cdots \times \mathbb{R}[\bar{X}]_{d_m}^{\Lambda(d_m)} \rightarrow \mathbb{R}[\bar{X}]_k \\ ((q_{0\alpha})_{\alpha \in \Lambda(d_0)}, \dots, (q_{m\alpha})_{\alpha \in \Lambda(d_m)}) \mapsto \sum_{i=0}^m \sum_{\alpha \in \Lambda(d_i)} q_{i\alpha}^2 g_i. \end{cases}$$

This map is quadratically homogeneous, i.e., $\varphi(\lambda z) = \lambda^2 \varphi(z)$ for all $\lambda \in \mathbb{R}$ and z in the domain of φ . Let us assume without loss of generality that none of the g_i are the zero polynomial. Then there is some $x \in S$ with $g_i(x) > 0$ for all $i \in \{1, \dots, m\}$; cf. the preceding remark. All g_i are positive on a neighborhood U of x . For z in the domain of φ , we have that $\varphi(z) \geq 0$ on U , but $\varphi(z)$ cannot vanish on U unless $z = 0$. In other words, the map φ is injective. Now equip the $\mathbb{R}$ -vector space on which φ is defined with an arbitrary norm. Then the image V of the unit sphere with respect to this norm is a compact set not containing the origin. From the homogeneity of φ and the fact that M_k is a cone, it is easy to see that M_k is the set of nonnegative multiples of elements of V . Finally, it follows from $0 \notin V$ that M_k is closed. Indeed, any sequence in M_k is of the form $(\lambda_i v_i)_{i \in \mathbb{N}}$, where $\lambda_i \in [0, \infty)$ and $0 \neq v_i \in V$ for all $i \in \mathbb{N}$. Suppose that $p = \lim_{i \rightarrow \infty} \lambda_i v_i$. We have to show that $p \in M_k$. Since V is compact, the sequence $(v_i)_{i \in \mathbb{N}}$ has a convergent subsequence. Without loss of generality we may assume that $(v_i)_{i \in \mathbb{N}}$ itself is convergent, say $v := \lim_{i \rightarrow \infty} v_i$. Then $v \in V$ and a fortiori $v \neq 0$. Now the limit

$$\lim_{i \rightarrow \infty} \lambda_i = \lim_{i \rightarrow \infty} \frac{\|\lambda_i v_i\|}{\|v_i\|} = \frac{\|\lim_{i \rightarrow \infty} \lambda_i v_i\|}{\|\lim_{i \rightarrow \infty} v_i\|} = \frac{\|p\|}{\|v\|}$$

exists, such that

$$\lim_{i \rightarrow \infty} \lambda_i v_i = \left(\lim_{i \rightarrow \infty} \lambda_i \right) \left(\lim_{i \rightarrow \infty} v_i \right) = \frac{\|p\|}{\|v\|} v \in [0, \infty) V = M_k. \quad \square$$

Following Marshall [Mar], we draw the following corollary. Note that a feasible solution L of (P_k) has to obey the constraint $L(1) = 1$. So the corollary is not really

just an instance of the well-known fact that the bidual of a closed convex cone is the cone itself.

COROLLARY 4.6 (Marshall). *If S has nonempty interior, then M_k is the intersection of all half-spaces $\{p \in \mathbb{R}[\bar{X}]_k \mid L(p) \geq 0\}$, where L ranges over all feasible solutions of (P_k) .*

Proof. Consider an arbitrary $p \in \mathbb{R}[\bar{X}]_k \setminus M_k$. We have to show that there is a feasible solution of (P_k) such that $L(p) < 0$.

Fix some scalar product on the $\mathbb{R}$ -vector space $\mathbb{R}[\bar{X}]_k$. Since M_k is closed by the preceding theorem, we are sure that there exists $q \in M_k$ with minimal distance to p . Choose a linear map $L_0 : \mathbb{R}[\bar{X}]_k \rightarrow \mathbb{R}$ whose kernel is the hyperplane perpendicular to $p - q$ satisfying $L_0(p - q) < 0$. Note that $L_0(q) = 0$ either because of $q = 0$ or because q is the orthogonal projection of p on the line $\mathbb{R}q$. So we have $L_0(p) = L_0(p - q) + L_0(q) = L_0(p - q) < 0$.

It is easy to see that $L_0(M_k) \subseteq [0, \infty)$. Indeed, if there were a $q' \in M_k$ such that $L_0(q') < 0$, then all points q'' on the line segment from q' to q in a neighborhood of q would be even closer to p than it is q . At the same time, the convexity of M_k would imply $q'' \in M_k$, contradicting the choice of q .

Of course, we have $L_0(1) \geq 0$ since $1 \in M_k$. If $L_0(1)$ happens to be positive, then we can scale L_0 with a positive factor in order to obtain an L with the desired properties. If $L_0(1) = 0$, then we choose any point $x \in S$ and set $L := \varepsilon_x + \lambda L_0$ for sufficiently large $\lambda \in [0, \infty)$, where $\varepsilon_x : \mathbb{R}[\bar{X}]_k \rightarrow \mathbb{R}$ is evaluation at the point x . $\square$

We can now prove strong duality. In the case that (D_k) has a feasible solution, this has been first proved by Lasserre using the duality theory of semidefinite programming [Las].

COROLLARY 4.7 (Lasserre). *Suppose that S has nonempty interior. Then $P_k^* = D_k^*$. In addition, if (D_k) has a feasible solution, then it also has an optimal solution.*

Proof. Suppose that $a \in \mathbb{R}$ is such that $f - a \notin M_k$. Then Corollary 4.6 ensures the existence of a feasible solution L of (P_k) such that $L(f - a) < 0$, whence $D_k^* \leq P_k^* \leq L(f) < a$. This implies all our claims. $\square$

The next example shows that (P_k^*) might fail to have an optimal solution, even if S has nonempty interior.

Example 4.8. We consider an example in $n := 2$ variables that we denote by X and Y instead of X_1 and X_2 . To prepare our example, consider a linear map $L : \mathbb{R}[X, Y]_2 \rightarrow \mathbb{R}$ satisfying $L(1) = 1$. We claim that the condition

$$(4.2) \quad L((aX + bY + c)^2) \geq 0 \quad \text{for all } a, b, c \in \mathbb{R}$$

is equivalent to the three simultaneous inequalities

$$(4.3) \quad (L(XY) - L(X)L(Y))^2 \leq \underbrace{(L(X^2) - L(X)^2)}_{\geq 0} \underbrace{(L(Y^2) - L(Y)^2)}_{\geq 0}.$$

Indeed, using a scaling argument, we can split condition (4.2) into the two conditions

$$(4.4) \quad L((X + bY + c)^2) \geq 0 \quad \text{for all } b, c \in \mathbb{R}$$

and

$$(4.5) \quad L((Y + c)^2) \geq 0 \quad \text{for all } c \in \mathbb{R}.$$

We can rewrite (4.5) as

$$c^2 + 2L(Y)c + L(Y^2) \geq 0 \quad \text{for all } c \in \mathbb{R},$$

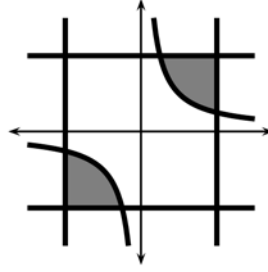


FIG. 4.1.

which is equivalent to

$$(4.6) \quad (L(Y))^2 \leq L(Y^2).$$

In a similar way, we see that (4.4) is equivalent to

$$(L(X + bY))^2 \leq L((X + bY)^2),$$

i.e.,

$$(L(X))^2 + 2bL(X)L(Y) + b^2(L(Y))^2 \leq L(X^2) + 2bL(XY) + b^2L(Y^2)$$

for all $b \in \mathbb{R}$. Now take all the terms to the right-hand side and look at them as a (perhaps degenerate) parabola in b . This parabola opens up by (4.6). Writing down the condition that this parabola has at most one zero (taking care of degeneration) and combining this with (4.6) shows the equivalence of (4.2) and (4.3).

Having finished the preparation, consider now the set S defined by $m := 5$ polynomials $g_1 = XY - 1$, $g_2 = 2 + X$, $g_3 = 2 - X$, $g_4 = 2 + Y$, and $g_5 = 2 - Y$. It is the filled region in Figure 4.1 and clearly has interior points. Note that our general assumption made in section 1 is satisfied since $4(4 - X^2) = (g_2 + g_3)g_2g_3 = g_2^2g_3 + g_3^2g_2 \in M$, analogously $4(4 - Y^2) \in M$, and therefore $8 - (X^2 + Y^2) \in M$. We look at the second primal and dual relaxation, (P_2) and (D_2) . We observe that

$$M_2 = \sum \mathbb{R}[X, Y]_1^2 + [0, \infty)(XY - 1) + [0, \infty)(2 + X) + [0, \infty)(2 - X) \\ + [0, \infty)(2 + Y) + [0, \infty)(2 - Y).$$

Therefore, the feasible solutions of (P_2) are the linear maps $L : \mathbb{R}[X, Y]_2 \rightarrow \mathbb{R}$ satisfying $L(1) = 1$, condition (4.3), $L(XY) \geq 1$, and $(L(X), L(Y)) \in [-2, 2] \times [-2, 2]$. Suppose that we want to minimize $f := X^2$. For any feasible solution L of (P_2) , we have that $L(f) > 0$. Otherwise, as in dominoes, all the subexpressions of (4.3) involving X would have to vanish, contradicting $L(XY) \geq 1$. On the other hand, for all $\varepsilon > 0$, the linear map $L : \mathbb{R}[X, Y]_2 \rightarrow \mathbb{R}$ defined by

$$L(X^2) = \varepsilon^2, \quad L(Y^2) = \frac{1}{\varepsilon^2}, \quad L(XY) = 1, \quad L(X) = L(Y) = 0, \quad \text{and} \quad L(1) = 1$$

is feasible for (P_k) . This proves that $P_2^* = 0$, and (P_2) has no optimal solution.

We conclude the section with an example where a big duality gap occurs. By Corollary 4.7, S cannot have interior points in such an example.

Example 4.9. Take again $n := 2$ variables X and Y . The set $S = [-1, 1] \times \{0\}$ defined by the $m := 3$ polynomials $g_1 := -Y^2$, $g_2 := 1 + X$, and $g_3 := 1 - X$ has empty interior. The general assumption from section 1 is trivially satisfied since

$$2(1 - (X^2 + Y^2)) = (g_2 + g_3)g_2g_3 + 2g_1 = g_2^2g_3 + g_3^2g_2 + (\sqrt{2})^2g_1 \in M.$$

This time, we want to minimize $f := XY$ on this set. Once more, we look at the second relaxations. It is easy to see that

$$M_2 = \sum \mathbb{R}[X, Y]_1^2 - [0, \infty)Y^2 + [0, \infty)(1 + X) + [0, \infty)(1 - X).$$

Every feasible solution L of (P_2) satisfies $L(\varepsilon X^2 + f) \geq 0$ for all $\varepsilon > 0$ since

$$\varepsilon X^2 + f = \left(\sqrt{\varepsilon}X + \frac{1}{2\sqrt{\varepsilon}}Y \right)^2 - \frac{1}{4\varepsilon}Y^2 \in M_2.$$

But then $\varepsilon L(X^2) + L(f) \geq 0$ for all $\varepsilon > 0$, i.e., $L(f) \geq 0$. This shows $0 \leq P_2^* \leq f^* = 0$, whence $P_2^* = f^* = 0$.

On the other hand, $D_2^* = -\infty < 0 = P_2^*$ because there is no $a \in \mathbb{R}$ such that $f - a \in M_2$. If there were $s \in \mathbb{N}$, $a, b_i, c_i, d_i, e_i \in \mathbb{R}$ such that

$$f - a = \sum_{i=1}^s (b_i X + c_i Y + d_i)^2 - e_1 Y^2 + e_2(1 + X) + e_3(1 - X),$$

then comparing the coefficients of X^2 on both sides would yield $\sum_{i=1}^s b_i^2 = 0$. Hence the coefficient of XY on the right-hand side would be zero.

5. Formulation as a semidefinite program. We fix again some $k \in \mathcal{N}$ (as defined in (1.7) on page 809). In this section, we show how to solve the problems (P_k) and (D_k) using semidefinite programming. This translation of (P_k) and (D_k) into semidefinite programs has become a quite common technique now (see, e.g., [Las], [Mar], [PPP]). Though this short section therefore does not contain new ideas, we include it in order to make this article a complete introduction to Lasserre's method. Also, other articles use quite different notations and a technically different setup. (For example, Lasserre [Las] considers the relaxations (P_k) and (D_k) only for even k . This is in our opinion not natural, at least if one of the g_i has odd degree.) Semidefinite programs are generalizations of linear programs, which can be solved efficiently (see, for example, [Tod]). Before we translate (P_k) and (D_k) into semidefinite programs, we will carry out three modifications, which are harmless. First, we will pass over from f to $f - f(0)$. This amounts to just adding the same constant to the goal functions of both problems. Second, we will negate both goal functions and at the same time exchange minimization and maximization. Third, we will exchange primal and dual. So we will forget about (P_k) and (D_k) and instead look at the following pair of modified problems (P_k^{mod}) and (D_k^{mod}) :

$$\begin{aligned} (P_k^{\text{mod}}) \quad & \text{minimize} \quad a && \text{subject to} \quad a \in \mathbb{R} \text{ and} \\ & && (f - f(0)) + a \in M_k, \\ (D_k^{\text{mod}}) \quad & \text{maximize} \quad -L(f - f(0)) && \text{subject to} \quad L : \mathbb{R}[\bar{X}]_k \rightarrow \mathbb{R} \text{ is linear,} \\ & && L(1) = 1, \text{ and} \\ & && L(M_k) \subseteq [0, \infty). \end{aligned}$$

Define $d_0, \dots, d_m \in \mathbb{N}$ as in (1.8) and recall Notations 3.3 and 4.1. The key lemma is the following.

LEMMA 5.1. *Suppose $L : \mathbb{R}[\bar{X}]_k \rightarrow \mathbb{R}$ is a linear map. Then $L(M_k) \subseteq [0, \infty)$ if and only if the $m + 1$ matrices*

$$(L(\bar{X}^{\beta+\gamma} g_i))_{(\beta, \gamma) \in \Lambda(d_i) \times \Lambda(d_i)} \quad (i \in \{0, \dots, m\})$$

are positive semidefinite. Moreover,

$$M_k = \left\{ \sum_{i=0}^m \langle (\bar{X}^{\beta+\gamma} g_i)_{(\beta,\gamma) \in \Lambda(d_i) \times \Lambda(d_i)}, G_i \rangle \mid G_0, \dots, G_m \in S\mathbb{R}_+^{\Lambda(d_i) \times \Lambda(d_i)} \right\}.$$

Proof. The first part follows from the observation that the $m+1$ matrices under consideration represent the bilinear forms

$$\mathbb{R}[\bar{X}]_{d_i} \times \mathbb{R}[\bar{X}]_{d_i} \rightarrow \mathbb{R} : (p, q) \mapsto L(pqg_i) \quad (i \in \{0, \dots, m\})$$

with respect to the bases $(\bar{X}^\alpha)_{\alpha \in \Lambda(d_i)}$ of $\mathbb{R}[\bar{X}]_{d_i}$. The positive semidefiniteness of such a matrix expresses the positive semidefiniteness of the corresponding bilinear form, that is, $L(p^2g_i) \geq 0$ for all $p \in \mathbb{R}[\bar{X}]_{d_i}$. The second part follows easily from (1.6) and Lemma 4.3. $\square$

Using this lemma, we can reformulate (P_k^{mod}) and (D_k^{mod}) using positive semidefinite matrices:

$$\begin{aligned} (P_k^{\text{I}}) \quad & \text{minimize} \quad a \\ & \text{subject to} \quad a \in \mathbb{R}, \quad G_0, \dots, G_m \in S\mathbb{R}_+^{\Lambda(d_i) \times \Lambda(d_i)}, \text{ and} \\ & \quad \sum_{i=0}^m \langle (\bar{X}^{\beta+\gamma} g_i)_{(\beta,\gamma) \in \Lambda(d_i) \times \Lambda(d_i)}, G_i \rangle = f - f(0) + a, \\ (D_k^{\text{I}}) \quad & \text{maximize} \quad -L(f - f(0)) \\ & \text{subject to} \quad L : \mathbb{R}[\bar{X}]_k \rightarrow \mathbb{R} \text{ is linear, } L(1) = 1, \text{ and} \\ & \quad (L(\bar{X}^{\beta+\gamma} g_i))_{(\beta,\gamma) \in \Lambda(d_i) \times \Lambda(d_i)} \text{ is positive semidefinite, } i = 0, \dots, m. \end{aligned}$$

The matrix $(L(\bar{X}^{\beta+\gamma}))_{(\beta,\gamma) \in \Lambda(d_0, d_0)}$ appearing in (D_k^{I}) (recall that $g_0 = 1$) is often called a *moment matrix* of L . This is because, if L is integration with respect to a measure, then its entries are moments of this measure. Such matrices seem to be very interesting objects (see [CF], [Lau], and the references therein). The same is true for the so-called *localizing matrices* $(L(\bar{X}^{\beta+\gamma} g_i))_{(\beta,\gamma) \in \Lambda(d_i) \times \Lambda(d_i)}$, $i = 1, \dots, m$.

Now for $i \in \{0, \dots, m\}$ and $\alpha \in \Lambda(k)$, define a matrix $A_{\alpha i} \in S\mathbb{R}^{\Lambda(d_i) \times \Lambda(d_i)}$ by requiring

$$(5.1) \quad \bar{X}^{\beta+\gamma} g_i = \sum_{\alpha \in \Lambda(k)} A_{\alpha i}(\beta, \gamma) \bar{X}^\alpha,$$

that is, $A_{\alpha i}(\beta, \gamma)$ is the coefficient of $\bar{X}^\alpha$ in $\bar{X}^{\beta+\gamma} g_i$. Also, we define b_α to be the coefficient of $\bar{X}^\alpha$ in f for each $0 \neq \alpha \in \Lambda(k)$, i.e.,

$$(5.2) \quad f - f(0) = \sum_{\alpha \in \Lambda(k) \setminus \{0\}} b_\alpha \bar{X}^\alpha.$$

We plug (5.1) and (5.2) into (P_k^{I}) and get

$$\begin{aligned} (P_k^{\text{II}}) \quad & \text{minimize} \quad a \\ & \text{subject to} \quad a \in \mathbb{R}, \quad G_0, \dots, G_m \in S\mathbb{R}_+^{\Lambda(d_i) \times \Lambda(d_i)}, \text{ and} \\ & \quad \sum_{\alpha \in \Lambda(k)} \bar{X}^\alpha \sum_{i=0}^m \langle A_{\alpha i}, G_i \rangle = \sum_{\alpha \in \Lambda(k) \setminus \{0\}} b_\alpha \bar{X}^\alpha + a, \end{aligned}$$

$$\begin{aligned}
(D_k^{\text{II}}) \quad & \text{maximize} \quad \sum_{\alpha \in \Lambda(k) \setminus \{0\}} -b_\alpha L(\bar{X}^\alpha) \\
& \text{subject to} \quad L : \mathbb{R}[\bar{X}]_k \rightarrow \mathbb{R} \text{ is linear, } L(1) = 1, \text{ and} \\
& \quad \sum_{\alpha \in \Lambda(k)} L(\bar{X}^\alpha) A_{\alpha i} \text{ is positive semidefinite, } i = 0, \dots, m.
\end{aligned}$$

Finally, we rewrite the polynomial equality constraint in (P_k^{II}) by splitting it up, coefficient by coefficient, and incorporating the equation for the constant coefficient into the goal function. Concerning the dual problem (D_k^{II}) , we exploit that a linear map $L : \mathbb{R}[\bar{X}]_k \rightarrow \mathbb{R}$ with $L(1) = 1$ can be identified with its values $y_\alpha := -L(\bar{X}^\alpha)$, $0 \neq \alpha \in \Lambda(k)$:

$$\begin{aligned}
(P_k^{\text{sdp}}) \quad & \text{minimize} \quad \sum_{i=0}^m \langle A_{0i}, G_i \rangle \\
& \text{subject to} \quad G_0, \dots, G_m \in S\mathbb{R}_+^{\Lambda(d_i) \times \Lambda(d_i)} \text{ and} \\
& \quad \sum_{i=0}^m \langle A_{\alpha i}, G_i \rangle = b_\alpha, \quad 0 \neq \alpha \in \Lambda(k), \\
(D_k^{\text{sdp}}) \quad & \text{maximize} \quad \sum_{\alpha \in \Lambda(k) \setminus \{0\}} b_\alpha y_\alpha \\
& \text{subject to} \quad y_\alpha \in \mathbb{R}, \quad 0 \neq \alpha \in \Lambda(k), \text{ and} \\
& \quad A_{0i} - \sum_{\alpha \in \Lambda(k) \setminus \{0\}} y_\alpha A_{\alpha i} \text{ is positive semidefinite, } i = 0, \dots, m.
\end{aligned}$$

This is in fact a primal-dual pair of semidefinite programs with matrices in block diagonal structure (see [Tod]). For each $\alpha \in \Lambda(k)$, one could define A_α to be the matrix consisting of the $m+1$ diagonal blocks $A_{\alpha i}$, $i = 0, \dots, m$, and G to be a positive semidefinite matrix variable of the same size. Then (P_k^{sdp}) could be written as the minimization of $\langle A_0, G \rangle$ subject to $\langle A_\alpha, G \rangle = b_\alpha$ for $0 \neq \alpha \in \Lambda(k)$. Since semidefinite programming solvers usually take advantage of such a block diagonal structure, we stick to the above formulation.

Acknowledgment. The author would like to thank Chip Delzell for his numerous useful comments concerning the exposition of the present work.

REFERENCES

- [BW] R. BERR AND T. WÖRMANN, *Positive polynomials on compact sets*, Manuscripta Math., 104 (2001), pp. 135–143.
- [CF] R. E. CURTO AND L. FIALKOW, *The truncated complex K -moment problem*, Trans. Amer. Math. Soc., 352 (2000), pp. 2825–2855.
- [CLR] M. D. CHOI, T. Y. LAM, AND B. REZNICK, *Sums of squares of real polynomials*, in *K-theory and Algebraic Geometry: Connections with Quadratic Forms and Division Algebras*, Proc. Sympos. Pure Math. 58, Part 2, AMS, Providence, RI, 1995, pp. 103–126.
- [HL] D. HENRION AND J. B. LASSERRE, *GloptiPoly: Global optimization over polynomials with Matlab and SeDuMi*, ACM Trans. Math. Software, 29 (2003), pp. 165–194.
- [Jac] T. JACOBI, *A representation theorem for certain partially ordered commutative rings*, Math. Z., 237 (2001), pp. 259–273.
- [JP] T. JACOBI AND A. PRESTEL, *Distinguished representations of strictly positive polynomials*, J. Reine Angew. Math., 532 (2001), pp. 223–235.

- [Las] J. B. LASSERRE, *Global optimization with polynomials and the problem of moments*, SIAM J. Optim., 11 (2001), pp. 796–817.
- [Lau] M. LAURENT, *Revisiting two theorems of Curto and Fialkow on moment matrices*, Proc. Amer. Math. Soc., to appear.
- [Mar] M. MARSHALL, *Optimization of polynomial functions*, Canad. Math. Bull., 46 (2003), pp. 575–587.
- [PD] A. PRESTEL AND C. DELZELL, *Positive Polynomials*, Springer Monogr. Math., Springer-Verlag, Berlin, 2001.
- [Pól] G. PÓLYA, *Über positive Darstellung von Polynomen*, Vierteljahresschrift der Naturforschenden Gesellschaft in Zürich, 73 (1928), pp. 141–145; reprinted in Collected Papers, Vol. 2, MIT Press, Cambridge, MA, 1974, pp. 309–313.
- [PPP] S. PRAJNA, A. PAPACHRISTODOULOU, P. A. PARRILO, AND P. SEILER, *SOSTOOLS: Sum of squares optimization toolbox for MATLAB*, in Proceedings of the 41st IEEE Conference on Decision and Control, Las Vegas, 2002; also available from <http://www.cds.caltech.edu/sostools> and <http://www.mit.edu/~parrilo/sostools>.
- [PR] V. POWERS AND B. REZNICK, *A new bound for Pólya's theorem with applications to polynomials positive on polyhedra*, J. Pure Appl. Algebra, 164 (2001), pp. 221–229.
- [PS] V. POWERS AND C. SCHEIDERER, *The moment problem for non-compact semialgebraic sets*, Adv. Geom., 1 (2001), pp. 71–88.
- [Put] M. PUTINAR, *Positive polynomials on compact semi-algebraic sets*, Indiana Univ. Math. J., 42 (1993), pp. 969–984.
- [Rud] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill Series in Higher Mathematics, McGraw-Hill, New York, 1966.
- [S1] M. SCHWEIGHOFER, *An algorithmic approach to Schmüdgen's Positivstellensatz*, J. Pure Appl. Algebra, 166 (2002), pp. 307–319.
- [S2] M. SCHWEIGHOFER, *On the complexity of Schmüdgen's Positivstellensatz*, J. Complexity, 20 (2004), pp. 529–543.
- [Sch] K. SCHMÜDGEN, *The K -moment problem for compact semi-algebraic sets*, Math. Ann., 289 (1991), pp. 203–206.
- [Ste] G. STENGLE, *Complexity estimates for the Schmüdgen Positivstellensatz*, J. Complexity, 12 (1996), pp. 167–174.
- [Tod] M. TODD, *Semidefinite optimization*, Acta Numer., 10 (2001), pp. 515–560.

OPTIMALITY CONDITIONS FOR TRUST-REGION SUBPROBLEMS INVOLVING A CONIC MODEL*

QIN NI†

Abstract. The problem of minimizing a conic function subject to an ellipsoidal constraint is discussed in this paper. The feasible set of this problem is divided into three different cases. Necessary and sufficient optimality conditions are obtained for different cases where there is a characterization of optimality for the basic case. These conditions provide a sound basis for the further development of efficient methods for trust-region subproblems of a conic model.

Key words. conic model, trust regions, optimality conditions

AMS subject classifications. 49K10, 90C30

DOI. 10.1137/S1052623402418991

1. Introduction. In 1980, Davidon [2] proposed a class of methods for solving the unconstrained minimization problem based on a local collinear scaling of variables and a conic model. The conic model is a generalization of the quadratic model and can take into account more information from previous iterations. Some methods based on the conic model have been developed for nonlinear programming (see [1], [3], [9], [11]).

A typical conic model is as follows:

$$(1.1) \quad \psi(s) = \frac{g^T s}{1 - a^T s} + \frac{s^T A s}{2(1 - a^T s)^2},$$

where $a, g \in R^n$, $A \in R^{n \times n}$, and $s \in R^n$. The conic function ψ generally represents a local model to the objective function defined by interpolatory data at an iterate. In most cases, a linear search is used to develop some methods for the conic model. However, theoretical and numerical results show that the methods with linear search are not very efficient for the conic model. Powerful trust-region algorithms have been developed during the past 20 years and have attracted the attention of more and more researchers (see [6], [7], [13]). It is believed that combining trust-region techniques with a conic model would be appealing and the unboundedness of the conic function would be efficiently controlled by trust regions.

The trust-region subproblem is defined as

$$(1.2) \quad \min \psi(s)$$

$$(1.3) \quad \text{s.t. } s \in B,$$

where $\psi(s)$ is given in (1.1), $B = \{s \in R^n : \|s\| \leq \Delta\}$, Δ is a positive parameter that is called a trust-region radius, and $\|\cdot\|$ is the Euclidean norm for R^n . This problem was first proposed by Di and Sun in [3]. They obtained some optimality conditions for

*Received by the editors December 2, 2002; accepted for publication (in revised form) August 17, 2004; published electronically April 22, 2005. This work was done when the author was visiting the Academy of Mathematics and System Science in the Chinese Academy of Sciences and was supported by the National Natural Science Foundation of China.

<http://www.siam.org/journals/siopt/15-3/41899.html>

†Nanjing University of Aeronautics and Astronautics, Nanjing 210016, People's Republic of China (njqin6@jlonline.com).

(1.2)–(1.3) when A is symmetric and positive semidefinite and $1 - a^T s > 0$. However, the problem (1.2)–(1.3) is not well defined, since $\psi(s)$ may be undefined on B . In addition, if this conic function is used in the derivative-free optimization, A is generally not positive semidefinite. Hence, it is desirable that the problem (1.2)–(1.3) can be modified, and the modified problem can be solved for any symmetric matrix A .

In this paper, a new feasible set is added to (1.2)–(1.3) such that the trust-region subproblem of the conic model is well defined and is divided into three different cases. For any symmetric matrix A , any trust-region radius Δ , and any vectors a and g , there always exists a finite solution for the new trust-region subproblems. Necessary and sufficient optimality conditions are obtained for all different cases. These conditions provide a sound basis for the further development of efficient methods for trust-region subproblems involving a conic model.

The rest of this paper is organized as follows. In section 2, a new feasible set is proposed, and some trust-region subproblems are established. Optimality conditions for the problem of minimizing a conic function are discussed in section 3. A characterization of optimality is given for the basic trust-region subproblem in section 4, and optimality conditions for other trust-region subproblems are derived in section 5.

2. Trust-region subproblems with new feasible sets. For some conic models, the feasible set of (1.1) is defined as $\{s \in R^n, 1 - a^T s > 0\}$ or $\{s \in R^n, 1 - a^T s \neq 0\}$ (see [2] and [3]). There may not be a finite solution for (1.2)–(1.3) on these feasible sets because $\psi(s)$ may not be bounded on B and on these feasible sets. In order to avoid this situation, we define a new feasible set of (1.1) as

$$S = \{s \in R^n, |1 - a^T s| \geq \epsilon_0\},$$

where $\epsilon_0 \in (0, 1)$. Hence, combining the new set with (1.2), we obtain the following new trust-region subproblem:

$$(2.1) \quad \min \psi(s)$$

$$(2.2) \quad \text{s.t. } s \in \Omega,$$

where $\Omega = B \cap S$. A solution of (2.1)–(2.2) always exists since $\psi(s)$ is bounded on Ω .

Let

$$(2.3) \quad S_1 = \{s \in R^n, a^T s \leq 1 - \epsilon_0\}, \quad S_2 = \{s \in R^n, a^T s \geq 1 + \epsilon_0\};$$

then $\Omega = (B \cap S_1) \cup (B \cap S_2)$, since $S = S_1 \cup S_2$. There are three different expressions for Ω described by the following lemma.

LEMMA 2.1. Assume that $\epsilon_0 \in (0, 1)$, $a \in R^n$, $\Delta > 0$. Then

$$\Omega = \begin{cases} B & \text{if } 1 - \epsilon_0 \geq \Delta \|a\|, \\ B \cap S_1 & \text{if } |1 - \Delta \|a\|| < \epsilon_0, \\ (B \cap S_1) \cup (B \cap S_2) & \text{if } \Delta \|a\| \geq 1 + \epsilon_0. \end{cases}$$

Proof. Suppose that $1 - \epsilon_0 \geq \Delta \|a\|$. If $s \in B$, i.e., $\|s\| \leq \Delta$, then $|1 - a^T s| \geq 1 - \|a\| \|s\| \geq 1 - \Delta \|a\| \geq \epsilon_0$, i.e., $s \in S$. Hence, $B \subset S$ and $\Omega = B \cap S = B$.

Now suppose that $|1 - \Delta \|a\|| < \epsilon_0$. If $s \in B$, i.e., $\|s\| \leq \Delta$, then $a^T s \leq \Delta \|a\| < 1 + \epsilon_0$, $s \notin S_2$. Hence $B \cap S_2 = \emptyset$, and $\Omega = B \cap S_1$.

Finally, if $\Delta \|a\| \geq 1 + \epsilon_0$, then $a \neq 0$. Set $s_1 = \min\{1, \Delta\} \frac{(1-\epsilon_0)a}{\|a\| \max\{1, \|a\|\}}$. Then $s_1 \in B$, $a^T s_1 = (1 - \epsilon_0) \frac{\min\{1, \Delta\} \|a\|}{\max\{1, \|a\|\}} \leq 1 - \epsilon_0$, i.e., $s_1 \in S_1$. Set $s_2 = \frac{a}{\|a\|} \Delta$. Then

$s_2 \in B$, $a^T s_2 = \|a\|\Delta \geq 1 + \epsilon_0$, i.e., $s_2 \in S_2$. Hence,

$$B \cap S_1 \neq \emptyset, \quad B \cap S_2 \neq \emptyset.$$

This means that Ω is expressed only by its definition $\Omega = (B \cap S_1) \cup (B \cap S_2)$. $\square$

Therefore, we consider three different cases of (2.1)–(2.2):

$$(2.4) \quad \min \psi(s)$$

$$(2.5) \quad \text{s.t. } \|s\| \leq \Delta$$

with $1 - \epsilon_0 \geq \Delta\|a\|$;

$$(2.6) \quad \min \psi(s)$$

$$(2.7) \quad \text{s.t. } \|s\| \leq \Delta, \quad a^T s \leq 1 - \epsilon_0$$

with $|1 - \Delta\|a\|| < \epsilon_0$;

$$(2.8) \quad \min \psi(s)$$

$$(2.9) \quad \text{s.t. } \|s\| \leq \Delta, \quad a^T s \leq 1 - \epsilon_0 \text{ or } a^T s \geq 1 + \epsilon_0$$

with $\Delta\|a\| \geq 1 + \epsilon_0$.

It is noted that the choice of a in Algorithm 1 in [3] satisfies $\|a\|\Delta \leq 1 - \epsilon_0$ (i.e., set $\alpha_0 = 1 - \epsilon_0$ in [3]) so that (2.4)–(2.5) is a unique subproblem in Algorithm 1. However, the choice of a strongly limits the range of a when Δ is not small, and the conic function may generate a bad local model of the objective function. The limitation is eliminated in this paper since the three problems (2.4)–(2.9) include all cases for the parameters Δ, a , and A . Hence these parameters are properly chosen such that the original function is well approximated by a conic function.

If $(1 - \epsilon_0) \geq \Delta\|a\|$ is not satisfied, we have to solve (2.6)–(2.7) or (2.8)–(2.9) which includes two similar disjunctive problems. However, we can solve the problem (2.8)–(2.9) by using the algorithm for the problem (2.6)–(2.7) with only a few extra computations. Moreover, when (2.1)–(2.2) is used to approximate the original objective function at an iterate point near to the optimal solution, $(1 - \epsilon_0) \geq \Delta\|a\|$ is easily satisfied and (2.1)–(2.2) is reduced to (2.4)–(2.5). Hence, it should not be difficult to design an algorithm for solving the well-defined trust-region subproblem (2.1)–(2.2).

3. Optimality conditions for minimizing conic functions. Davidon [2] proposed optimality conditions for the problem of minimizing a conic function (1.1). In this section, a detailed proof is presented. For this purpose, a lemma from [2] is needed.

LEMMA 3.1. *The gradient of the function in (1.1) is*

$$(3.1) \quad \nabla \psi(s) = \frac{1}{v^2}(I - as^T)^{-1}(vg + As) = \frac{1}{v^3}(vI + as^T)(vg + As),$$

where $v = 1 - a^T s$. This gradient vanishes at s_* , i.e., s_* is a critical point of $\psi(s)$, if and only if

$$(3.2) \quad (A - ga^T)s_* = -g, \quad 1 - a^T s_* \neq 0.$$

The value of $\psi(s)$ is the same at all critical points s_* and equals

$$(3.3) \quad \psi(s_*) = \frac{g^T s_*}{2v_*} = -\frac{s_*^T A s_*}{2v_*^2},$$

where s_* is some critical point, $v_* = 1 - a^T s_*$.

The following lemma is given in [2] without a proof. We provide a proof of the lemma here.

LEMMA 3.2. *Assume that s_* is a critical point of $\psi(s)$. Then s_* is a local minimizer if and only if s_* is a global minimizer if and only if $A \geq 0$ (positive semidefinite). The minimizer is unique if and only if $A > 0$ (positive definite).*

Proof. Consider the quadratic function

$$(3.4) \quad q(s) = (1 - a^T s)^2 (\psi(s) - \psi(s_*)).$$

From (1.1), (3.3), and some calculation, it follows that

$$(3.5) \quad \begin{aligned} q(s) &= (1 - a^T s)g^T s + \frac{1}{2}s^T A s - \psi(s_*)(1 - a^T s)^2 \\ &= \frac{1}{2}s^T \tilde{A}s + \tilde{g}^T s - \psi(s_*), \end{aligned}$$

where

$$(3.6) \quad \tilde{g} = g + 2\psi(s_*)a, \quad \tilde{A} = \nabla^2 q(s) = A - ga^T - ag^T - 2\psi(s_*)aa^T.$$

By (3.2) and (3.3), we have that

$$\begin{aligned} \nabla q(s_*) &= \tilde{A}s_* + \tilde{g} \\ &= (A - ga^T - ag^T)s_* - 2\psi(s_*)a^T s_* a + g + 2\psi(s_*)a \\ &= -g^T s_* a + 2\psi(s_*)(1 - a^T s_*)a = 0 \end{aligned}$$

and

$$(3.7) \quad \tilde{A} = \frac{1}{v_*^2} Q_*^T A Q_*,$$

where

$$(3.8) \quad Q_* = v_* I + s_* a^T, \quad v_* = 1 - a^T s_*.$$

Hence,

$$q(s) = q(s_*) + \nabla q(s_*)(s - s_*) + \frac{1}{2}(s - s_*)^T \tilde{A}(s - s_*) = \frac{1}{2}(s - s_*)^T \tilde{A}(s - s_*),$$

which implies that

$$(3.9) \quad \psi(s) - \psi(s_*) = \frac{1}{2(1 - a^T s)^2} (s - s_*)^T \tilde{A}(s - s_*).$$

Since $1 - a^T s_* \neq 0$, we know that Q_* is nonsingular. Thus this lemma follows from (3.7) and (3.9). $\square$

The introduction of the quadratic function $q(s)$ is very important, and this quadratic function will be used many times later. Thus, we need the following significant result for a trust-region quadratic model that was proved independently by Gay [4] and Sorensen [10].

THEOREM 3.3. A vector s_* is a global solution to the problem

$$(3.10) \quad \min c^T s + \frac{1}{2} s^T H s$$

$$(3.11) \quad \text{s.t. } s \in B$$

if and only if there exists $\lambda_* \geq 0$ such that $H + \lambda_* I$ is positive semidefinite,

$$(3.12) \quad (H + \lambda_* I)s_* = -c, \quad \lambda_*(\|s_*\| - \Delta) = 0, \quad \text{and} \quad \|s_*\| \leq \Delta,$$

where λ_* is unique. If $H + \lambda_* I$ is positive definite, then s_* is a unique global solution of (3.10)–(3.11).

4. Characterization of optimality for basic trust-region subproblem.

When Δ is not large, $(1 - \epsilon_0) \geq \Delta\|a\|$ is easily satisfied. Hence, we call the problem (2.4)–(2.5) a basic trust-region subproblem of a conic model. The following theorem shows a characterization of optimality for (2.4)–(2.5).

THEOREM 4.1. Assume that $\epsilon_0 \in (0, 1)$, $\Delta > 0$, $a \in R^n$ that satisfy $\Delta\|a\| \leq 1 - \epsilon_0$. Then s_* solves (2.4)–(2.5) if and only if there exists $\lambda^* \geq 0$ such that

$$(4.1) \quad (A - ga^T + \lambda^* I)s_* = -g + \lambda^* \Delta^2 a,$$

$$(4.2) \quad \lambda^*(\|s_*\| - \Delta) = 0, \quad \|s_*\| \leq \Delta$$

with $A + \lambda^*(I - \Delta^2 aa^T)$ positive semidefinite. If $A + \lambda^*(I - \Delta^2 aa^T)$ is positive definite, then s_* is a unique solution of (2.4)–(2.5).

Proof. Let s_* be a global solution of the problem (2.4)–(2.5). We define a trust-region subproblem with quadratic function as follows:

$$(4.3) \quad \min q(s) = (1 - a^T s)^2 (\psi(s) - \psi(s_*))$$

$$(4.4) \quad \text{s.t. } \|s\| \leq \Delta.$$

It is easy to find that s_* is also a global solution of the problem (4.3)–(4.4). If $\tilde{s}$ is another global solution of the problem (4.3)–(4.4), then $q(\tilde{s}) = q(s_*) = 0$ and

$$a^T \tilde{s} \leq \|a\|\|\tilde{s}\| \leq \|a\|\Delta \leq 1 - \epsilon_0, \quad 1 - a^T \tilde{s} \geq \epsilon_0,$$

which implies that $\psi(\tilde{s}) - \psi(s_*) = 0$, and that $\tilde{s}$ is also a global solution of the problem (2.4)–(2.5). This means that s_* is a global solution of the problem (2.4)–(2.5) if and only if s_* is a global solution of the problem (4.3)–(4.4).

Applying Theorem 3.3 to the problem (4.3)–(4.4), we obtain that s_* is a global solution of the problem (2.4)–(2.5) if and only if there exists $\lambda^* \geq 0$ such that $\tilde{A} + \lambda^* I$ is positive semidefinite,

$$(4.5) \quad (\tilde{A} + \lambda^* I)s_* = -\tilde{g},$$

and

$$(4.6) \quad \lambda^*(\|s_*\| - \Delta) = 0, \quad \|s_*\| \leq \Delta,$$

where

$$(4.7) \quad \tilde{g} = g + 2\psi(s_*)a, \quad \tilde{A} = \nabla^2 q(s) = \nabla^2 q(s_*) = A - ga^T - ag^T - 2\psi(s_*)aa^T.$$

If $\tilde{A} + \lambda^* I$ is positive definite, then s_* is a unique global solution of (2.4)–(2.5).

Direct calculations from (4.5) show that

$$(4.8) \quad (A - ga^T + \lambda^* I)s_* = -g + (-2\psi(s_*)v_* + g^T s_*)a,$$

where $v_* = 1 - a^T s_*$. Multiplying the above equality by s_*^T , we obtain

$$(4.9) \quad (-2\psi(s_*)v_* + g^T s_*)s_*^T a = s_*^T A s_* + \lambda^* s_*^T s_* + v_* g^T s_*.$$

From the definition of $\psi(s)$, we have

$$(4.10) \quad 2v_*^2 \psi(s_*) = 2v_* g^T s_* + s_*^T A s_*.$$

Combing (4.9) with the above equality, we obtain that

$$(4.11) \quad -2\psi(s_*)v_* + g^T s_* = \lambda^* \Delta^2,$$

which implies that

$$(4.12) \quad (A - ga^T + \lambda^* I)s_* = -g + \lambda^* \Delta^2 a \quad (\text{cf. (4.8)}).$$

Denote

$$(4.13) \quad \bar{A} = A + \lambda^*(I - \Delta^2 aa^T), \quad Q_* = v_* I + s_* a^T.$$

Then from $Q_*^T a = a$ it follows that

$$\begin{aligned} \frac{1}{v_*^2} Q_*^T \bar{A} Q_* &= \frac{1}{v_*^2} (Q_*^T A Q_* + \lambda^* v_*^2 I + \lambda^* v_* (s_* a^T + a s_*^T)) \\ &= \frac{1}{v_*^2} (v_*^2 A + s_*^T A s_* a a^T + \lambda^* v_*^2 I + v_* ((A s_* + \lambda^* s_*) a^T + a (A s_* + \lambda^* s_*)^T)) \\ &= A + \lambda^* I + \frac{s_*^T A s_*}{v_*^2} a a^T + \frac{1}{v_*} ((-v_* g + \lambda^* \Delta^2 a) a^T \\ &\quad + a (-v_* g + \lambda^* \Delta^2 a)^T) \quad (\text{cf. (4.12)}) \\ &= A + \lambda^* I - ga^T - ag^T + \left(\frac{s_*^T A s_* + 2\lambda^* v_* \Delta^2}{v_*^2} \right) a a^T \\ &= A + \lambda^* I - ga^T - ag^T - 2\psi(s_*) a a^T \quad (\text{cf. (4.10) and (4.11)}) \\ (4.14) \quad &= \tilde{A} + \lambda^* I \quad (\text{cf. (4.7)}). \end{aligned}$$

Since Q_* is nonsingular, $\bar{A}$ is positive semidefinite (definite) if and only if $\tilde{A} + \lambda^* I$ is positive semidefinite (definite). Hence, the theorem follows from (4.6), (4.12), and (4.14). $\square$

We note that a similar sufficient condition was proposed in Theorem 3 of [3]. However it is important that in Theorem 4.1 the assumption that $A \geq 0$ is moved out and the simplified condition is a characterization of optimality. In addition, using the dominator of the conic model as a constraint, we replace (2.4)–(2.5) by a problem which minimizes a quadratic with two quadratic constraints (see [5] and [12]). Applying Theorem 4.2 in [5] to this problem, we can obtain a weak necessary condition where $A + \lambda^*(I - \Delta^2 aa^T)$ has at most one negative eigenvalue. This means that the results in Theorem 4.1 in this section are better than others and the proof of this theorem is simple.

It is convenient to solve problem (2.4)–(2.5) when (4.1) becomes

$$(4.15) \quad (A + \lambda^*(I - \Delta^2 aa^T))s_* = v_*(-g + \lambda^* \Delta^2 a).$$

Using the Sherman–Morrison theorem, we can solve this problem by using the algorithms for symmetric coefficient matrices.

5. Optimality conditions for other trust-region subproblems. In this section we discuss the necessary and sufficient conditions for the problems (2.6)–(2.7) and (2.8)–(2.9).

THEOREM 5.1. *Assume that $-\epsilon_0 < \Delta\|a\| - 1 < \epsilon_0$. If s_* solves the problem (2.6)–(2.7), then there exist $\lambda_1^*, \lambda_2^* \geq 0$ such that*

$$(5.1) \quad (A - ga^T + \lambda_1^* I)s_* = -g - \tilde{\lambda}a,$$

$$(5.2) \quad \lambda_1^*(\|s_*\| - \Delta) = 0, \quad \lambda_2^*(a^T s_* - 1 + \epsilon_0) = 0,$$

and $d^T Q_*^T \hat{A} Q_* d \geq 0$ for all d with $a^T d = 0$, where

$$(5.3) \quad \hat{A} = A + \lambda_1^* I + \tilde{\lambda}aa^T, \quad \tilde{\lambda} = \lambda_2^* \epsilon_0 - \lambda_1^* \Delta^2.$$

Proof. Let s_* be a global solution of the problem (2.6)–(2.7). We define a trust-region subproblem with quadratic function as follows:

$$(5.4) \quad \min q(s) = (1 - a^T s)^2 (\psi(s) - \psi(s_*))$$

$$(5.5) \quad \text{s.t. } \|s_*\| \leq \Delta, \quad a^T s \leq 1 - \epsilon_0.$$

It is easy to see that s_* is a global solution of the problem (2.6)–(2.7) if and only if s_* is a global solution of the problem (5.4)–(5.5).

Let $\eta(s) = \frac{1}{2}(s^T s - \Delta^2)$ and $\xi(s) = a^T s - 1 + \epsilon_0$. Then $\nabla\eta(s) = s$ and $\nabla\xi(s) = a$. In order to apply Karush–Kuhn–Tucker (KKT) optimality conditions to the problem (5.4)–(5.5), we need to prove the linear independence of the gradients of active constraints. Suppose that only $\eta(s)$ is active at s_* , i.e., $\|s_*\| = \Delta > 0$. Then $\nabla\eta(s_*)$ is linearly independent. Suppose that only $\xi(s)$ is active at s_* , i.e., $a^T s_* = 1 - \epsilon_0$. Then $\nabla\xi(s_*) = a \neq 0$ is also linearly independent. Now suppose that $\eta(s)$ and $\xi(s)$ are active at s_* , i.e., $\|s_*\| = \Delta > 0$, $a^T s_* = 1 - \epsilon_0$. If $\nabla\eta(s_*)$ and $\nabla\xi(s_*)$ are linearly dependent, then $|a^T s_*| = \|a\|\|s_*\| = \|a\|\Delta > 1 - \epsilon_0$, which is a contradiction to $s_*^T a = 1 - \epsilon_0$. This means that $\nabla\eta(s_*)$ and $\nabla\xi(s_*)$ are linearly independent.

Hence from the KKT first order optimality necessary conditions it follows that there exist $\lambda_1^*, \lambda_2^* \geq 0$ such that

$$(5.6) \quad \nabla q(s_*) + \lambda_1^* s_* + \lambda_2^* a = 0,$$

$$(5.7) \quad \lambda_1^*(\|s_*\|^2 - \Delta^2) = 0, \quad \lambda_2^*(a^T s_* - 1 + \epsilon_0) = 0,$$

where

$$(5.8) \quad \nabla q(s_*) = \tilde{A}s_* + \tilde{g}, \quad \tilde{g} = g + 2\psi(s_*)a, \quad \tilde{A} = A - ga^T - ag^T - 2\psi(s_*)aa^T.$$

From (5.6) it follows that

$$(5.9) \quad (A - ga^T + \lambda_1^* I)s_* = -g + (-2\psi(s_*)v_* + g^T s_* - \lambda_2^*)a.$$

Multiplying the above equality with s_*^T , we obtain

$$(5.10) \quad (-2\psi(s_*)v_* + g^T s_* - \lambda_2^*)s_*^T a = s_*^T A s_* + \lambda_1^* s_*^T s_* + v_* g^T s_*.$$

From (4.10), (5.9), and (5.10) we have

$$(5.11) \quad -2\psi(s_*)v_* + g^T s_* = \lambda_1^* \Delta^2 + \lambda_2^* s_*^T a$$

and

$$(5.12) \quad \begin{aligned} (A - ga^T + \lambda_1^* I)s_* &= -g + (\lambda_1^* \Delta^2 - \lambda_2^* \epsilon_0)a \\ &= -g - \tilde{\lambda}a \quad (\text{cf. (5.3) and (5.7)}). \end{aligned}$$

Hence, (5.1) and (5.2) follow from (5.7) and (5.12).

If we denote

$$(5.13) \quad L(s, \lambda_1^*, \lambda_2^*) = q(s) + \lambda_1^* \eta(s) + \lambda_2^* \xi(s),$$

then from (5.6), (5.8), and $Q_*^T a = a$ we obtain

$$(5.14) \quad \begin{aligned} \nabla L(s_*, \lambda_1^*, \lambda_2^*) &= \nabla q(s_*) + \lambda_1^* s_* + \lambda_2^* a = 0, \\ \nabla^2 L(s_*, \lambda_1^*, \lambda_2^*) &= \nabla^2 q(s_*) + \lambda_1^* I = \tilde{A} + \lambda_1^* I \\ &= A - ga^T - ag^T - 2\psi(s_*)aa^T + \lambda_1^* I. \end{aligned}$$

Hence we have

$$(5.15) \quad \begin{aligned} \frac{1}{v_*^2} Q_*^T \hat{A} Q_* &= \frac{1}{v_*^2} (Q_*^T A Q_* + \lambda_1^* v_*^2 I + \lambda_2^* \epsilon_0 v_*^2 aa^T \\ &\quad + \lambda_1^* v_* (s_* a^T + a s_*^T)) \quad (\text{cf. (3.8), (5.3), and (5.7)}) \\ &= \frac{1}{v_*^2} (v_*^2 A + s_*^T A s_* aa^T + \lambda_1^* v_*^2 I + \lambda_2^* \epsilon_0 v_*^2 aa^T \\ &\quad + v_* ((A s_* + \lambda_1^* s_*) a^T + a (A s_* + \lambda_1^* s_*)^T)) \\ &= A + \lambda_1^* I + \lambda_2^* \epsilon_0 aa^T + \frac{s_*^T A s_*}{v_*^2} aa^T - \frac{1}{v_*} ((v_* g + \tilde{\lambda} a) a^T \\ &\quad + a (v_* g + \tilde{\lambda} a)^T) \quad (\text{cf. (5.12)}) \\ &= A + \lambda_1^* I - ga^T - ag^T + \lambda_2^* \epsilon_0 aa^T + \left(\frac{s_*^T A s_* - 2\tilde{\lambda} v_*}{v_*^2} \right) aa^T \\ &= \tilde{A} + \lambda_1^* I + \lambda_2^* \epsilon_0 aa^T + \left(\frac{4\psi(s_*)v_*^2 - 2v_* g^T s_* - 2\tilde{\lambda} v_*}{v_*^2} \right) aa^T \\ &\quad (\text{cf. (4.10) and (5.8)}) \\ &= \tilde{A} + \lambda_1^* I + \lambda_2^* \epsilon_0 aa^T - \frac{2\lambda_2^*}{v_*} aa^T \quad (\text{cf. (5.7) and (5.11)}) \\ &= \nabla^2 L(s_*, \lambda_1^*, \lambda_2^*) + \lambda_2^* \left(\epsilon_0 - \frac{2}{\epsilon_0} \right) aa^T, \end{aligned}$$

where the last equality follows from (5.14) and $\frac{\lambda_2^*}{v_*} = \frac{\lambda_2^*}{\epsilon_0}$ (see (5.7)).

In order to prove the nonnegativeness property, we need to consider the following three cases.

(1) Suppose that $a^T s_* < 1 - \epsilon_0$. Then $\lambda_2^* = 0$, and s_* is a solution of (4.3)–(4.4). By a similar deduction process in the proof of Theorem 4.1 and (5.3), we see that

$$\frac{1}{v_*^2} Q_*^T \hat{A} Q_* = \frac{1}{v_*^2} Q_*^T (A + \lambda_1^* (I - \Delta aa^T)) Q_*$$

is positive semidefinite. Thus $d^T Q_*^T \hat{A} Q_* d \geq 0$ for all d with $d^T a = 0$.

(2) Suppose that $\|s_*\| < \Delta$ and $a^T s_* = 1 - \epsilon_0$. Then $\lambda_1^* = 0$ and s_* is a local minimizer of

$$\begin{aligned} \min \quad & q(s) \\ \text{s.t.} \quad & \xi(s) = a^T s - (1 - \epsilon_0) = 0. \end{aligned}$$

According to the second order necessary condition of the equality constraint optimization, we obtain that $d^T \nabla^2 L(s_*, 0, \lambda_2^*) d \geq 0$ for all $d \in R^n$ with $d^T a = 0$. Thus from (5.15) we have

$$d^T Q_*^T \hat{A} Q_* d = v_*^2 d^T \nabla^2 L(s_*, 0, \lambda_2^*) d + v_*^2 \lambda_2^* \left(\epsilon_0 - \frac{2}{\epsilon_0} \right) (a^T d)^2 \geq 0$$

for all $d \in R^n$ with $d^T a = 0$.

(3) Suppose that $\|s_*\| = \Delta$ and $a^T s_* = 1 - \epsilon_0$. According to the second order necessary condition of the equality constraint optimization, we have

$$d^T \nabla^2 L(s_*, \lambda_1^*, \lambda_2^*) d \geq 0$$

for all $d \in R^n$ with $d^T a = 0$, $d^T s_* = 0$. Hence,

$$\begin{aligned} d^T Q_*^T \hat{A} Q_* d &= v_*^2 d^T \nabla^2 L(s_*, \lambda_1^*, \lambda_2^*) d + v_*^2 \lambda_2^* \left(\epsilon_0 - \frac{2}{\epsilon_0} \right) (a^T d)^2 \\ &= v_*^2 d^T \nabla^2 L(s_*, \lambda_1^*, \lambda_2^*) d \geq 0 \end{aligned}$$

for all $d \in R^n$ with $d^T a = 0$, $d^T s_* = 0$.

For $d \in R^n$, $d^T a = 0$, $d^T s_* \neq 0$, set $t = -\frac{2d^T s_*}{\|d\|^2}$. Then

$$\eta(s_* + td) = 0, \quad \xi(s_* + td) = (s_* + td)^T a - 1 + \epsilon_0 = 0$$

and

$$\begin{aligned} 0 &\leq q(s_* + td) - q(s_*) \\ &= L(s_* + td, \lambda_1^*, \lambda_2^*) - L(s_*, \lambda_1^*, \lambda_2^*) \quad (\text{cf. (5.13)}) \\ &= t \nabla L(s_*, \lambda_1^*, \lambda_2^*)^T d + \frac{1}{2} t^2 d^T \nabla^2 L(s_*, \lambda_1^*, \lambda_2^*) d \\ &= \frac{1}{2} t^2 d^T \nabla^2 L(s_*, \lambda_1^*, \lambda_2^*) d. \end{aligned}$$

Hence,

$$d^T Q_*^T \hat{A} Q_* d = v_*^2 d^T \nabla^2 L(s_*, \lambda_1^*, \lambda_2^*) d + v_*^2 \lambda_2^* \left(\epsilon_0 - \frac{2}{\epsilon_0} \right) (a^T d)^2 \geq 0$$

for all $d \in R^n$ with $d^T a = 0$, $d^T s_* \neq 0$. $\square$

A bit stronger sufficient optimality condition for (2.6)–(2.7) is given in the following. However, it is easy to check this condition.

THEOREM 5.2. *Let $A \in R^{n \times n}$ be symmetric and assume that $-\epsilon_0 < \Delta \|a\| - 1 < \epsilon_0$. If there exist $\lambda_1^*, \lambda_2^* \geq 0$, $s_* \in R^n$ such that*

$$(5.16) \quad (A - ga^T + \lambda_1^* I) s_* = -g - \tilde{\lambda} a,$$

$$(5.17) \quad \lambda_1^* (\|s_*\| - \Delta) = 0, \quad \lambda_2^* (a^T s_* - 1 + \epsilon_0) = 0, \quad a^T s_* \leq 1 - \epsilon_0, \quad \text{and} \quad \|s_*\| \leq \Delta$$

with $\hat{A} = A + \lambda_1^* I + \tilde{\lambda} a a^T \geq 0$, where $\tilde{\lambda} = \lambda_2^* \epsilon_0 - \lambda_1^* \Delta^2$, then s_* solves (2.6)–(2.7); i.e., s_* satisfies

$$\begin{aligned} \min \quad & \psi(s) \\ \text{s.t.} \quad & \|s\| \leq \Delta, \quad a^T s \leq 1 - \epsilon_0. \end{aligned}$$

Proof. We consider two different cases.

(1) Suppose that $\lambda_1^* = 0$, $\lambda_2^* = 0$. Then (5.16) means

$$(A - g a^T) s_* = -g.$$

Since $A = \hat{A} \geq 0$, it follows from Lemma 3.2 that s_* is a global minimizer of $\psi(s)$. Hence, (5.16) and (5.17) guarantee that s_* is a solution to (2.6)–(2.7).

(2) Suppose that $\lambda_1^* + \lambda_2^* > 0$. Define

$$(5.18) \quad \psi_{\lambda_1 \lambda_2}(s) = \frac{(g + \tilde{\lambda} a)^T s}{1 - a^T s} + \frac{1}{2} \frac{s^T (A + \lambda_1^* I + \tilde{\lambda} a a^T) s}{(1 - a^T s)^2}.$$

The critical point of $\psi_{\lambda_1 \lambda_2}(s)$ should satisfy

$$[A + \lambda_1^* I + \tilde{\lambda} a a^T - (g + \tilde{\lambda} a) a^T] s = -(g + \tilde{\lambda} a), \quad 1 - a^T s \neq 0,$$

i.e.,

$$(A + \lambda_1^* I - g a^T) s = -g - \tilde{\lambda} a, \quad 1 - a^T s \neq 0.$$

Hence, (5.16), (5.17), and $\hat{A} = A + \lambda_1^* I + \tilde{\lambda} a a^T \geq 0$ imply that s_* is a global minimizer of $\psi_{\lambda_1 \lambda_2}(s)$. Hence we obtain

$$(5.19) \quad \psi_{\lambda_1 \lambda_2}(s) \geq \psi_{\lambda_1 \lambda_2}(s_*) \quad \text{for all } s \in R^n, \quad 1 - a^T s \geq \epsilon_0, \quad \|s\| \leq \Delta.$$

After some calculation, we have

$$(5.20) \quad \psi(s) = \psi_{\lambda_1 \lambda_2}(s) + \bar{\psi}(s),$$

where

$$(5.21) \quad \bar{\psi}(s) = \frac{1}{2} \tilde{\lambda} + \frac{\lambda_1^* (\Delta^2 - s^T s)}{2(1 - a^T s)^2} - \frac{\lambda_2^* \epsilon_0}{2(1 - a^T s)^2}.$$

In addition, from (5.17) it follows that

$$(5.22) \quad \lambda_2^* \epsilon_0^2 = \epsilon_0 \lambda_2^* (1 - a^T s_*) = \lambda_2^* (1 - a^T s_*)^2.$$

Now (5.19)–(5.22) imply that

$$\psi(s) \geq \psi_{\lambda_1 \lambda_2}(s_*) + \bar{\psi}(s) \geq \psi_{\lambda_1 \lambda_2}(s_*) + \frac{1}{2} \tilde{\lambda} - \frac{\lambda_2^*}{2\epsilon_0} = \psi(s_*)$$

for all $s \in R^n$, $1 - a^T s \geq \epsilon_0$, $\|s\| \leq \Delta$. Hence, s_* is a global minimizer of (2.6)–(2.7). $\square$

For the problem (2.8)–(2.9), there is only one point in set S_2 (see (2.3)) when $\Delta\|a\| = 1 + \epsilon_0$. It is an interesting case which has to be separately considered.

THEOREM 5.3. Assume that $\Delta\|a\| \geq 1 + \epsilon_0$. If s_* solves (2.8)–(2.9), then there exist $\lambda_1^*, \lambda_2^* \geq 0$ such that

$$(5.23) \quad (A - ga^T + \lambda_1^* I)s_* = -g - \tilde{\lambda}a,$$

$$(5.24) \quad \lambda_1^*(\|s_*\| - \Delta) = 0,$$

and

$$(5.25) \quad \begin{cases} \lambda_2^*(a^T s_* - 1 + \epsilon_0) = 0 & \text{if } a^T s_* \leq 1 - \epsilon_0, \\ \lambda_2^*(a^T s_* - 1 - \epsilon_0) = 0 & \text{if } a^T s_* \geq 1 + \epsilon_0 \text{ and } \Delta\|a\| > 1 + \epsilon_0, \end{cases}$$

and $d^T Q_*^T \hat{A} Q_* d \geq 0$ for all d with $a^T d = 0$, where $\tilde{\lambda}$ and $\hat{A}$ are defined by (5.3).

If $\Delta\|a\| = 1 + \epsilon_0$ and s_* solves (2.8)–(2.9) and $a^T s_* \geq 1 + \epsilon_0$, then $s_* = \frac{\Delta}{\|a\|}a$.

Proof. If s_* satisfies $a^T s_* \leq 1 - \epsilon_0$, then the proof of this theorem is the same as that of Theorem 5.1. If s_* satisfies $a^T s_* \geq 1 + \epsilon_0$ and $\Delta\|a\| > 1 + \epsilon_0$, then we need to prove that a and s_* are linearly independent when two constraint functions are active at s_* , i.e., $\|s_*\| = \Delta$ and $a^T s_* = 1 + \epsilon_0$. Suppose that it is not true. Then there exists a constant c such that $a = cs_*$. From $a^T s_* = 1 + \epsilon_0$ and $\|s_*\| = \Delta$ it follows that

$$c = \frac{1 + \epsilon_0}{\Delta^2} = \frac{\|a\|}{\Delta}.$$

This means that $\|a\|\Delta = 1 + \epsilon_0$, which is in contradiction with $\|a\|\Delta > 1 + \epsilon_0$. Hence we can use the KKT first order optimality necessary condition to obtain the conclusion of this theorem by a similar deduction process in the proof of Theorem 5.1.

Suppose that s_* solves (2.8)–(2.9), $\|a\|\Delta = 1 + \epsilon_0$, and $a^T s_* \geq 1 + \epsilon_0$. Since $a^T s_* \leq \|a\|\|s_*\| \leq \|a\|\Delta = 1 + \epsilon_0$,

$$a^T s_* = 1 + \epsilon_0 = \|a\|\Delta = \|a\|\|s_*\|,$$

which implies that s_* and a are linearly dependent, and

$$s_* = \frac{\Delta}{\|a\|}a. \quad \square$$

The following theorem gives a sufficient condition for a solution to the problem (2.8)–(2.9).

THEOREM 5.4. Let $A \in R^{n \times n}$ be symmetric and assume that $\Delta\|a\| \geq 1 + \epsilon_0$. If there exist $\lambda_1^*, \lambda_2^* \geq 0$, $s_* \in R^n$ such that

$$(5.26) \quad (A - ga^T + \lambda_1^* I)s_* = -g - \tilde{\lambda}a,$$

$$(5.27) \quad \lambda_1^*(\|s_*\| - \Delta) = 0, \quad \|s_*\| \leq \Delta,$$

and

$$(5.28) \quad \begin{cases} \lambda_2^*(a^T s_* - 1 + \epsilon_0) = 0 & \text{if } a^T s_* \leq 1 - \epsilon_0, \\ \lambda_2^*(a^T s_* - 1 - \epsilon_0) = 0 & \text{if } a^T s_* \geq 1 + \epsilon_0 \end{cases}$$

with $\hat{A} = A + \lambda_1^* I + \tilde{\lambda}aa^T \geq 0$, then s_* solves (2.8)–(2.9).

The proof of this theorem is similar to that of Theorem 5.2 and is omitted.

6. Comments. In this paper the optimality conditions of the trust-region subproblem of a conic model are obtained for different cases. These conditions may provide a sound basis for the further development of the efficient methods for trust-region subproblems of a conic model.

Acknowledgments. I am grateful to two anonymous referees for their careful reading of the original draft and their detailed comments on revising the original draft. I am also grateful to Professor Gould, Professor Rubinov, Professor Ding, and Dr. Deng for their help.

REFERENCES

- [1] K. A. ARIYAWANSA, *Deriving collinear scaling algorithms as extensions of quasi-Newton methods and the local convergence of DFP- and BFGS-related collinear scaling algorithms*, Math. Program., 49 (1990), pp. 23–48.
- [2] W. C. DAVIDON, *Conic approximations and collinear scalings for optimizers*, SIAM J. Numer. Anal., 17 (1980), pp. 268–281.
- [3] S. DI AND W. Y. SUN, *A trust region method for conic model to solve unconstrained optimization*, Optim. Methods Softw., 6 (1996), pp. 237–263.
- [4] D. M. GAY, *Computing optimal locally constrained steps*, SIAM J. Sci. Stat. Comput., 2 (1981), pp. 186–197.
- [5] J.-M. PENG AND Y.-X. YUAN, *Optimality conditions for the minimization of a quadratic with two quadratic constraints*, SIAM J. Optim., 7 (1997), pp. 579–594.
- [6] M. J. D. POWELL, *Convergence properties of a class of minimization algorithms*, in Nonlinear Programming 2, O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds., Academic Press, New York, 1975, pp. 1–27.
- [7] M. J. D. POWELL AND Y. X. YUAN, *A trust region algorithm for equality constrained optimization*, Math. Program., 49 (1991), pp. 189–211.
- [8] R. B. SCHNABEL, *Conic methods for unconstrained minimization and tensor methods for nonlinear equations*, in Mathematical Programming: The State of the Art, A. Bachem, M. Grotschel, and B. Korte, eds., Springer-Verlag, Berlin, 1983, pp. 417–438.
- [9] D. C. SORESENSEN, *The Q-superlinear convergence of a collinear scaling algorithm for unconstrained optimization*, SIAM J. Numer. Anal., 17 (1980), pp. 84–114.
- [10] D. C. SORESENSEN, *Newton's method with a model trust region modification*, SIAM J. Numer. Anal., 19 (1982), pp. 409–426.
- [11] W. Y. SUN AND Y. X. YUAN, *A conic trust-region method for nonlinearly constrained optimization*, Ann. Oper. Res., 103 (2001), pp. 175–191.
- [12] Y. X. YUAN, *A dual algorithm for minimizing a quadratic function with two quadratic constraints*, J. Comput. Math., 9 (1991), pp. 349–359.
- [13] Y. X. YUAN, *A review of trust region algorithms for optimization*, in ICIAM99: Proceedings of the Fourth International Congress on Industrial and Applied Mathematics, J. M. Ball and J. C. R. Hunt, eds., Oxford University Press, Oxford, UK, 2000, pp. 271–282.

A COLUMN GENERATION APPROACH TO RADIATION THERAPY TREATMENT PLANNING USING APERTURE MODULATION*

H. EDWIN ROMEIJN[†], RAVINDRA K. AHUJA[†], JAMES F. DEMPSEY[‡], AND
ARVIND KUMAR[†]

Abstract. This paper considers the problem of radiation therapy treatment planning for cancer patients. During radiation therapy, beams of radiation pass through a patient. This radiation kills both cancerous and normal cells, so the radiation therapy must be carefully planned to deliver a clinically prescribed dose to certain targets while sparing nearby organs and tissues. Currently, a technique called *intensity modulated radiation therapy* (IMRT) is considered to be the most effective radiation therapy for many forms of cancer. In IMRT, the patient is irradiated from several different directions. From each direction, one or more irregularly shaped radiation beams of uniform intensity are used to deliver the treatment. This paper deals with the problem of designing a treatment plan for IMRT that determines an optimal set of such shapes (called apertures) and their corresponding intensities. This is in contrast with established two-stage approaches where, in the first phase, each radiation beam is viewed as consisting of a set of individual beamlets, each with its own intensity. A second phase is then needed to approximate and decompose the optimal intensity profile into a set of apertures with corresponding intensities. The problem is formulated as a large-scale convex programming problem, and a column generation approach to deal with its dimensionality is developed. The associated pricing problem determines, in each iteration, one or more apertures to be added to our problem. Several variants of this pricing problem are discussed, each corresponding to a particular set of constraints that the apertures must satisfy in one or more of the currently available types of commercial IMRT equipment. Polynomial-time algorithms are presented for solving each of these variants of the pricing problem to optimality. Finally, the effectiveness of our approach is demonstrated on clinical data.

Key words. radiation therapy, aperture modulation, convex optimization, column generation, pricing problem

AMS subject classifications. 90C06, 90C25, 90C90, 92C50

DOI. 10.1137/040606612

1. Introduction. Every year, approximately 300,000 of the U.S. citizens who are newly diagnosed with cancer may benefit from radiation therapy (American Cancer Society [3], Murphy, Lawrence, and Lenard [24], Perez and Brady [28], and Steel [42]). Despite sophisticated radiation therapy treatments, a significant number of patients who are initially considered curable die of their disease. Others may see their quality of life seriously reduced due to unintended side effects from radiation therapy. The main cause of this is that radiation therapy treatment plans often underdose the targets, overdose healthy organs, or both. Thus, the preservation of healthy or functional tissues, and hence the quality of a patient's life, must be balanced against the probability of the eradication of the patient's disease. In this paper, we consider a novel approach to the efficient planning of radiation therapy treatments that integrates treatment design with delivery issues that have previously been handled separately.

*Received by the editors August 24, 2003; accepted for publication (in revised form) August 13, 2004; published electronically April 22, 2005.

<http://www.siam.org/journals/siopt/15-3/60661.html>

[†]Department of Industrial and Systems Engineering, University of Florida, 303 Weil Hall, P.O. Box 116595, Gainesville, FL 32611-6595 (romeijn@ise.ufl.edu, ahuja@ufl.edu, arvind22@ufl.edu).

[‡]Department of Radiation Oncology, University of Florida, P.O. Box 100385, Gainesville, FL 32610-0385 (dempsey@ufl.edu).

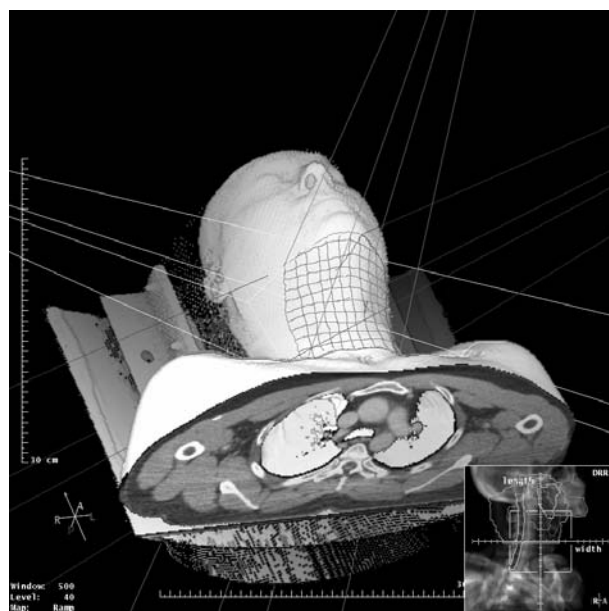


FIG. 1.1. Irradiating a head-and-neck cancer patient using beams from different directions.

1.1. Treatment delivery. During radiation therapy, beams of radiation pass through a patient. These beams deposit energy along their path, which kills both cancerous and normal cells. Delivering a clinically prescribed radiation dose to cancerous cells while sparing normal cells in nearby organs and tissues requires a careful design of the radiation therapy treatment. Typically, there are several clinical *targets* that we wish to irradiate, and there are several nearby organs, called *critical structures*, that we wish to spare. Note that we usually treat targets which contain known tumors, as well as additional regions to which the disease may have spread or that account for patient motion during the treatment. It may be possible to kill all cells in the targets with a single beam of radiation. However, such a treatment can be expected to significantly damage normal cells in critical structures located along the path of the beam. When beams are delivered from a number of different orientations spaced around the patient, the intersection of these beams defines a region that receives the highest radiation dose, whereas regions covered by a single beam or only a few beams receive much lower radiation doses. The goal is now to design the treatment plan, i.e., the shapes and the radiation intensities of the beams, in such a way that the intersecting regions receiving the highest radiation dose conform closely to the targets, while the critical structures receive dose from some, but not all, beams and can thus be spared. (See Figure 1.1.)

Patients receiving radiation therapy are typically treated on a clinical radiation delivery device called a linear accelerator. The radiation source is contained in the accelerator head and can be viewed as a point source of high-energy photons. The patient is immobilized with restraints on a couch, and the accelerator head can rotate for a full 360° about a point that is 1 meter from the source, called the isocenter. It is common in practice to use a single, fixed couch location, and use a very limited number of locations (3–9) of the accelerator head in the circle around the isocenter.

Conformal radiation therapy seeks to conform the geometric shape of the delivered radiation dose as closely as possible to that of the intended targets. In *conventional*

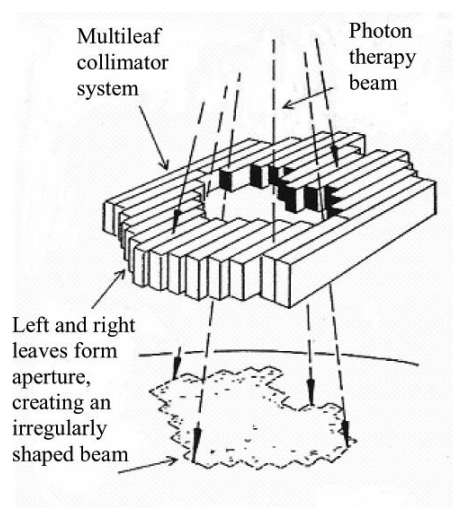


FIG. 1.2. A multileaf collimator (MLC) system.

conformal radiation therapy, one attempts to achieve this by delivering a radiation dose by using a relatively small number of different beam shapes, called *apertures*, each with uniform intensity level. Such apertures are formed by partially blocking the radiation source. In this way we may, for example, create an aperture whose shape conforms to the beam's-eye-view of the targets in the patient as seen from a certain beam orientation. Since the geometry of the targets and critical structure is different in each patient, customized apertures have to be manufactured for each patient. Despite this practical limitation, there are several advantages to conventional conformal radiation therapy. First, the fact that only a limited set of apertures is used leads to a short patient treatment time. Shorter treatment times not only limit the discomfort of the patient, but also ensure that the exposure to radiation of the patient as a whole is limited. Additional advantages lie in the ability to accurately deliver a planned treatment, due to the relative ease with which the cumulative dose received by the patient for a given set of apertures and intensities can be determined. Finally, treatment plans using only a few (relatively large) apertures are typically insensitive to patient setup errors as well as patient motion during treatment.

Recent technological advancements have lead to the rapid development and widespread clinical implementation of an external-beam radiation delivery technique known as *intensity modulated radiation therapy* (IMRT) (see Webb [47]). An IMRT capable linear accelerator is equipped with a so-called *multileaf collimator* (MLC) system (see Figure 1.2), which is used to generate a desired dose distribution. By dynamically blocking different parts of the beam, the MLC system is able to efficiently deliver treatment plans that use a much larger number of different apertures than conventional conformal radiation therapy. IMRT thus allows for the creation of very complex nonuniform dose distributions that allow the delivery of sufficiently high radiation doses to targets while limiting the radiation dose delivered to healthy tissues. The advent of IMRT has dramatically improved treatment efficiency and offers the potential for large improvements in treatment quality.

1.2. Traditional two-phase approach to treatment planning. In IMRT we often model each beam as a collection of hundreds of small beamlets, the intensities of which can be controlled individually. The problem of finding an optimal intensity

profile (also called fluence map) for each beam is called the *fluence map optimization* (FMO) problem. This problem has received a lot of attention in recent years from both operations researchers as well as medical physicists (see, e.g., Shepard et al. [38] for a review, and also Hamacher and Küfer [16]; Lee, Fox, and Crocker [22, 23]; Bortfeld et al. [12]; and Romeijn et al. [31, 32]). However, MLC-based IMRT delivery systems cannot directly deliver such fluence maps since the radiation source can only deliver a uniform intensity at any time. Therefore, a fluence map needs to be decomposed into (and approximated by) a number of apertures, each with its own associated uniform intensity. This decomposition problem is a nontrivial optimization problem in its own right and is often solved in a separate phase (see, e.g., Bortfeld et al. [11], Xia and Verhey [50], Siochi [40], Boland, Hamacher, and Lenzen [10], Kamath et al. [18], and Ahuja and Hamacher [1]). To allow a decomposition of the fluence map into a manageable number of apertures, the beamlet intensities in a given beam are first discretized to levels that are typically chosen to be multiples of 5%–20% of their maximal value. It is easy to see that this discretization may cause a significant deterioration of the treatment plan quality, in particular when not integrated into the FMO problem. Moreover, despite the discretization step, the number of apertures required to deliver the fluence map is often relatively large (on the order of 20–30 per beam, for a total of up to 140–210 apertures), which may cause undesirably long treatment times, as well as an unnecessarily high level of sensitivity of the treatment plan to patient setup errors or patient motion. Finally, a two-stage approach is usually rigid, in the sense that a sound trade-off between treatment time and treatment plan quality cannot be made due to the decoupling of the decomposition phase from the treatment plan optimization phase.

1.3. Aperture modulation approach to treatment planning. An approach to the FMO problem that deals explicitly with apertures rather than individual beamlets is called *aperture modulation*. Recently, Siebers et al. [39] have proposed incorporating the aperture choice into a heuristic local search method for FMO. Bednarz et al. [7] developed a mixed-integer linear programming model that incorporates a predetermined pool of apertures from which to choose. Finally, Shepard et al. [37] have proposed a simulated annealing algorithm for finding a treatment plan using only a limited number of apertures. These approaches have been applied to head-and-neck and prostate cancer cases, and empirical experiments suggest that one can obtain high-quality treatment plans with only a limited number of apertures. However, all these approaches are heuristic in nature, and cannot guarantee that an optimal solution will be found. The key to obtaining the best apertures seems to be to formally integrate the FMO and decomposition problems.

In this paper, we propose a new optimization formulation of the aperture modulation approach to FMO. This formulation is based on a convex optimization formulation of the traditional beamlet-based FMO problem as developed in Romeijn et al. [31, 32]. However, in contrast with these earlier papers, in this paper we represent a treatment plan explicitly in terms of apertures and their associated intensities. This model thus integrates the first stage problem of obtaining a set of fluence maps (i.e., a set of individual beamlet intensities) with the second-stage problem of decomposing these fluence maps into a set of deliverable apertures and intensities. However, the number of allowable apertures, and therefore the number of decision variables and constraints in the new problem formulation, is typically enormous. To overcome this, we propose a column generation approach to solving the convex programming problem. A column generation approach requires the repeated solution of a subprob-

lem called the pricing problem. Solving this problem identifies one or more suitable apertures to add to a pool of allowable apertures.

The very nature of our algorithm thus allows us to study the effect of adding apertures on the quality of the treatment plan, thereby enabling a better trade-off between the number of apertures (i.e., treatment time) and treatment plan quality. One of the goals of our paper is to determine whether high-quality treatment plans may be obtained using a limited number of apertures, which would mean that the higher quality and efficiency of IMRT treatments can be achieved without sacrificing the advantages of conventional conformal radiation therapy mentioned above.

The set of feasible apertures from which the pricing problem may choose is dictated by the particular characteristics of the MLC system that is used to deliver the treatment. We study three different variants of the pricing problem, each incorporating a set of constraints on deliverable apertures that is commonly encountered in one or more of the commercially available IMRT systems. We provide algorithms that solve these subproblems in polynomial time in the number of beamlets.

1.4. Treatment plan quality and optimization. The goal of radiation therapy treatment planning is to design a treatment plan that delivers a specific level of radiation dose, a so-called *prescription dose*, to the targets, while on the other hand sparing critical structures by ensuring that the level of radiation dose received by these structures does not exceed some structure-specific *tolerance dose*. These two goals are inherently contradictory if the targets are located in the vicinity of critical structures. This is especially problematic for certain cancers, such as tumors in the head-and-neck area, which are often located very close to, for instance, the spinal cord, brainstem, and salivary glands. In order to formulate a (beamlet- or aperture-based) FMO problem, a quantitative measure of treatment plan quality needs to be devised that appropriately makes the trade-off between the contradictory goals. This is a difficult issue since there is no fundamental basis for treatment plan evaluation.

Many models for FMO that have been proposed to date use a convex objective function that penalizes the doses received by individual points in the patient (see, e.g., Shepard et al. [38] for a review, and also Lee, Fox, and Crocker [22, 23] and Romeijn et al. [31, 32]). An alternative approach that has received some attention in recent years is to view the FMO problem as a multicriteria optimization problem (see, e.g., Küfer, Hamacher, and Bortfeld [19], Hamacher and Küfer [16], Thieke, Bortfeld, and Küfer [44], Thieke et al. [45], Bortfeld et al. [12], and Lahanas, Schreibmann, and Baltas [20]). In this approach several treatment plan evaluation criteria are identified, usually one or two per structure. These criteria can be based on biological considerations (such as measures of equivalent uniform dose or estimated probabilities of tumor control and normal tissue complications) or physical considerations (such as the convex penalty functions mentioned above). In Romeijn, Dempsey, and Li [33] we have shown that most of the treatment plan evaluation criteria proposed in the medical physics literature are *equivalent* to convex penalty function criteria when viewed as a multicriteria optimization problem, in the sense that for each set of treatment plan evaluation criteria from a very large class there exists a class of convex penalty functions that produces an identical Pareto efficient frontier. In this paper, we have therefore chosen to use a convex penalty function-based approach to evaluating treatment plans. Clearly, our approach could be used in a multicriteria setting. However, our experience in Romeijn et al. [31, 32] and in section 5 of this paper suggests that it is possible to quantify a trade-off between the different evaluation criteria that produces high-quality treatment plans for a population of patients, eliminating the need

to solve the FMO problem as a multicriteria optimization problem for each individual patient.

A notable class of treatment plan evaluation criteria that is not equivalent to penalty function criteria is formed by criteria that are based on properties of the entire dose distribution in a given structure. We will enhance our basic convex penalty function model for FMO with such criteria.

1.5. Organization of the paper. This paper is organized as follows. In section 2 we develop our basic model formulation. In section 3 we then develop a column generation approach to solving the optimization problem, including a detailed discussion of variants of the pricing problems and solution methods to solve these. In section 4 we enhance the model, incorporating additional constraints that can control the quality of the obtained treatment plans. In section 5 we discuss the results of our approach and we end the paper with some concluding remarks and directions for future research in section 6.

2. Model formulation.

2.1. Notation and dose calculation. The *structures* (which include both targets and critical structures) are irradiated using a predetermined set of *beams*, denoted by B . Each beam is decomposed into a rectangular grid of *beamlets* with m rows and n columns (where m and n are usually on the order of 10–20, yielding on the order of 100–400 beamlets per beam). A typical beamlet is of size $1 \times 1 \text{ cm}^2$. The position and intensity of all beamlets in a beam can be represented by a two-dimensional $(m \times n)$ -matrix, often called the *intensity matrix*. Each element in this matrix is called a *bixel*, and its value represents the *intensity* (or, more correctly, fluence; see Webb and Lomax [48]) of the corresponding beamlet. We denote the set of all bixels by $N \equiv B \times \{1, \dots, m\} \times \{1, \dots, n\}$. The set of all bixels contained in beam $\ell \in B$ is then denoted by $N_\ell \equiv \{\ell\} \times \{1, \dots, m\} \times \{1, \dots, n\}$. The core task in IMRT treatment planning is to find radiation intensities (sometimes also called *weights*) for all beamlets.

We denote the total number of structures by S , and the number of targets by T . For convenience, we assume that the targets are indexed by $s = 1, \dots, T$, and the critical structures by $s = T + 1, \dots, S$. Furthermore, each of the structures s is discretized into a finite number v_s of cubes, typically of size $4 \times 4 \times 4 \text{ mm}^3$. In the context of radiation therapy, these cubes are usually called *voxels*. The number of voxels required to accurately represent the targets and surrounding critical structures is typically around 100,000. Note that, in general, structures may overlap. For instance, if a target has invaded a critical structure, some voxels will be in both the target and the critical structure. However, in our model we associate a unique (i.e., dominant) structure with each voxel, based on a priority list of all structures (where targets usually have the highest priorities, followed by the critical structures). Note that this is not carried over to our treatment planning system, where the obtained treatment plans are always appropriately evaluated using the original, overlapping, structures.

In our aperture modulation formulation of the FMO problem, we assume that we are given a set of apertures, say K , and choose the decision variables to be the intensities of these apertures, say y_k ($k \in K$). Note that the total number of potential apertures that needs to be included in the model is very large. For example, consider an MLC that allows all combinations of left and right leaf settings. With the $m \times n$ bixel grid for a single beam, there are thus $\frac{1}{2}n(n-1) + 1$ potential combinations of

right and left leaf settings for each beamlet row, for a total of $(\frac{1}{2}n(n-1) + 1)^m - 1$ potential apertures per beam angle. With $n = m = 10$ and 7 beams, this would yield approximately 3×10^{17} apertures to consider. Nevertheless, we will formulate the problem taking all potential apertures into account, and then propose a column generation approach for solving the problem.

Let K_ℓ denote the set of allowable apertures for beam $\ell \in B$, let $K \equiv \cup_{\ell \in B} K_\ell$ denote the set of all apertures, and let $A_k \subset N$ denote the set of beamlets that are exposed in aperture $k \in K$. As a beam of radiation passes through a patient, it deposits an amount of energy along its path that decreases with the distance travelled, but is linearly dependent on the intensity of the aperture. Denoting the dose received by voxel j in structure s from aperture k at unit intensity by $\mathcal{D}_{kjs}$, we can express the dose received by voxel j in structure s as a function of the vector of aperture intensities $y \in \mathbb{R}^{|K|}$ as $\sum_{k \in K} \mathcal{D}_{kjs} y_k$. Next, we will express the so-called dose deposition coefficients $\mathcal{D}_{kjs}$ in terms of the beamlets comprising each aperture. To this end, we first denote the dose received by voxel j in structure s from beamlet i (or, more accurately, an aperture exposing this beamlet only) at unit intensity by D_{ijs} . All traditional beamlet-based FMO formulations then proceed by assuming that the amount of dose deposited to a given voxel per unit intensity of any aperture can be determined by adding the dose deposited by each of the individual beamlets comprising the aperture. However, this assumption is not entirely accurate since the actual dose delivered to each voxel by a given beamlet depends on the shape of the aperture in which it is contained. This dependence cannot be incorporated into traditional beamlet-based FMO models since the decomposition of the fluence maps into a set of deliverable apertures is done in a second phase, outside of the model. A major additional advantage of using an aperture-based approach is that it is possible to improve on the accuracy of the dose calculation by accounting for this dependency. To this end, we define nonnegative intensity correction factors γ_k for all apertures $k \in K$, representing the dependence of the delivered dose on the shape of the aperture. (See, e.g., Clarkson [15], Bjärngård and Siddon [9], and Sanz [36] for methods to compute these factors.) We can then express the dose deposition coefficients as $\mathcal{D}_{kjs} = \gamma_k \sum_{i \in A_k} D_{ijs}$, leading to the following expression for the dose z_{js} received by voxel j in structure s :

$$z_{js} = \sum_{k \in K} \gamma_k \left(\sum_{i \in A_k} D_{ijs} \right) y_k, \quad j = 1, \dots, v_s, \quad s = 1, \dots, S.$$

2.2. A basic model. Our basic model for FMO uses a voxel-based objective function, where a penalty is assigned to each voxel based on the dose it receives under a given set of beamlet intensities. A convex penalty function for voxel j in structure s is denoted by F_{js} . Typically, these functions are chosen to be piecewise polynomial functions, allowing for a different penalization of dose received above or below some threshold. In addition, the same penalty function is often, though not necessarily, applied to all voxels that are in the same structure. For voxels that are in critical structures and not in targets, the penalty function is commonly chosen to be one-sided, where the penalty is equal to zero for doses below a threshold and positive above the threshold.

An aperture modulation formulation of the FMO problem then reads as follows:

$$\text{minimize } \sum_{s=1}^S \sum_{j=1}^{v_s} F_{js}(z_{js})$$

subject to (AM₁)

$$(2.1) \quad \sum_{k \in K} \gamma_k \left(\sum_{i \in A_k} D_{ijs} \right) y_k = z_{js}, \quad j = 1, \dots, v_s, \quad s = 1, \dots, S,$$

$$(2.2) \quad y_k \geq 0, \quad k \in K.$$

Note that if all apertures that consist of a single beamlet are deliverable (which is typically the case), the aperture modulation formulation (AM₁) is equivalent to the beamlet-based formulation of the FMO problem as presented in Romeijn et al. [31, 32]. Apart from being able to solve for apertures and associated intensities directly without a postprocessing phase, this equivalence will allow us to test the hypothesis that using relatively few apertures often yields near-optimal solutions to the FMO problem.

3. A column generation approach.

3.1. Introduction. As was noted above, the number of variables in this formulation is much too large for practical purposes. However, it can be expected that in the optimal solution to this problem relatively few apertures will actually be used (i.e., have positive intensity). In fact, empirical evidence of this conjecture exists; see, e.g., Siebers et al. [39], Bednarz et al. [7], and Shepard et al. [37]. This means that if we were able to identify a set of “good” apertures, we could obtain a (near-) optimal solution to the problem by incorporating only these apertures. Our approach to solving (AM₁) is to use column generation, which is a formalization of this idea. In this method, we start by choosing a limited set of apertures and solve the restricted (AM₁) using only that set of apertures. (Implicitly, this means that we restrict the intensities of the remaining apertures to be equal to zero.) Solving the restricted problem to optimality, we use the Karush–Kuhn–Tucker (KKT)-conditions for optimality to check whether the corresponding dual solution is feasible for the dual problem of the complete problem (AM₁). If so, we have obtained a pair of feasible primal and dual solutions with the same objective function value, and therefore the optimal solution. If the dual solution is infeasible, any dual constraint that is violated provides a candidate aperture that may be added to the restricted version of (AM₁). Using the common terminology from the simplex method and column generation for linear programming problems, we say that these apertures price out, or are profitable, in the sense that adding them to the restricted problem will (likely) yield an improved fluence map. We can reoptimize the new restricted problem, and repeat the procedure. Below, we formally derive an approach that either verifies that a particular fluence map is optimal or provides one or more profitable apertures.

3.2. The pricing problem. The KKT-conditions (see, e.g., Bazaraa, Sherali, and Shetty [6] and Hiriart-Urruty and Lemarechal [17]) for optimality of (AM₁) are

$$(3.1) \quad \pi_{js} \in \partial F_{js}(z_{js}), \quad j = 1, \dots, v_s, \quad s = 1, \dots, S,$$

$$(3.2) \quad \rho_k = \sum_{s=1}^S \sum_{j=1}^{v_s} \gamma_k \left(\sum_{i \in A_k} D_{ijs} \right) \pi_{js}, \quad k \in K,$$

$$(3.3) \quad y_k \geq 0, \quad k \in K,$$

$$(3.4) \quad z_{js} = \sum_{k \in K} \gamma_k \left(\sum_{i \in A_k} D_{ijs} \right) y_k, \quad j = 1, \dots, v_s, \quad s = 1, \dots, S,$$

$$(3.5) \quad \rho_k y_k = 0, \quad k \in K,$$

$$(3.6) \quad \rho_k \geq 0, \quad k \in K,$$

where π_{js} is the Lagrange multiplier corresponding to the dose definition constraint (2.1) for voxel j in structure s , and ρ_k is the Lagrange multiplier corresponding to the nonnegativity constraint (2.2) of the intensity of aperture k . Now suppose that we have a solution to the aperture modulation problem with only a limited number of apertures included, say $\hat{K} \subseteq K$. That is, we have a solution $(\hat{z}, \hat{y})$ where the elements of $\hat{y}$ corresponding to omitted apertures have been set to zero and the remaining values are optimal given these constraints. We then choose the corresponding values of $\hat{\pi}_{js}$ according to KKT condition (3.1):

$$(3.7) \quad \hat{\pi}_{js} \in \partial F_{js}(\hat{z}_{js}), \quad j = 1, \dots, v_s, \quad s = 1, \dots, S;$$

i.e., $\hat{\pi}_{js}$ is a subgradient of the penalty function for voxel j in structure s at the current dose. Since the solution is optimal to the restricted problem, the pair $((\hat{z}, \hat{y}), (\hat{\pi}, \hat{\rho}))$ (with $\hat{\rho}$ defined by (3.2)) satisfies the KKT-conditions for the restricted formulation, i.e., it satisfies conditions (3.1) and (3.4), and it satisfies conditions (3.2), (3.3), (3.5), and (3.6) for $k \in \hat{K}$. The pair is optimal for the complete problem if and only if the remainder of the KKT-conditions are satisfied as well. Since for all $k \notin \hat{K}$ we have that $\hat{y}_k = 0$, conditions (3.3) and (3.5) are satisfied for all $k \in K$. It thus remains to verify whether conditions (3.2) and (3.6) are satisfied. These conditions can be summarized as

$$\sum_{s=1}^S \sum_{j=1}^{v_s} \gamma_k \left(\sum_{i \in A_k} D_{ijs} \right) \hat{\pi}_{js} \geq 0, \quad k \notin \hat{K},$$

or, equivalently,

$$\min_{k \in K} \sum_{s=1}^S \sum_{j=1}^{v_s} \left(\sum_{i \in A_k} D_{ijs} \right) \hat{\pi}_{js} \geq 0.$$

As noted above, the condition is satisfied for all $k \in \hat{K}$, which allows us to extend the range over which is minimized to K . Also note that the intensity correction factors γ_k disappear from the problem.

Representing an aperture by a binary vector w , where $w_i = 1$ if beamlet i is in the aperture and $w_i = 0$ otherwise, we wish to check that

$$\min_{w \in W} \sum_{s=1}^S \sum_{j=1}^{v_s} \left(\sum_{i \in N} D_{ijs} w_i \right) \hat{\pi}_{js} \geq 0,$$

where we optimize over the set W of all binary vectors w representing an allowable aperture. Finally, regrouping terms we may rewrite this as

$$(3.8) \quad \min_{w \in W} \sum_{i \in N} \left(\sum_{s=1}^S \sum_{j=1}^{v_s} D_{ijs} \hat{\pi}_{js} \right) w_i \geq 0.$$

If this condition is satisfied, the current solution is optimal. Otherwise, any vector w that violates this condition corresponds to an aperture that is likely to improve the fluence map when added to the restricted version of (AM₁). Clearly, the condition

can be verified by solving the optimization problem on the left-hand side of (3.8) to optimality.

When applied to large-scale linear programming problems, the latter optimization subproblem, which either verifies that the current solution is optimal or provides a previously omitted variable that would improve the solution when added to the problem, is called the *pricing problem*, since it determines the reduced cost of each (omitted) primal variable (or, equivalently, the shadow price of the corresponding nonnegativity constraint). The intuition behind the pricing problem for the aperture modulation problem is that, for each bixel, we compute the net unit effect of increasing its intensity by some small value. If the aggregate of these effects is nonnegative for all possible apertures, the current solution cannot be improved and is optimal. Otherwise, any aperture with negative aggregate effect can be added to the set of apertures and improve the solution.

As a final remark, note that (3.7) may allow for a choice of values for $\hat{\pi}_{js}$ when the set of subgradients of the penalty function for voxel j in structure s at the current dose contains more than a singleton. Now recall that the pricing problem can conclude that the current solution is optimal if the optimal solution value to the pricing problem is nonnegative. Thus, due to the convexity of the functions F_{js} the best choice for $\hat{\pi}_{js}$ will be

$$\hat{\pi}_{js} = (F_{js})'_+(\hat{z}_{js});$$

i.e., $\hat{\pi}_{js}$ is the right derivative of the penalty function for voxel j in structure s at the current dose.

3.3. Solving the pricing problem. The difficulty of the pricing problem depends critically on the structure of the feasible region W . First, note that we will allow only apertures that open beamlets at a single beam angle. This means that we in fact obtain a smaller scale pricing problem for each beam $\ell \in B$:

$$\text{minimize } \sum_{i \in N_\ell} \left(\sum_{s=1}^S \sum_{j=1}^{v_s} D_{ijs} \hat{\pi}_{js} \right) w_i$$

subject to

$$w \in W_\ell,$$

where W_ℓ represents the set of apertures that are allowed for beam ℓ . Suppose first that any configuration of beamlets in a beam forms an allowable aperture. In that case, the feasible region W_ℓ reduces to $W_\ell = \{0, 1\}^{|N_\ell|}$ and the pricing problem reduces to

$$\text{minimize } \sum_{r=1}^m \sum_{c=1}^n \left(\sum_{s=1}^S \sum_{j=1}^{v_s} D_{(\ell,r,c)js} \hat{\pi}_{js} \right) w_{(\ell,r,c)}$$

subject to

$$w_{(\ell,r,c)} \in \{0, 1\}, \quad r = 1, \dots, m, \quad c = 1, \dots, n.$$

The pricing problem then decomposes by beamlet, and can trivially be solved by, for each beam, simply choosing all beamlets for which the coefficient in the pricing

problem is negative, i.e.,

$$w_{(\ell,r,c)}^* = \begin{cases} 1 & \text{if } \sum_{s=1}^S \sum_{j=1}^{v_s} D_{(\ell,r,c)js} \hat{\pi}_{js} < 0, \\ 0 & \text{otherwise.} \end{cases}$$

The running time for solving a pricing problem for a given beam is then $O(mn)$. However, apertures thus obtained may not be deliverable using an MLC, and a so-called sequencing phase is needed to decompose the apertures and intensities into truly deliverable apertures and intensities.

In the next three sections, we develop algorithms to solve the pricing problems for three nested sets of constraints on deliverable apertures. These constraints are common in one or more of the commercially available MLC systems. For all systems, the physical limitations of the MLC (in particular, the leaf structure that is used to form the apertures) imply that, in each row of the beamlet grid corresponding to a particular beam, the beamlets that are exposed should be *consecutive*. Figure 3.1 shows examples of three different apertures that satisfy this constraint. The first system that we will consider is an MLC system where this is the only requirement that apertures should satisfy. Such a system is thus able to deliver all three apertures illustrated in Figure 3.1. In some systems, overlapping of the left and right leaves of adjacent beamlet rows, often called *interdigitation*, is not allowed. The constraints disallowing interdigitation are sometimes also called interleaf motion constraints or leaf collision constraints. The second system that we consider disallows interdigitation, but allows the leaves to be closed in the middle of the beam to prevent leaf collisions. Such a system can thus deliver apertures such as the ones in Figures 3.1(a) and (b), but not an aperture as the one in Figure 3.1(c) (where in that figure the leaves violating the interdigitation constraint are indicated by the dashed oval). Finally, the third system that we consider does not allow the leaves to be closed in the middle of an aperture; i.e., it requires the aperture to be formed by a *connected* set of beamlets. This system can only deliver apertures of the type illustrated in Figure 3.1(a).

3.3.1. Allowing interdigitation. The most widely used commercial MLC system is a system that allows interdigitation. In the simple approach discussed above the pricing problems decompose by beamlet. However, if we take into account that an aperture is formed by pairs of left and right leaves for each beamlet row, this decomposition is not valid anymore. We can reformulate the pricing problem for beam ℓ by letting $c_1(r)$ and $c_2(r)$ denote the index of the last beamlet that is blocked by the left leaf and the first beamlet that is blocked by the right leaf in row r of beam ℓ , respectively. When interdigitation is allowed, the pricing problem becomes

$$\text{minimize } \sum_{r=1}^m \sum_{c=c_1(r)+1}^{c_2(r)-1} \left(\sum_{s=1}^S \sum_{j=1}^{v_s} D_{(\ell,r,c)js} \hat{\pi}_{js} \right)$$

subject to

$$\begin{aligned} c_1(r) &< c_2(r), & r &= 1, \dots, m, \\ c_1(r) &\in \{0, \dots, n\}, & r &= 1, \dots, m, \\ c_2(r) &\in \{1, \dots, n+1\}, & r &= 1, \dots, m. \end{aligned}$$

It immediately follows that the pricing problem for beam $\ell \in B$ in this case decomposes by beamlet row. In particular, for beamlet row r in beam ℓ we need to solve

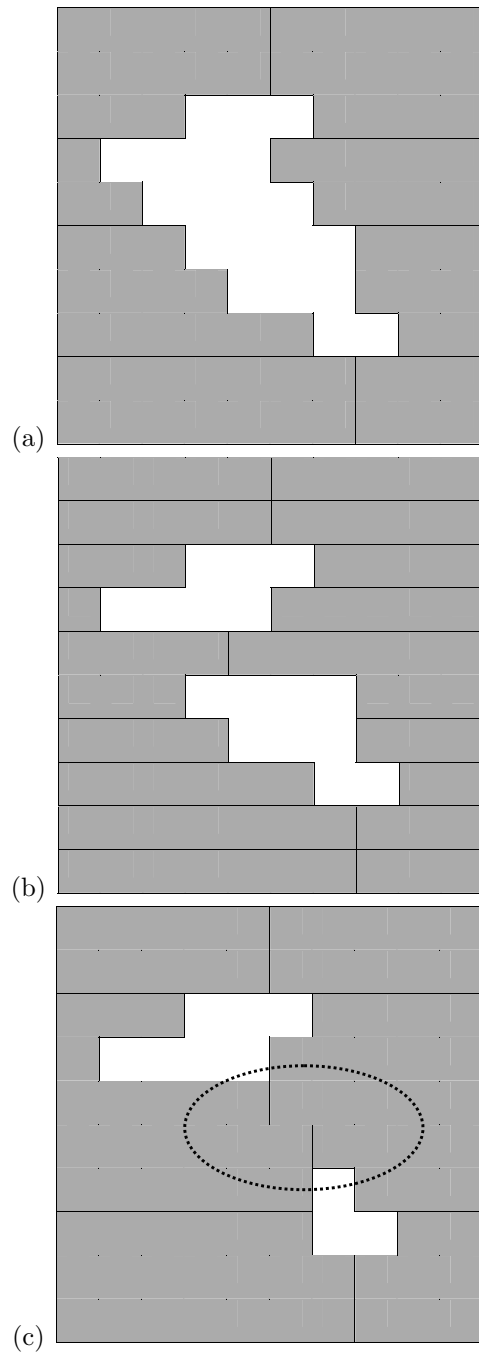


FIG. 3.1. *Three types of apertures. Aperture (a) can be delivered by all three considered systems. Aperture (b) can be delivered by systems that allow disconnected apertures. Aperture (c) can only be delivered by systems allowing interdigitation.*

the problem

$$\begin{aligned} & \text{minimize} \quad \sum_{c=c_1+1}^{c_2-1} \left(\sum_{s=1}^S \sum_{j=1}^{v_s} D_{(\ell,r,c)js} \hat{\pi}_{js} \right) \\ & \text{subject to} \end{aligned} \tag{P}$$

$$\begin{aligned} c_1 &< c_2, \\ c_1 &\in \{0, \dots, n\}, \\ c_2 &\in \{1, \dots, n+1\}. \end{aligned}$$

If the optimal value of this problem is zero, all bixels in the beamlet row are blocked. Otherwise, denoting the optimal solution to the problem by (c_1^*, c_2^*) , beamlets $1, \dots, c_1^*$ in beamlet row r in beam ℓ are blocked by the left leaf, and beamlets $c_2^*, \dots, n$ in the same beamlet row are blocked by the right leaf. For beam ℓ , the collection of beamlets chosen for each row in that beam forms an aperture for that beam. Solving the pricing problem in a straightforward way for each beamlet row using simple enumeration takes $O(n^2)$ time. We can thus find a new aperture that prices out (or determine that none exists at this stage) in $O(mn^2)$ time per beam.

However, we next present an alternative approach that reduces the computational complexity of the pricing problem. Note that we are looking for a consecutive set of beamlets in a given row for which the sum of their values is minimal. We can find such a set of beamlets by passing through the n beamlets in a given row and beam from left to right only once, keeping track of (i) the cumulative value over all beamlets considered so far (v) and (ii) the maximum cumulative value found so far ($\bar{v}$). Now note that, at any point, the difference $v - \bar{v}$ between these two is a candidate for the best solution value found so far. The algorithm is more formally given by:

- Step 0.** Set $v = \bar{v} = v^* = 0$, $c_1^* = c_1 = 0$, $c_2^* = c_2 = 1$.
- Step 1.** Set $v = v + \sum_{s=1}^S \sum_{j=1}^{v_s} D_{(\ell,r,c_2)js} \hat{\pi}_{js}$.
- Step 2.** If $v \geq \bar{v}$, set $\bar{v} = v$ and set $c_1 = c_2$. If $v - \bar{v} < v^*$, set $v^* = v - \bar{v}$, $c_1^* = c_1$, $c_2^* = c_2 + 1$.
- Step 3.** If $c_2 < n$, increment c_2 and return to Step 1. Otherwise, stop; c_1^* and c_2^* give the optimal aperture for row r in beam ℓ , and v^* the corresponding reduced cost.

Bates and Constable [5] and Bentley [8] have shown that this algorithm is correct, and that its running time is linear in the number of beamlets in a row. We conclude that we can find a new aperture that prices out (or determine that none exists at this stage) in $O(mn)$ time per beam when interdigitation is allowed.

3.3.2. Disallowing interdigitation. Some MLC systems do not allow interdigitation; that is, the left leaf in a row cannot overlap the right leaf of an adjacent row. This additional constraint causes a dependency between the rows in a beam that prevents the pricing problem from decomposing by beamlet row. In this section, we will formulate a network flow problem that efficiently solves the pricing problem for a given beam in case interdigitation is not allowed. Our network flow formulation is related to a network flow formulation of the problem of decomposing a given fluence map into apertures that was introduced by Boland, Hamacher, and Lenzen [10].

In this network, we create a node for each potential pair of left and right leaf settings in each beamlet row. For a fixed beam $\ell \in B$, this means that we have nodes of the form (r, c_1, c_2) , where, as in section 3.3.1, r ($r = 1, \dots, m$) indicates the beamlet

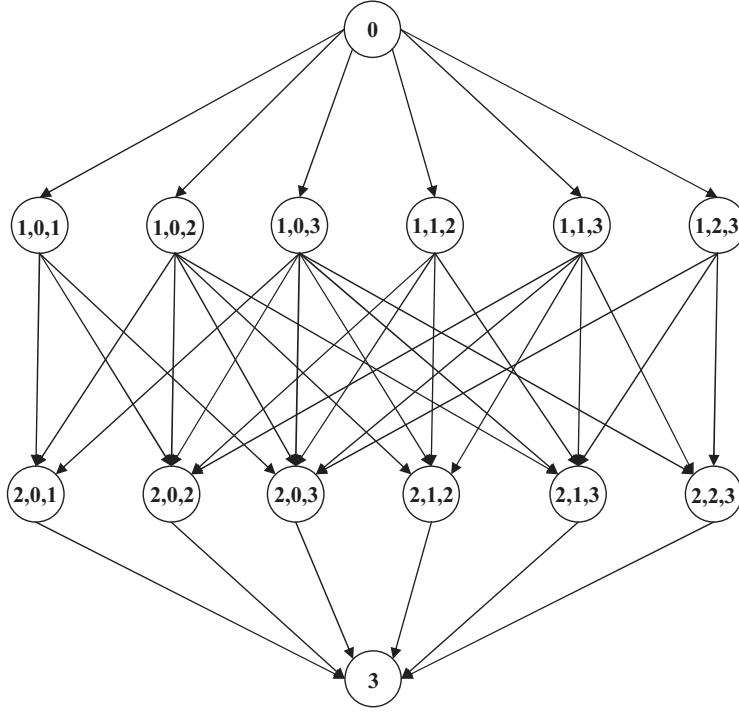


FIG. 3.2. Network for the pricing problem in case interdigitation is not allowed.

row, c_1 ($c_1 = 0, \dots, n$) indicates the last beamlet covered by the left leaf in row r , c_2 ($c_2 = 1, \dots, n+1$) indicates the first beamlet covered by the right leaf in row r , and $c_1 < c_2$. In addition, we define a source node, say 0, and a sink node, say $m+1$. The total number of nodes in the network is $O(mn^2)$.

We next create arcs between nodes in consecutive beamlet rows if the two rows may appear in a single aperture. Note that this will allow us to prevent interdigitation, by simply not including arcs that would cause interdigitation. For $r = 1, \dots, m-1$, this means that we create a directed arc from node (r, c_1, c_2) to node $(r+1, c'_1, c'_2)$ only when $c'_1 < c_2$ and $c_1 < c'_2$, which ensures that no leaf collisions take place. In addition, we create directed arcs from the source node to all nodes for row $r = 1$, and also from all nodes for row $r = m$ to the sink node. The total number of arcs in the network is $O(mn^4)$.

Figure 3.2 illustrates the structure of the network for a small case with $n = 2$ rows and $m = 2$ columns. This network provides a one-to-one correspondence between paths from the source 0 to the sink $m+1$ and apertures that satisfy the interdigitation constraints. If we now let all arcs leading to a node of the form (r, c_1, c_2) have cost

$$\sum_{c=c_1+1}^{c_2-1} \sum_{s=1}^S \sum_{j=1}^{v_s} D_{(\ell, r, c)js} \hat{\pi}_{js}$$

and all arcs to the sink have cost 0, the cost of each path is equal to the cost of the corresponding aperture in the pricing problem. The candidate aperture for beam ℓ can now be found by solving for the minimum cost path from the source to the sink in the network. Since the network is acyclic, the shortest path problem can be solved

in time proportional to the number of arcs in the network (see, e.g., Ahuja, Magnanti, and Orlin [2]). For each beam ℓ , we can thus find a new aperture that prices out (or determine that none exists at this stage) in $O(mn^4)$ time.

3.3.3. Requiring connected apertures. The previous section solves the pricing problem when apertures need to satisfy interdigitation constraints but are allowed to consist of a disconnected set of beamlets. However, there exist systems that cannot deliver such apertures. For such systems, it is not allowed that leaves close in the middle of the aperture. This, in particular, means that the beamlet rows that do not cover all beamlets should be consecutive. This additional constraint can be incorporated into the network given in section 3.3.2 with slight modifications.

The first modification is that we represent only beamlet rows that expose at least one beamlet by nodes in the network. Put differently, we remove all nodes of the form $(r, c_1, c_1 + 1)$ from the network described in section 3.3.2. Then, to represent the fact that the first and last row(s) in an aperture may be closed, we create arcs from the source 0 to all nodes, as well as from each node to the sink $m + 1$. It is easy to see that each path in this network corresponds to a connected aperture. Moreover, if we again let all arcs leading to node (r, c_1, c_2) have cost

$$\sum_{c=c_1+1}^{c_2-1} \sum_{s=1}^S \sum_{j=1}^{v_s} D_{(\ell, r, c)js} \hat{\pi}_{js}$$

and all remaining arcs have cost 0, the cost of each path is equal to the cost of the corresponding aperture in the pricing problem.

As the network in section 3.3.2, the modified network contains $O(mn^2)$ nodes and $O(mn^4)$ arcs. For each beam ℓ , we can thus find a new aperture that prices out (or determine that none exists at this stage) in $O(mn^4)$ time by solving for the minimum cost path from the source to the sink in this network.

4. Extending the model with partial volume constraints. A tool that is commonly used by physicians to judge the quality of a treatment plan is the so-called (cumulative) *dose-volume histogram (DVH)*. This histogram specifies, for a given target or critical structure, the fraction of its volume that receives at least a certain amount of dose. Many successful optimization models for FMO include some type of constraints (often called DVH constraints or partial volume constraints) on the shape and location of the DVH of a structure. In this section, we will extend our basic model with a class of partial volume constraints.

4.1. Modeling partial volume constraints. As in Romeijn et al. [31], we will employ the following type of partial volume constraints:

(P₁) The *average* dose received by the subset of a target of relative volume $1 - \alpha$ receiving the *lowest* amount of dose must be at least equal to L^α .

(P₂) The *average* dose received by the subset of a structure of relative volume $1 - \alpha$ receiving the *highest* amount of dose may be no more than U^α .

If, for a given structure s , αv_s is integral, these partial volume constraints can be reformulated as bounds on the sum of the $(1 - \alpha)v_s$ largest or smallest voxel doses in a structure. This concept has been applied to facility location models by Andreatta and Mason [4], who proposed choosing a facility location for which the sum of the k largest distances between customers and facility, called the k -eccentricity, is minimized. They showed that the k -eccentricity is a convex function and proposed algorithms for solving the corresponding location problems on tree graphs (see also Slater [41], who studied

similar problems in a more abstract graph theoretical setting). More recently, Tamir [43] and Ogryczak and Zawadzki [27] have developed efficient optimization approaches for various location models employing such an objective. Furthermore, Ogryczak and Tamir [26] have developed a more general approach for efficiently minimizing the sum of the k largest of n function values.

When formulating partial volume constraints for the FMO problem, care must be taken when αv_s is nonintegral. In this case, we could round up or down the number of voxels that we are interested in and apply the methods referred to above. However, this could have undesirable round-off effects, in particular for smaller structures. This is due to the fact that although we model the FMO problem on a discretization of the patient using a finite number of voxels, the underlying structures are continuous objects. Therefore, we have instead chosen to use the related concept of tail-averages, which applies to both continuous and discrete populations. This concept has in recent years received much attention in the risk management literature (see Rockafellar and Uryasev [29, 30] and Ogryczak and Ruszczyński [25]), and is often referred to as Conditional Value-at-Risk (CVaR) or Tail Value-at-Risk (TVaR). Applying these concepts to discrete distributions allows for a correction for nonintegral values of αv_s by including a fractional voxel in the tail average, and can as such be expected to much better approximate the continuous tail-average than a rounding approach would. In Romeijn et al. [31, 32] we have demonstrated the power of such constraints in the context of FMO. In addition, in contrast with other types of constraints on the DVH proposed in the literature (see, e.g., Langer et al. [21], Bortfeld, Stein, and Preiser [13], Carol et al. [14], Wu and Mohan [49], Lee, Fox, and Crocker [22, 23]), these constraints can be incorporated into a convex programming formulation of the problem.

Constraints of the type (P₁) can be formulated as lower bound constraints on the so-called lower CVaR at level α (for short, the lower α -CVaR). Rockafellar and Uryasev [29] and Ogryczak and Ruszczyński [25] show that the lower α -CVaR can be expressed as

$$\underline{\phi}_s^\alpha = \max_{\zeta \in \mathbb{R}} \left\{ \zeta - \frac{1}{(1-\alpha)v_s} \sum_{j=1}^{v_s} (\zeta - z_{js})^+ \right\}.$$

Similarly, constraints of the type (P₂) can be formulated as upper bounds on the upper α -CVaR, which is equal to

$$\overline{\phi}_s^\alpha = \min_{\zeta \in \mathbb{R}} \left\{ \zeta + \frac{1}{(1-\alpha)v_s} \sum_{j=1}^{v_s} (z_{js} - \zeta)^+ \right\}.$$

For each structure, we may now define a set of bounds on low and high tail averages. For targets, these will be in the form of lower bounds L_s^α for $\alpha \in \underline{A}_s$, where $\underline{A}_s$ is a (finite) subset of $(0, 1)$, and upper bounds U_s^α for $\alpha \in \overline{A}_s$, where $\overline{A}_s$ is a (finite) subset of $(0, 1)$ ($s = 1, \dots, T$). For the critical structures, these will be in the form of only upper bounds U_s^α for $\alpha \in \overline{A}_s$, where $\overline{A}_s$ is a (finite) subset of $(0, 1)$ ($s = 1, \dots, S$).

4.2. An enhanced problem formulation. To ensure that a feasible solution to our enhanced model exists, we will not incorporate hard DVH constraints but instead allow them to be violated and incorporate a penalization of such violations into the objective function. If $SA^- \subset S \times (0, 1)$ denotes the set of lower CVaR constraints,

we let $\underline{G}_s^\alpha$ denote a convex and nonincreasing penalty function associated with the lower α -CVaR constraint for structure s . Similarly, if $SA^+ \subset S \times (0, 1)$ denotes the set of upper CVaR constraints, we let $\overline{G}_s^\alpha$ denote a convex and nondecreasing penalty function associated with the upper α -CVaR constraint for structure s . This leads to the following aperture modulation formulation of the FMO problem:

$$\text{minimize } \sum_{s=1}^S \sum_{j=1}^{v_s} F_{js}(z_{js}) + \sum_{(s,\alpha) \in SA^+} \overline{G}_s^\alpha(\overline{\phi}_s^\alpha) + \sum_{(s,\alpha) \in SA^-} \underline{G}_s^\alpha(\underline{\phi}_s^\alpha)$$

subject to (AM₂)

$$\sum_{\ell \in B} \sum_{k \in K_\ell} \left(\sum_{i \in A_k} D_{ijs} \right) y_k = z_{js}, \quad j = 1, \dots, v_s, \quad s = 1, \dots, S,$$

$$(4.1) \quad \overline{\zeta}_s^\alpha + \frac{1}{(1-\alpha)v_s} \sum_{j=1}^{v_s} (z_{js} - \overline{\zeta}_s^\alpha)^+ \leq \overline{\phi}_s^\alpha, \quad (s, \alpha) \in SA^+,$$

$$(4.2) \quad \underline{\zeta}_s^\alpha - \frac{1}{(1-\alpha)v_s} \sum_{j=1}^{v_s} (\underline{\zeta}_s^\alpha - z_{js})^+ \geq \underline{\phi}_s^\alpha, \quad (s, \alpha) \in SA^-,$$

$$y_k \geq 0, \quad k \in K.$$

Note that it is easy to see that there will exist an optimal solution to (AM₂) for which constraints (4.1) and (4.2) are binding.

4.3. The pricing problem. The KKT-conditions for optimality of (AM₂) are

$$(4.3) \quad \begin{aligned} \pi_{js} &\in \partial F_{js}(z_{js}) \\ &+ \sum_{(s,\alpha) \in SA^+ : j=1, \dots, v_s, z_{js} \geq \overline{\zeta}_s^\alpha} \overline{\delta}_s^\alpha + \sum_{(s,\alpha) \in SA^- : j=1, \dots, v_s, z_{js} \leq \underline{\zeta}_s^\alpha} \underline{\delta}_s^\alpha, \\ & \quad j = 1, \dots, v_s, \quad s = 1, \dots, S, \end{aligned}$$

$$(4.4) \quad \overline{\delta}_s^\alpha \in \partial \overline{G}_s^\alpha(\overline{\phi}_s^\alpha), \quad (s, \alpha) \in SA^+,$$

$$(4.5) \quad \underline{\delta}_s^\alpha \in \partial \underline{G}_s^\alpha(\underline{\phi}_s^\alpha), \quad (s, \alpha) \in SA^-,$$

$$\rho_k = \sum_{s=1}^S \sum_{j=1}^{v_s} \left(\sum_{i \in A_k} D_{ijs} \right) \pi_{js}, \quad k \in K,$$

$$y_k \geq 0, \quad k \in K,$$

$$z_{js} = \sum_{k \in K} \left(\sum_{i \in A_k} D_{ijs} \right) y_k, \quad j = 1, \dots, v_s, \quad s = 1, \dots, S,$$

$$\overline{\phi}_s^\alpha \geq \overline{\zeta}_s^\alpha + \frac{1}{(1-\alpha)v_s} \sum_{j=1}^{v_s} (z_{js} - \overline{\zeta}_s^\alpha)^+, \quad (s, \alpha) \in SA^+,$$

$$\underline{\phi}_s^\alpha \leq \underline{\zeta}_s^\alpha - \frac{1}{(1-\alpha)v_s} \sum_{j=1}^{v_s} (\underline{\zeta}_s^\alpha - z_{js})^+, \quad (s, \alpha) \in SA^-,$$

$$\rho_k y_k = 0, \quad k \in K,$$

$$(4.6) \quad \begin{aligned} \rho_k &\geq 0, & k &\in K, \\ \overline{\delta}_s^\alpha &\geq 0, & (s, \alpha) &\in SA^+, \end{aligned}$$

$$(4.7) \quad \underline{\delta}_s^\alpha \leq 0, \quad (s, \alpha) \in SA^-,$$

where π_{js} and ρ_k are defined as in the basic model, and $\bar{\delta}_s^\alpha$ and $\underline{\delta}_s^\alpha$ are the Lagrange multipliers of the CVaR constraints (4.1) and (4.2). Note that the sign conditions (4.6) and (4.7) are automatically satisfied by the monotonicity assumptions on the penalty functions $\bar{G}_s^\alpha$ and $\underline{G}_s^\alpha$, respectively. Then proceeding as in the basic model, given a solution $(\hat{z}, \hat{y})$ to the problem using a limited set of apertures $\hat{K} \subseteq K$, we use (4.3)–(4.5) to determine corresponding values of $\hat{\pi}$. As in the basic model, if any set of subgradients contains more than one element, the right derivative (i.e., the largest value in the set of subgradients) is chosen. With these choices, we obtain a pricing problem of exactly the same form as in the basic model, which can then be solved using the same approaches.

5. Computational results. To study the effect of the number of apertures used on the quality of the treatment plans obtained, we have studied cases of head-and-neck cancer patients where the primary reason for using IMRT is to preserve salivary function while obtaining adequate target coverage. Three-dimensional treatment planning data for ten head-and-neck cancer patients was exported from a commercial patient imaging and anatomy segmentation system (VoxelQ, Philips Medical Systems) at the University of Florida Department of Radiation Oncology. This data was then imported into the University of Florida Optimized Radiation Therapy (UFORT) treatment planning system and used to generate the data required by the models described in sections 2 and 4.

Our treatment planning system generated a voxel grid with voxels of size $4 \times 4 \times 4 \text{ mm}^3$ for the targets and critical structures except for unspecified tissue, for which we used voxels of size $8 \times 8 \times 8 \text{ mm}^3$. Table 5.1 shows the problem dimensions for the ten cases. As mentioned in section 1.4, there is no fundamental basis for quantifying the quality of a treatment plan. Therefore, to assess the ability of our model to generate high-quality treatment plans, we have evaluated our treatment plans using a set of external planning criteria that were not included in our model. Note that in order to be able to make an objective assessment of model quality, it is important that these external planning criteria were not included in our model, to avoid bias of the conclusions. In particular, we have used the criteria published in two studies of the Radiation Therapy Oncology Group (RTOG) [34, 35] with minor modifications (where the modifications made the criteria more demanding). The main goal was to treat two targets: Planning Target Volume 1 (PTV1) and Planning Target Volume 2 (PTV2), which contains PTV1. PTV1 consists of actual identified disease, while PTV2 is an expansion of PTV1 that also contains regions where the disease may have spread, as well as a margin to account for the possibility of patient motion and setup

TABLE 5.1
Model sizes.

Case	# beamlets	# voxels
1	1,017	17,471
2	1,195	19,543
3	1,745	36,136
4	1,804	38,724
5	1,182	17,466
6	1,015	14,147
7	1,178	22,300
8	1,090	19,417
9	921	15,416
10	2,073	30,664

uncertainty. PTV1 should be treated to a prescription dose level of $D_{R_x}^{\text{PTV1}} = 73.8$ Gray (Gy), and PTV2 should be treated to a prescription dose level of $D_{R_x}^{\text{PTV2}} = 54$ Gy. More precisely, this means that at least 95% of both targets should receive the prescribed dose, no more than 1% of both targets should be underdosed by more than 7%, and no more than 20% of PTV1 should be overdosed by more than 10%. Each of the up to four salivary glands (LPG: left parotid gland, RPG: right parotid gland, LSG: left submandibular gland, and RSG: right submandibular gland) is considered spared if either no more than 50% of the voxels in the gland receive a dose exceeding 30 Gy, or the mean dose to the gland does not exceed 26 Gy. No voxels in the spinal cord (SC) should receive a dose exceeding 45 Gy, and no voxels in the brainstem (BS) should receive a dose exceeding 54 Gy. Less than 1% of the voxels in unspecified tissue should receive more than 65 Gy to avoid side effects of treatment due to an unacceptable level of overdosing of that structure. For all ten cases, we designed plans using seven equispaced beams at nominal settings; i.e., their orientation was not optimized for the individual patients.

The structure-dependent penalty functions were chosen to be piecewise-linear approximations to piecewise-polynomial functions of the following form:

$$F_{js}(z) = \frac{1}{v_s} \left(\beta_s^U \max(0, T_s - z)^{p_s^U} + \beta_s^O \max(0, z - T_s)^{p_s^O} \right),$$

where the superscript ^{U} refers to underdosing and the superscript ^{O} refers to overdosing. The coefficients $\beta_s^U, \beta_s^O \geq 0$ and the powers $p_s^U, p_s^O \geq 1$ to ensure convexity of the function. Finally, the parameters T_s represent thresholds, to allow the shape of the penalty functions to be different for underdosing and overdosing. We employed piecewise-linear penalty functions in order to allow us to solve our models using a linear programming solver. The piecewise-linear functions used two segments for underdosing and four segments for overdosing for targets, and two segments for overdosing for critical structures. In addition, we defined lower bounds L_s and upper bounds U_s on the dose for each structure, with very steep slopes (in the range 10^8 – 10^{12}) in the piecewise-linear functions for dose values below the lower bound and above the upper bound.

Based on three of the ten patient cases and the beamlet-based model for FMO, we tuned the problem parameters by manual adjustment. This resulted in the parameter values shown in Tables 5.2 and 5.3. Similar to Tsien et al. [46], we found that high powers of dose difference lead to excellent results. We next solved the beamlet formulation of the FMO problem with partial volume constraints for all ten patient cases using this set of parameters to obtain benchmarks.

Next, we solved all ten problem instances using the column generation approach to the aperture modulation formulation of the FMO problem as developed in this paper. Since the most commonly used MLC system allows for interdigitation, we have used the pricing algorithm developed in section 3.3.1. Recall that when the column generation algorithm is run until convergence, the optimal beamlet-based FMO solution is obtained. To determine the number of apertures required by our approach to obtain a high-quality treatment plan, we examined the solution obtained by the column generation algorithm in each iteration. We then evaluated which of the external planning criteria were satisfied for each intermediate solution. The results for each of the ten cases are summarized in Table 5.4. In particular, this table provides the number of used apertures in our solution that ensured that each of the external planning criteria was satisfied. Note that we have added several criteria that are

TABLE 5.2
Values of the coefficients of the voxel-based penalty functions.

s	T_s	L_s	β_s^U	p_s^U	U_s	β_s^O	p_s^O
PTV1	$D_{R_x}^{\text{PTV1}} + 2.5$	$D_{R_x}^{\text{PTV1}} - 0.5$	7,500	12	$D_{R_x}^{\text{PTV1}} + 5.5$	7,500	6
PTV2	$D_{R_x}^{\text{PTV2}} + 2$	$0.93D_{R_x}^{\text{PTV2}}$	75,000	12	$D_{R_x}^{\text{PTV1}} + 5.5$	75,000	6
Tissue	35	0	0	–	$D_{R_x}^{\text{PTV1}} + 5.5$	8,000,000	5
LPG	0	0	0	–	$D_{R_x}^{\text{PTV1}} + 5.5$	350,000	4
RPG	0	0	0	–	$D_{R_x}^{\text{PTV1}} + 5.5$	350,000	4
LSG	0	0	0	–	$D_{R_x}^{\text{PTV1}} + 5.5$	150,000	4
RSG	0	0	0	–	$D_{R_x}^{\text{PTV1}} + 5.5$	150,000	4
SC	28	0	0	–	40	150	3
BS	33.6	0	0	–	48	200	3

TABLE 5.3
Values of the coefficients for the CVaR constraints.

	Lower CVaR-constraints		Upper CVaR-constraints	
s	α	L_s^α	α	U_s^α
PTV1	0.90	$0.97D_{R_x}^{\text{PTV1}}$	0.99	$1.07D_{R_x}^{\text{PTV1}}$
PTV2	0.90	$0.96D_{R_x}^{\text{PTV2}}$	0.95	$1.1D_{R_x}^{\text{PTV2}}$
LPG	–	–	0.60	26
RPG	–	–	0.60	26
LSG	–	–	0.60	26
RSG	–	–	0.60	26

TABLE 5.4
Number of (used) apertures needed to satisfy each criterion.

		Case									
s	Criterion	1	2	3	4	5	6	7	8	9	10
PTV1	$99\% \geq 68.6 \text{ Gy}$	9	7	10	10	10	6	10	12	12	12
	$80\% \leq 81.2 \text{ Gy}$	16	19	17	17	18	17	19	17	18	23
	$90\% \leq 81.2 \text{ Gy}$	47	55	55	48	47	46	51	50	49	59
	$95\% \leq 81.2 \text{ Gy}$	111	126	130	102	108	111	108	116	119	118
PTV2	$95\% \geq 54 \text{ Gy}$	★	48	56	43	38	84	209	85	175	43
	$99\% \geq 50.2 \text{ Gy}$	117	45	18	30	88	65	55	19	77	21
Tissue	$95\% \leq 50 \text{ Gy}$	51	27	76	31	35	92	26	43	14	105
	$99\% \leq 65 \text{ Gy}$	12	19	18	10	17	17	11	15	12	12
LPG	$50\% \leq 30 \text{ Gy}$	28	28	10	10	20	19	10	24	12	12
	mean $\leq 26 \text{ Gy}$	36	28	10	10	45	34	10	43	12	12
RPG	$50\% \leq 30 \text{ Gy}$	28	–	10	10	10	17	67	12	28	39
	mean $\leq 26 \text{ Gy}$	34	–	10	10	10	46	×	15	52	55
LSG	$50\% \leq 30 \text{ Gy}$	36	29	–	33	37	–	33	–	30	33
	mean $\leq 26 \text{ Gy}$	80	30	–	54	53	–	39	–	52	36
RSG	$50\% \leq 30 \text{ Gy}$	–	–	–	78	19	–	–	–	–	–
	mean $\leq 26 \text{ Gy}$	–	–	–	×	22	–	–	–	–	–
SC	$100\% \leq 45 \text{ Gy}$	73	30	37	44	26	80	34	40	83	39
BS	$100\% \leq 54 \text{ Gy}$	9	15	–	10	10	19	10	12	12	42
FMO	# apertures	166	140	194	192	170	157	168	185	160	157

TABLE 5.5
Regression results of the relationship between generated and used apertures.

	Case									
	1	2	3	4	5	6	7	8	9	10
θ_0	22.0	27.2	31.4	18.8	17.5	21.7	19.3	25.9	24.1	18.9
θ_1	0.55	0.60	0.61	0.54	0.58	0.53	0.59	0.57	0.56	0.61
R^2 (%)	99.0	98.0	97.9	99.2	99.3	98.3	99.7	98.7	98.5	99.1

stronger than the ones generally used in clinical practice, such as stronger overdosing criteria for PTV1 (no more than 5% or 10% (instead of 20%) of its volume may be overdosed by more than 10%) and unspecified tissue (no more than 5% of its volume may receive more than 50 Gy). In case 1, 95% coverage of PTV2 with the prescription dose of 54 Gy was not obtained (indicated by $\star$), even using the beamlet-based FMO formulation. In this case, 94% coverage was obtained using 108 apertures. In cases 4 and 7, the mean dose constraint could not be satisfied (indicated by $\times$). However, in these cases the partial volume constraint was satisfied using a limited number of apertures, ensuring sparing of these structures. Cases where a saliva gland was either not spareable (since it was largely contained in the target) or not in the vicinity of a target (and therefore not included in the problem) are indicated by $-$. Note that the ability of the models to obtain high-quality treatment plans (as measured by the external planning criteria) with a single parameter setting suggests that parameter tuning (or, equivalently, a multicriteria solution approach) is not needed for each individual patient.

The last line in the table corresponds to the number of apertures that were found using a traditional two-stage approach, where an intensity profile found by our beamlet-based FMO model was discretized and decomposed into apertures.

To compare the number of apertures that were generated to the number of apertures that were actually used, i.e., that received a positive intensity in the optimal solution given a particular number of apertures, we performed the following linear regression characterizing this relationship for all ten cases:

$$\text{Number of apertures used} = \theta_0 + \theta_1 \times \text{Number of apertures generated}.$$

Table 5.5 shows the estimates for the regression coefficients, as well as the goodness-of-fit. The positive intercept θ_0 represents that a particular number of apertures (apparently on the order of 2–4 per beam) should always be used for treatment. Out of all additional apertures generated, about 58%, on average, turn out to be useful in the long run, while the other ones are eventually replaced by one or more better apertures.

For one of the cases (case 5), Figures 5.1–5.4 illustrate the performance of our method in more detail. Figure 5.1(a) shows the behavior of the objective function value as a function of the number of apertures generated, as well as the number of apertures used (i.e., the number of apertures with nonzero intensity in the corresponding solution). Figure 5.1(b) shows the number of apertures used versus the number of apertures generated. Figures 5.2–5.4 show how the number of apertures used affects the level of satisfaction of the external criteria. For example, the curves in Figure 5.3(a) show the relative volume of each of the four saliva glands that receives in excess of 30 Gy as a function of the number of apertures used. Similarly, the curves in Figure 5.3(b) show the mean dose received by each of the four saliva glands as a function of the number of apertures used.

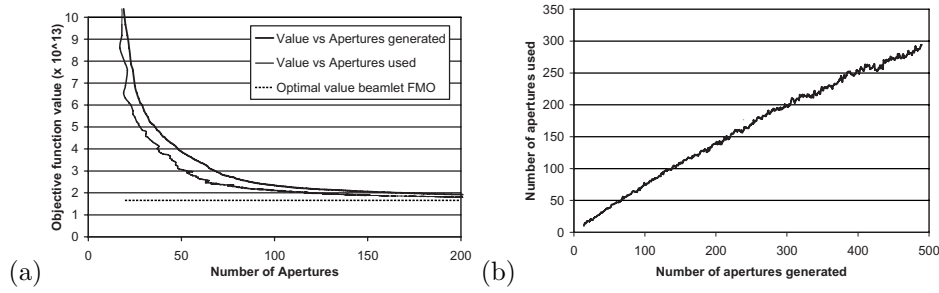


FIG. 5.1. (a) Objective function value as a function of number of apertures generated and used, and comparison with the optimal value to the beamlet FMO problem. (b) Number of apertures used as a function of number of apertures generated.

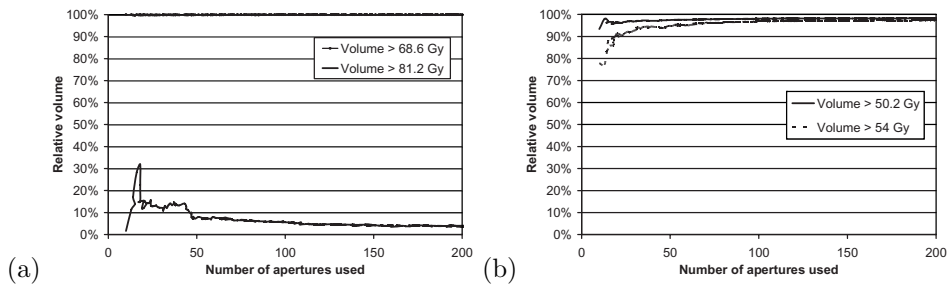


FIG. 5.2. Coverage of (a) PTV1 and (b) PTV2 as a function of number of apertures used.

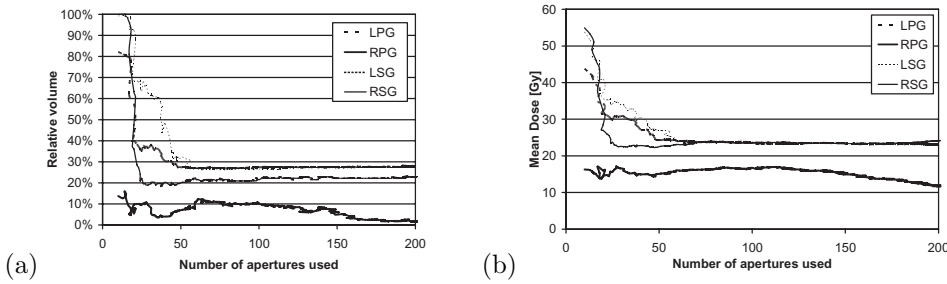


FIG. 5.3. Sparing of saliva glands according to (a) DVH criterion (relative volume > 30 Gy) and (b) mean dose as a function of number of apertures used.

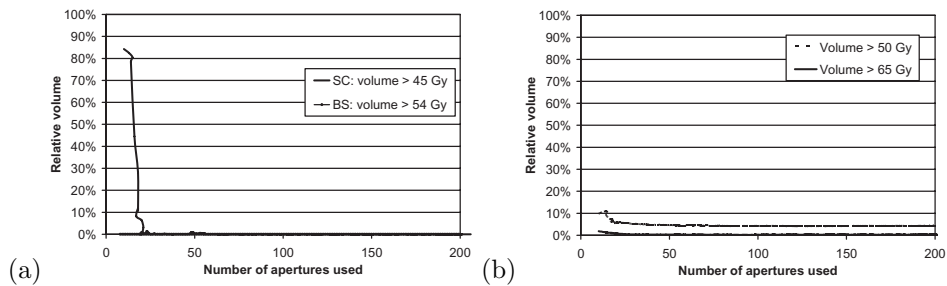


FIG. 5.4. Sparing of (a) spinal cord (volume > 45 Gy) and brainstem (volume > 54 Gy) and (b) unspecified tissue as a function of number of apertures used.

From our results, we conclude that all external planning criteria except for the criteria for PTV2 can be satisfied with, on average, 65 apertures (which corresponds to 9–10 apertures per beam). In clinical practice, physicians may decide to treat using these more efficient plans at the cost of lower coverage of PTV2. This trade-off between treatment plan efficiency and PTV2-coverage needs to be made on a case-by-case basis based on the extent of the disease and the individual level of clinical risk associated with underdosing PTV2. To obtain a good coverage and a smaller level of underdosing of PTV2, an average of 101 apertures is needed. In fact, in 7 out of the 10 cases fewer than 100 apertures (approximately 14 per beam) suffice. Even when the external planning criteria that are much stricter than the published RTOG criteria are taken into account, the average required number of apertures is only 131. Note that although it may take more apertures than this to achieve the same objective function value as the optimal value obtained using the beamlet model for FMO, the suboptimal solutions obtained using this number of apertures are of clinically acceptable quality. Most current commercial treatment planning systems typically require 20–30 apertures per beam, or a total of 140–210 apertures in order to achieve a clinically acceptable treatment plan, which corresponds to about 50%–100% more apertures than in our solutions. This is illustrated by the last line in Table 5.4, which shows that in all cases but one a traditional two-stage approach yields significantly more apertures than the aperture modulation approach. Our methodology thus suggests that patients can be treated in much less time than in current practice. In addition, and perhaps even more importantly, our approach will provide the physicians with a tool to make a sound trade-off between the number of apertures (i.e., treatment time) and the level of satisfaction of the treatment plan criteria.

6. Concluding remarks and future research. In this paper, we have developed a new approach to the design of radiation therapy treatment plans. Our approach incorporates delivery issues into the design phase that to date have usually been handled in a second optimization phase. Using 10 problem instances obtained from clinical data, we have shown that patient treatment times can be reduced while maintaining treatment plan quality by integrating these two optimization phases.

In our approach, we are currently using a simple technique for finding an initial solution for the column generation procedure. Future research will focus on developing new heuristic techniques for finding better initial solutions, in order to reduce the number of steps required by the column generation procedure.

Our model currently assumes that the beam orientations are fixed in advance. Another direction of our future research will incorporate this beam orientation selection problem into the optimization problem.

Acknowledgments. We would like to thank Dr. Jonathan G. Li of the Department of Radiation Oncology at the University of Florida for many helpful discussions regarding the clinical implementation of our models. We would also like to thank two anonymous referees for their insightful suggestions that significantly improved the presentation of this paper.

REFERENCES

- [1] R. AHUJA AND H. HAMACHER, *Linear Time Network Flow Algorithm to Minimize Beam-On-Time for Unconstrained Multileaf Collimator Problems in Cancer Radiation Therapy*, tech. report, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, 2003.
- [2] R. AHUJA, T. MAGNANTI, AND J. ORLIN, *Network Flows: Theory, Algorithms, and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

- [3] AMERICAN CANCER SOCIETY, *Cancer Facts & Figures*, report, <http://www.cancer.org/>, 2004.
- [4] G. ANDREATTA AND F. MASON, *k-Eccentricity and absolute k-centrum of a probabilistic tree*, European J. Oper. Res., 19 (1985), pp. 114–117.
- [5] J. BATES AND R. CONSTABLE, *Proofs as programs*, ACM Trans. Progr. Lang. Syst., 7 (1985), pp. 113–136.
- [6] M. BAZARAA, H. SHERALI, AND C. SHETTY, *Nonlinear Programming: Theory and Algorithms*, 2nd ed., John Wiley & Sons, New York, 1993.
- [7] G. BEDNARZ, D. MICHALSKI, C. HOUSER, M. HUQ, Y. XIAO, P. ANNE, AND J. GALVIN, *The use of mixed-integer programming for inverse treatment planning with pre-defined field segments*, Phys. Med. Biol., 47 (2002), pp. 2235–2245.
- [8] J. BENTLEY, *Programming Pearls*, Addison-Wesley, Reading, MA, 1986.
- [9] B. BJÄRNGÅRD AND R. SIDDON, *A note on equivalent circles, squares, and rectangles*, Med. Phys., 9 (1982), pp. 258–160.
- [10] N. BOLAND, H. HAMACHER, AND F. LENZEN, *Minimizing beam-on time in cancer radiation treatment using multileaf collimators*, Networks, 43 (2004), pp. 226–240.
- [11] T. BORTFELD, D. KÄHLER, T. WALDRON, AND A. BOYER, *X-ray field compensation with multileaf collimators*, Int. J. Radiat. Oncol. Biol. Phys., 28 (1994), pp. 723–370.
- [12] T. BORTFELD, K.-H. KÜFER, M. MONZ, A. SCHERRER, C. THIEKE, AND H. TRINKHAUS, *Intensity-Modulated Radiotherapy: A Large Scale Multi-criteria Programming Problem*, Berichte des Fraunhofer ITWM 43, Fraunhofer Insitut Techno- und Wirtschaftsmathematik, Kaiserslautern, Germany, 2003.
- [13] T. BORTFELD, J. STEIN, AND K. PREISER, *Clinically relevant intensity modulation optimization using physical criteria*, in Proceedings of the XIIth ICCR, Salt Lake City, Utah, 1997, pp. 1–4.
- [14] M. CAROL, R. NASH, R. CAMPBELL, R. HUBER, AND E. STERNICK, *The development of a clinically intuitive approach to inverse treatment planning: Partial volume prescription and area cost function*, in Proceedings of the XIIth ICCR, Salt Lake City, Utah, 1997, pp. 317–319.
- [15] J. CLARKSON, *A note on depth doses in fields of irregular shape*, Br. J. Radiol., 14 (1941), pp. 265–268.
- [16] H. HAMACHER AND K.-H. KÜFER, *Inverse radiation therapy planning—a multiple objective optimization approach*, Discrete Appl. Math., 118 (2002), pp. 145–161.
- [17] J.-B. HIRIART-URRUTY AND C. LEMARECHAL, *Convex Analysis and Minimization Algorithms I: Fundamentals*, Springer-Verlag, Berlin, 1996.
- [18] S. KAMATH, S. SAHNI, J. LI, J. PALTA, AND S. RANKA, *Leaf sequencing algorithms for segmented multileaf collimation*, Phys. Med. Biol., 48 (2003), pp. 307–324.
- [19] K.-H. KÜFER, H. HAMACHER, AND T. BORTFELD, *A multicriteria optimization approach for inverse radiotherapy planning*, in Proceedings of the XIIIth ICCR, T. Bortfeld and W. Schlegel, eds., Heidelberg, Germany, 2000, pp. 26–29.
- [20] M. LAHANAS, E. SCHREIBMANN, AND D. BALTAS, *Multiobjective inverse planning for intensity modulated radiotherapy with constraint-free gradient-based optimization algorithms*, Phys. Med. Biol., 48 (2003), pp. 2843–2871.
- [21] M. LANGER, R. BROWN, M. URUE, J. LEONG, M. STRACHER, AND J. SHAPIRO, *Large-scale optimization of beam weights under dose-volume restrictions*, Int. J. Radiat. Oncol. Biol. Phys., 18 (1990), pp. 887–893.
- [22] E. LEE, T. FOX, AND I. CROCKER, *Optimization of radiosurgery treatment planning via mixed integer programming*, Med. Phys., 27 (2000), pp. 995–1004.
- [23] E. LEE, T. FOX, AND I. CROCKER, *Integer programming applied to intensity-modulated radiation treatment planning*, Ann. Oper. Res., 119 (2003), pp. 165–181.
- [24] G. MURPHY, W. LAWRENCE, AND R. LENLARD, EDS., *American Cancer Society Textbook on Clinical Oncology*, American Cancer Society, Atlanta, 1995.
- [25] W. OGRYCZAK AND A. RUSZCZYŃSKI, *Dual stochastic dominance and related mean-risk models*, SIAM J. Optim., 13 (2002), pp. 60–78.
- [26] W. OGRYCZAK AND A. TAMIR, *Minimizing the sum of the k largest functions in linear time*, Inform. Process. Lett., 85 (2003), pp. 117–122.
- [27] W. OGRYCZAK AND M. ZAWADZKI, *Conditional mean: A parametric solution concept for location problems*, Ann. Oper. Res., 110 (2002), pp. 167–181.
- [28] C. PEREZ AND L. BRADY, *Principles and Practice of Radiotherapy*, 3rd ed., Lippincott-Raven, Philadelphia, 1998.
- [29] R. ROCKAFELLAR AND S. URYASEV, *Optimization of conditional value-at-risk*, J. Risk, 2 (2000), pp. 21–41.
- [30] R. ROCKAFELLAR AND S. URYASEV, *Conditional value-at-risk for general loss distributions*, J. Banking Finance, 26 (2002), pp. 1443–1471.

- [31] H. ROMEIJN, R. AHUJA, J. DEMPSEY, AND A. KUMAR, *A new linear programming approach to radiation therapy treatment planning problems*, Oper. Res., to appear.
- [32] H. ROMEIJN, R. AHUJA, J. DEMPSEY, A. KUMAR, AND J. LI, *A novel linear programming approach to fluence map optimization for intensity modulated radiation therapy treatment planning*, Phys. Med. Biol., 48 (2003), pp. 3521–3542.
- [33] H. ROMEIJN, J. DEMPSEY, AND J. LI, *A unifying framework for multi-criteria fluence map optimization models*, Phys. Med. Biol., 49 (2004), pp. 1991–2013.
- [34] RTOG, H-0022: *A phase I/II study of conformal and intensity modulated irradiation for oropharyngeal cancer*, <http://www.rtog.org/members/protocols/h0022/h0022.pdf> (2000).
- [35] RTOG, H-0225: *A phase II study of intensity modulated radiation therapy (IMRT) +/- chemotherapy for nasopharyngeal cancer*, <http://www.rtog.org/members/protocols/0225/0225.pdf> (2002).
- [36] D. SANZ, *A mathematical study of the area over perimeter rule using the sector-integration equation*, Med. Phys., 27 (2000), pp. 2376–2379.
- [37] D. SHEPARD, M. EARL, X. LI, S. NAQVI, AND C. YU, *Direct aperture optimization: A turnkey solution for step-and-shoot IMRT*, Med. Phys., 29 (2002), pp. 1007–1018.
- [38] D. M. SHEPARD, M. C. FERRIS, G. H. OLIVERA, AND T. R. MACKIE, *Optimizing the delivery of radiation therapy to cancer patients*, SIAM Rev., 41 (1999), pp. 721–744.
- [39] J. SIEBERS, M. LAUTERBACH, P. KEALL, AND R. MOHAN, *Incorporating multi-leaf collimator leaf sequencing into iterative IMRT optimization*, Med. Phys., 29 (2002), pp. 952–959.
- [40] R. SIOCHI, *Minimizing static intensity modulation delivery time using an intensity solid paradigm*, Int. J. Radiat. Oncol. Biol. Phys., 43 (1999), pp. 671–689.
- [41] P. SLATER, *Centers to centroids in a graph*, J. Graph Theory, 2 (1978), pp. 209–222.
- [42] G. STEEL, *Basic Clinical Radiobiology for Radiation Oncologists*, Edward Arnold Publishers, London, 1994.
- [43] A. TAMIR, *The k-centrum multi-facility location problem*, Discrete Appl. Math., 109 (2001), pp. 293–307.
- [44] C. THIEKE, T. BORTFELD, AND K. KÜFER, *New optimization concepts in inverse treatment planning*, Radiotherapy and Oncology, 68 (2003), p. S2.
- [45] C. THIEKE, T. BORTFELD, A. NIEMIERKO, K. KÜFER, AND S. NILL, *Multicriteria optimization in inverse radiotherapy planning*, Radiotherapy and Oncology, 68 (2003), p. S44.
- [46] C. TSIEH, A. EISBRUCH, D. MCSHAN, M. KESSLER, R. MARSH, AND B. FRAASS, *Intensity-modulated radiation therapy (IMRT) for locally advanced paranasal sinus tumors: Incorporating clinical decisions in the optimization process*, Int. J. Radiat. Oncol. Biol. Phys., 55 (2003), pp. 776–784.
- [47] S. WEBB, *Intensity Modulated Radiation Therapy*, Institute of Physics Publishing, Bristol, UK, 2001.
- [48] S. WEBB AND T. LOMAX, *There is no IMRT?*, Phys. Med. Biol., 46 (2001), pp. L7–L8.
- [49] Q. WU AND R. MOHAN, *Algorithms and functionality of an intensity modulated radiotherapy optimization problem*, Med. Phys., 27 (2000), pp. 701–711.
- [50] P. XIA AND L. VERHEY, *Multileaf collimator leaf sequencing algorithm for intensity modulated beams with multiple static segments*, Med. Phys., 25 (1998), pp. 1424–1434.

A GLOBALLY CONVERGENT LINEARLY CONSTRAINED LAGRANGIAN METHOD FOR NONLINEAR OPTIMIZATION*

MICHAEL P. FRIEDLANDER[†] AND MICHAEL A. SAUNDERS[‡]

Abstract. For optimization problems with nonlinear constraints, linearly constrained Lagrangian (LCL) methods solve a sequence of subproblems of the form “minimize an augmented Lagrangian function subject to linearized constraints.” Such methods converge rapidly near a solution but may not be reliable from arbitrary starting points. Nevertheless, the well-known software package MINOS has proved effective on many large problems. Its success motivates us to derive a related LCL algorithm that possesses three important properties: it is globally convergent, the subproblem constraints are always feasible, and the subproblems may be solved inexactly.

The new algorithm has been implemented in MATLAB, with an option to use either MINOS or SNOPT (Fortran codes) to solve the linearly constrained subproblems. Only first derivatives are required. We present numerical results on a subset of the COPS, HS, and CUTE test problems, which include many large examples. The results demonstrate the robustness and efficiency of the stabilized LCL procedure.

Key words. large-scale optimization, nonlinear programming, nonlinear inequality constraints, augmented Lagrangian

AMS subject classifications. 49M37, 65K05, 90C30

DOI. 10.1137/S1052623402419789

1. Introduction. For optimization problems with nonlinear constraints, *linearly constrained Lagrangian methods* (LCL methods) solve a sequence of subproblems that minimize an augmented Lagrangian function subject to linearizations of the problem constraints. (Typically some of the constraints are already linear, and some are simple bounds on the variables. They are included verbatim in the subproblems.) Existing LCL methods converge rapidly near a solution but sometimes may not converge from arbitrary starting points. (They might not be *globally convergent*.) Nevertheless, the well-known software package MINOS [36] employs an LCL method and has proved effective on many problems (large and small), especially within the GAMS [8] and AMPL [19] environments. It is widely used in industry and academia. Its success motivates us to propose an LCL-like method for which global convergence to a local minimizer or a stationary point can be proved under standard assumptions.

Our globally convergent LCL algorithm, henceforth referred to as *stabilized* LCL (sLCL), solves a sequence of linearly constrained (LC) subproblems as just described. Each subproblem minimizes an augmented Lagrangian function within a linear manifold that describes a current approximation to the nonlinear constraints (including any linear constraints and bounds). This manifold is nominally a linearization of the

*Received by the editors December 16, 2002; accepted for publication (in revised form) August 12, 2004; published electronically April 22, 2005.

<http://www.siam.org/journals/siopt/15-3/41978.html>

[†]Department of Computer Science, University of British Columbia, Vancouver V6T 1Z4, BC, Canada (mpf@cs.ubc.ca). This work was supported by U.S. National Science Foundation grant CCR-9988205 and by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, U.S. Department of Energy contract W-31-109-Eng-38.

[‡]Department of Management Science and Engineering, Stanford University, Stanford, CA 94305-4026 (saunders@stanford.edu). This work was supported by U.S. National Science Foundation grants CCR-9988205 and CCR-0306662 and U.S. Office of Naval Research grants N00014-96-1-0274 and N00014-02-1-0076.

constraint space but may be a relaxed (i.e., larger) space at any stage, particularly during early iterations. Few conditions are imposed on the nature of the subproblem solutions; consequently, the subproblems may be solved with any of a variety of optimization routines for linearly constrained problems, providing much flexibility.

In addition to global convergence, the sLCL method possesses two important properties: the subproblems are always feasible, and they may be solved inexactly. In the method used by MINOS, there is no guarantee that the subproblems will be feasible, and the criteria for their early termination is heuristic. Our method may be regarded as a generalization of sequential augmented Lagrangian methods (see, for example, [26, 1, 18]). The theory we develop provides a framework that unifies Robinson's LCL method [40] with the bound-constrained Lagrangian (BCL) method used, for example, by LANCELOT [11]. In the context of our theory, the proposed algorithm is actually a continuum of methods, with Robinson's LCL method and the BCL method at opposite ends of a spectrum. The sLCL method exploits this relationship, preserving the fast local convergence properties of LCL methods while inheriting the global convergence properties of BCL methods. This connection is explored in more detail by Friedlander [20].

Our intent is to develop a method that is effective in practice for large-scale problems and is also based on sound theory. We implemented the sLCL method as a MATLAB program that calls either the reduced-gradient part of MINOS [35] or the sequential quadratic programming code SNOPT [25] to solve the linearly constrained subproblems. These solvers are most efficient on problems with few degrees of freedom. Also, they use only first derivatives, and consequently our implementation requires only first derivatives. Following the theoretical development, we give computational results and a comparison with MINOS on a set of nontrivial problems.

1.1. The optimization problem. The proposed method solves nonlinearly constrained optimization problems of the form

(GNP)	$\begin{aligned} &\underset{x \in \mathbb{R}^n}{\text{minimize}} && f(x) \\ &\text{subject to} && c(x) = 0, \\ &&& x \geq 0, \end{aligned}$

where $f : \mathbb{R}^n \mapsto \mathbb{R}$ is a linear or nonlinear objective function and $c : \mathbb{R}^n \mapsto \mathbb{R}^m$ is a vector of nonlinear constraint functions. We assume that the constraints of (GNP) are feasible. In section 5.6 we explain how the proposed algorithm reveals infeasible constraints and we discuss properties of the points to which it converges. In section 7 we consider a formulation with more general constraints (for which our implementation and other solvers are designed).

One of the strengths of our method is that it does not explicitly require second-order information. However, the fast convergence rate of the algorithm relies on sufficient smoothness of the nonlinear functions, indicated by the existence of second derivatives. Second derivatives could be used by the subproblem solver if they were available, thus accelerating the subproblem solution process and changing the properties of the solutions obtained for (GNP). We discuss this further in section 5.7.

1.2. The LCL approach. The acronyms LCL and BCL are new. Methods belonging to the LCL class have previously been referred to as sequential linearized constraint (SLC) methods (cf. [26, 37]). The term SLC was chosen for compatibility with the terms sequential quadratic programming (SQP) and sequential linear

programming (SLP). Those methods also sequentially linearize the constraints. The term *linearly constrained Lagrangian* emphasizes that the Lagrangian itself (really the augmented Lagrangian) is used in the subproblems and not a linear or quadratic approximation. Moreover, there is a useful relationship (which we exploit) between LCL and BCL methods, and this is hinted at by the nomenclature.

The first LCL methods were proposed independently in 1972. Robinson [40] and Rosen and Kreuser [43] describe similar algorithms based on minimizing a sequence of Lagrangian functions subject to linearized constraints. Robinson is able to prove that, under suitable conditions, the sequence of subproblem solutions converges quadratically to a solution of (GNP). A practical strength is that efficient large-scale methods exist for the solution of the linearly constrained subproblems formed at each iteration. Any suitable subproblem solver may be called as a black box.

1.3. Other work on stabilizing LCL methods. Other approaches to stabilizing LCL algorithms include two-phase methods proposed by Rosen [42] and Van der Hoek [44]. In these approaches, a Phase 1 problem is formed by moving the non-linear constraints into the objective by means of a quadratic penalty function. The solution of the Phase 1 problem is used to initialize Robinson's method (Phase 2). With a sufficiently large penalty parameter, the Phase 1 solution will yield a starting point that allows Robinson's method to converge quickly to a solution. These two-phase methods choose the penalty parameter arbitrarily, however, and do not deal methodically with infeasible linearizations.

In 1981, Best et al. [2] describe a variant of the two-phase method whereby the Phase 1 penalty parameter is gradually increased by repeated return to the Phase 1 problem if the Phase 2 iterations are not converging. This two-phase method differs further from Rosen's and Van der Hoek's methods in that the Phase 2 iterations involve only those constraints identified as active by the Phase 1 problem. The authors are able to retain local quadratic convergence of the Phase 2 LCL iterations while proving global convergence to a stationary point. A drawback of their method is that it requires a fourth-order penalty term to ensure continuous second derivatives of the penalty objective. This requirement may introduce significant numerical difficulty for the solution of the Phase 1 problem (though probably a quadratic-penalty term would suffice in practice).

Both two-phase methods share the disadvantage that the Phase 1 penalty problems need to be optimized over a larger subspace than the subsequent LCL phase. We seek a method that retains all linearized constraints as part of each subproblem (one advantage being that it keeps the number of degrees of freedom small). We also allow the subproblems to predict the set of active constraints, as in Robinson's 1972 method. In particular, the active set for the final subproblem is taken to be the active set for the solution of the original problem.

1.4. Definitions. Unless otherwise specified, the function $\|v\|$ represents the Euclidean norm of a vector v . With vector arguments, the functions $\min(\cdot, \cdot)$ and $\max(\cdot, \cdot)$ are defined componentwise.

We define the *augmented Lagrangian function* corresponding to (GNP) as

$$\mathcal{L}(x, y, \rho) = f(x) - y^T c(x) + \frac{1}{2} \rho \|c(x)\|_2^2,$$

where x , the m -vector y , and the scalar ρ are independent variables. This function is particularly important for our analysis. When y_k and ρ_k are fixed, we often use the shorthand notation $\mathcal{L}_k(x) \equiv \mathcal{L}(x, y_k, \rho_k)$.

Let $g(x)$ denote the gradient of the objective function $f(x)$, and let $J(x)$ denote the Jacobian of $c(x)$: a matrix whose i th row is the gradient of $[c(x)]_i$, where $[\cdot]_i$ refers to the i th component of a vector. Let $H(x)$ and $H_i(x)$ be the Hessian matrices of $f(x)$ and $[c(x)]_i$, respectively, and define the *first-order multiplier estimate*

$$(1.1) \quad \tilde{y}(x, y, \rho) = y - \rho c(x).$$

The derivatives of $\mathcal{L}$ with respect to x may be written as follows:

$$(1.2) \quad \nabla_x \mathcal{L}(x, y, \rho) = g(x) - J(x)^T \tilde{y}(x, y, \rho),$$

$$(1.3) \quad \nabla_{xx}^2 \mathcal{L}(x, y, \rho) = H(x) - \sum_{i=1}^m [\tilde{y}(x, y, \rho)]_i H_i(x) + \rho J(x)^T J(x).$$

We assume that problem (GNP) is feasible and has at least one point (x_*, y_*, z_*) that satisfies the first-order Karush–Kuhn–Tucker (KKT) conditions.

DEFINITION 1.1 (first-order KKT conditions). *A triple (x_*, y_*, z_*) is a first-order KKT point for (GNP) if for some $\rho \geq 0$ all of the following hold:*

$$(1.4a) \quad c(x_*) = 0,$$

$$(1.4b) \quad \nabla_x \mathcal{L}(x_*, y_*, \rho) = z_*,$$

$$(1.4c) \quad \min(x_*, z_*) = 0.$$

The vector of Lagrange multipliers z_* corresponds to the bound constraints $x \geq 0$. Its elements are referred to as *reduced costs* in the linear programming literature.

KKT conditions are normally defined in terms of the Lagrangian function, not the augmented Lagrangian used in (1.4b). We take the perspective that the Lagrangian is just a special case of the augmented Lagrangian (with $\rho = 0$). Our definition is not more restrictive: if (1.4) holds for a particular ρ , then it holds for all ρ (including $\rho = 0$) because $c(x_*) = 0$.

Also, (1.4c) is shorthand for nonnegativity and complementarity conditions that are normally defined explicitly:

$$(1.5) \quad x_* \geq 0, \quad z_* \geq 0 \quad \text{and} \quad [x_*]_j = 0 \quad \text{or} \quad [z_*]_j = 0.$$

Let $\eta_* > 0$ and $\omega_* > 0$ be specified as primal and dual convergence tolerances. We regard the point (x, y, z) as an acceptable approximate solution of (GNP) if it satisfies (1.4) to within these tolerances:

$$(1.6a) \quad \|c(x)\| \leq \eta_*,$$

$$(1.6b) \quad \nabla_x \mathcal{L}(x, y, \rho) = z,$$

$$(1.6c) \quad \|\min(x, z)\|_\infty \leq \omega_*,$$

where (1.6c) is shorthand for

$$x \geq -\omega_* e, \quad z \geq -\omega_* e \quad \text{and} \quad [x]_j \leq \omega_* \quad \text{or} \quad [z]_j \leq \omega_*,$$

where e is a vector of ones. In practice, we might use a different tolerance for the nonnegativity of x , replacing the first inequality by $x \geq -\delta_* e$ for some $\delta_* > 0$.

The algorithms we discuss are structured around *major* and *minor* iterations. Each major iteration solves a linearly constrained subproblem to generate an element

of the sequence $\{(x_k, y_k, z_k)\}$. Under certain (desirable) circumstances, this sequence converges to a solution (x_*, y_*, z_*) . For each major iteration k , there is a corresponding set of minor iterates converging to $(x_k^*, \Delta y_k^*, z_k^*)$, a local solution of the current subproblem. In our development and analysis of a stabilized LCL method we are primarily concerned with the “outer”-level algorithm, comprising the major iterations. Unless stated otherwise, “iterations” refers to major iterations.

For the theoretical development of our method, it is convenient to use the normal augmented Lagrangian for the objective of the subproblems. The dual solutions of the subproblems then converge to 0 (not to y_*) as $k \rightarrow \infty$. Hence our notation Δy_k^* . (As mentioned in section 9.2, a *modified* augmented Lagrangian is desirable in practice, and the dual solutions then converge to y_* .)

Let $\mathcal{I}$ be the index set $\mathcal{I} = \{j \in 1, \dots, n \mid [x_*]_j > 0\}$ corresponding to inactive bounds at x_* . We define

$$(1.7) \quad \widehat{g}(x) \equiv [g(x)]_{\mathcal{I}} \quad \text{and} \quad \widehat{J}(x)$$

to be the elements of $g(x)$ and the columns of $J(x)$ corresponding to the indices in $\mathcal{I}$. Further notation follows, including some already introduced:

(x_*, y_*, z_*)	optimal primal and dual variables for (GNP)
(x_k, y_k, z_k)	the k th estimate of (x_*, y_*, z_*)
$(x_k^*, \Delta y_k^*, z_k^*)$	solution of the k th subproblem
f_k, g_k, c_k, J_k	functions and gradients evaluated at x_k
f_*, g_*, c_*, J_*	functions and gradients evaluated at x_*
$\bar{c}_k(x)$	$c_k + J_k(x - x_k)$, the linearization of $c(x)$ at x_k
$\mathcal{L}_k(x)$	$\mathcal{L}(x, y_k, \rho_k)$, the augmented Lagrangian with fixed y_k and ρ_k
$\tilde{y}(x, y, \rho)$	$y - \rho c(x)$, the first-order multiplier estimate in (1.1)–(1.3)
$\widehat{y}(x)$	the least-squares multiplier estimate (5.1).

The symbol x_* (used above) also denotes a limit point of the sequence $\{x_k^*\}$. For a single limit point, $x_k^* \rightarrow x_*$ for all k sufficiently large. When there are multiple limit points, we use $(x_*, \mathcal{K})$ to denote the limit point and the infinite set of indices associated with the corresponding convergent subsequence.

1.5. Assumptions. The first assumption holds implicitly throughout.

ASSUMPTION 1.2. *The functions f and c are twice continuously differentiable.* The following assumptions are used in sections 5–6.

ASSUMPTION 1.3. *The sequence $\{x_k^*\}$ lies in a closed and bounded set $\mathcal{B} \subset \mathbb{R}^n$.* (This ensures at least one convergent subsequence and at least one limit point x_* .)

ASSUMPTION 1.4. *The matrix $\widehat{J}(x_*)$ has full row rank at every limit point x_* of the sequence $\{x_k^*\}$.* (This is commonly known as the *linear independence constraint qualification* (LICQ); see Mangasarian [31] or Nocedal and Wright [37]. For infeasible problems, section 5.6 does not need this assumption.)

2. The canonical LCL method. Algorithm 1 below outlines what we regard to be the canonical LCL method for solving problem (GNP). It solves a sequence of linearly constrained subproblems that are parameterized by the latest estimates (x_k, y_k) and a fixed penalty parameter $\rho_k \equiv \bar{\rho}$ (which may be zero):

Algorithm 1: Canonical LCL.

Input: x_0, y_0, z_0
Output: x_*, y_*, z_*

Set the penalty parameter $\rho_0 = \bar{\rho} \geq 0$.
Set positive convergence tolerances $\omega_*, \eta_* \ll 1$.
 $k \leftarrow 0$
converged $\leftarrow$ false

repeat
Find a point $(x_k^*, \Delta y_k^*, z_k^*)$ that solves (LC_k) (2.1).
1 If there is more than one such point, choose $(x_k^*, \Delta y_k^*, z_k^*)$ closest to $(x_k, 0, z_k)$.
 $x_{k+1} \leftarrow x_k^*, y_{k+1} \leftarrow y_k + \Delta y_k^*, z_{k+1} \leftarrow z_k^*$ [update solution estimates]
if $(x_{k+1}, y_{k+1}, z_{k+1})$ satisfies (1.6) **then** converged $\leftarrow$ true
 $\rho_k \leftarrow \bar{\rho}$ [keep ρ_k fixed]
 $k \leftarrow k + 1$
until converged

$x_* \leftarrow x_k, y_* \leftarrow y_k, z_* \leftarrow z_k$

(LC_k)	$\begin{aligned} & \underset{x}{\text{minimize}} && \mathcal{L}_k(x) \\ & \text{subject to} && \bar{c}_k(x) = 0, \\ & && x \geq 0. \end{aligned}$

A local solution of (LC_k) is a triple $(x_k^*, \Delta y_k^*, z_k^*)$ that satisfies the first-order KKT conditions

$$(2.1a) \quad \bar{c}_k(x) = 0,$$

$$(2.1b) \quad \nabla \mathcal{L}_k(x) - J_k^T \Delta y = z,$$

$$(2.1c) \quad \min(x, z) = 0.$$

The original LCL method of Robinson [40] is equivalent to Algorithm 1 with $\bar{\rho} = 0$. A positive penalty parameter was introduced in the software package MINOS [36] to improve convergence from difficult starting points. Empirically, it has proved a helpful addition to Robinson's method for many problems, but sometimes it is ineffective. A theoretical understanding of when and how to modify ρ_k has been lacking.

For any set of parameters x_k, y_k , and ρ_k there may be many points $(x_k^*, \Delta y_k^*, z_k^*)$ that satisfy (2.1). Step 1 of Algorithm 1 (the line labeled 1) requires that the chosen point be closest in norm to $(x_k, 0, z_k)$. This step is found in the original formulation of Robinson's method, where it is used in the local convergence proof [40, Theorem 2], and we will continue to use it in our formulation of the sLCL method. Such a requirement cannot be verified in practice, but it may be argued that, depending on the choice of subproblem solver, the requirement may often be satisfied. Note that the requirement is used only in the local analysis (section 6) of the sLCL method, and that the point $(x_k^*, \Delta y_k^*, z_k^*)$ is locally unique under the assumptions of that analysis. However, in general the point may not be unique.

3. An elastic LC subproblem. We recognize two particular causes of failure in the canonical LCL method:

- The linear constraints in subproblem (LC_k) may be infeasible for some x_k , so that x_{k+1} may not be well defined.
- Even if the linearized constraints are feasible, $\|x_k^* - x_k\|$ may be arbitrarily large regardless of the values of y_k and ρ_k in the subproblem objective.

To remedy both deficiencies we modify the linearized constraints, allowing some degree of flexibility in their satisfaction. We introduce nonnegative *elastic variables* v and w into the constraints, with a penalty on these variables in the subproblem objective. The resulting elastic subproblem is *always feasible*:

$$\begin{array}{ll} \text{(ELC}_k\text{)} & \begin{array}{l} \text{minimize}_{x,v,w} \quad \mathcal{L}_k(x) + \sigma_k e^T(v + w) \\ \text{subject to} \quad \bar{c}_k(x) + v - w = 0, \\ \quad \quad \quad x, v, w \geq 0. \end{array} \end{array}$$

The solution of (ELC_k) is a 5-tuple $(x_k^*, v_k^*, w_k^*, \Delta y_k^*, z_k^*)$ that satisfies the first-order KKT conditions

$$(3.1a) \quad \bar{c}_k(x) + v - w = 0,$$

$$(3.1b) \quad \nabla \mathcal{L}_k(x) - J_k^T \Delta y = z,$$

$$(3.1c) \quad \min(x, z) = 0,$$

$$(3.1d) \quad \min(v, \sigma_k e - \Delta y) = 0,$$

$$(3.1e) \quad \min(w, \sigma_k e + \Delta y) = 0.$$

The last two conditions (3.1d)–(3.1e) imply that their arguments are nonnegative, and so $\sigma_k e \geq \Delta y \geq -\sigma_k e$. This pair of inequalities can be conveniently restated as

$$(3.2) \quad \|\Delta y\|_\infty \leq \sigma_k.$$

The sLCL method is based on subproblem (ELC_k) with judicious changes to ρ_k and σ_k for each k . We find later that the bound (3.2)—relaxed slightly in (3.4)—is crucial for the global convergence analysis of the method. In particular, setting σ_k small allows us to counteract the effect of a large $\|x_k^* - x_k\|$ (see (5.2)).

3.1. The ℓ_1 penalty function. The term $\sigma_k e^T(v + w)$ is the ℓ_1 penalty function. Together with the constraints $v, w \geq 0$, it is equivalent to a penalty on $\|v - w\|_1$ (see [28]). Eliminating $v - w$, we see that the elastic subproblem (ELC_k) can be stated as

$$\begin{array}{ll} \text{(ELC}'_k\text{)} & \begin{array}{l} \text{minimize}_x \quad \mathcal{L}_k(x) + \sigma_k \|\bar{c}_k(x)\|_1 \\ \text{subject to} \quad x \geq 0, \end{array} \end{array}$$

with solution (x_k^*, z_k^*) . This immediately reveals the stabilized LCL method's intimate connection with the augmented Lagrangian function and BCL methods. Far from a solution, the ℓ_1 penalty term $\sigma_k \|\bar{c}_k(x)\|_1$ gives the method an opportunity to deviate from the constraint linearizations and reduce $\mathcal{L}_k(x)$ more than otherwise. Near a solution, it keeps the iterates close to the linearizations. The following two extreme cases are of interest.

Recovering the BCL subproblem. Set $\sigma_k = 0$. Then (ELC_k) and (ELC'_k) reduce to the equivalent bound-constrained minimization problem

$$\begin{array}{ll} \text{(BC}_k\text{)} & \begin{array}{l} \text{minimize}_x \quad \mathcal{L}_k(x) \\ \text{subject to} \quad x \geq 0. \end{array} \end{array}$$

Subproblem (BC_k) is used by BCL methods (e.g., Hestenes [29], Powell [39], Bertsekas [1], Conn, Gould, and Toint [10]) and in particular by LANCELOT [11].

Recovering the LCL subproblem. The ℓ_1 penalty function is *exact*. If the linearization is feasible and σ_k is larger than a certain threshold, v and w are likely to be zero and the minimizers of the elastic problem (ELC_k) will coincide with the minimizers of the inelastic problem (LC_k) . Exact penalty functions have been studied by [28, 1, 17] among others. See Conn, Gould, and Toint [13] for a more recent discussion.

We are particularly interested in this feature when the iterates generated by the sLCL algorithm are approaching a solution (x_*, y_*, z_*) . Recovering the canonical LCL subproblem ensures that sLCL inherits the fast local convergence properties of the canonical LCL algorithm. In section 6.2 we give conditions under which the elastic variables will be forced to zero.

3.2. Early termination of the LC subproblems. Poor values of x_k , y_k , or ρ_k may imply subproblems whose accurate solutions are far from a solution of (GNP). We therefore terminate subproblems early by relaxing (3.1c)–(3.1e) by an amount ω_k . However, we enforce nonnegativity on x, v, w (also implied by (3.1c)–(3.1e)):

$$(3.3a) \quad x, v, w \geq 0,$$

$$(3.3b) \quad \bar{c}_k(x) + v - w = 0,$$

$$(3.3c) \quad \nabla \mathcal{L}_k(x) - J_k^T \Delta y = z,$$

$$(3.3d) \quad \|\min(x, z)\|_\infty \leq \omega_k,$$

$$(3.3e) \quad \|\min(v, \sigma_k - \Delta y)\|_\infty \leq \omega_k,$$

$$(3.3f) \quad \|\min(w, \sigma_k + \Delta y)\|_\infty \leq \omega_k.$$

Note that (3.3e) and (3.3f) imply

$$(3.4) \quad \|\Delta y\|_\infty \leq \sigma_k + \omega_k.$$

As discussed in connection with (1.5), in practice (3.3a) is typically relaxed by a fixed tolerance δ by instead requiring that $x, v, w \geq -\delta e$.

3.3. Relation to Sl_1QP and SNOPT. Subproblems (ELC_k) and (ELC'_k) are reminiscent of the QP subproblems arising in the Sl_1QP method of Fletcher [17] and the SQP method of SNOPT [25]. The main difference is that the QP objectives are *quadratic approximations to the Lagrangian* (not the Lagrangian itself, and not the augmented Lagrangian).

In Sl_1QP , trust-region bounds are used to ensure boundedness of $\|x_k^* - x_k\|_\infty$ and to permit the subproblem solution x_k^* to reduce the exact penalty function $\nu f(x) + \|c(x)\|_1$ for some (small enough) scalar ν . Also, second-order corrections are needed to maintain rapid convergence of the major iterations (avoiding the Maratos effect [32]).

In SNOPT, the QP subproblems are used to define *search directions* $(\Delta x, \Delta y)$ for the augmented Lagrangian $\mathcal{L}(x, y, \rho_k)$ regarded as a function of $x = x_k + \alpha \Delta x$ and $y = y_k + \alpha \Delta y$ for some steplength $\alpha \in (0, 1]$. The elastic slacks v and w are initially absent but become a permanent part of the original problem (and the QP subproblems) if the linearized constraints are infeasible at some x_k , or if $\|y_k\|$ becomes large. The net effect is that for problems whose constraints are infeasible, Sl_1QP and SNOPT minimize the 1-norm of $c(x)$, whereas sLCL minimizes the 2-norm (see section 5.6).

In all three methods (S ℓ_1 QP, SNOPT, and sLCL), the penalty term $\sigma_k e^T(v + w)$ may be nonzero at a particular subproblem solution, indicating that some of the linearized constraints are not satisfied. Thus, the active linearized constraints may be just a subset of the full set (cf. Fletcher [17, p. 32]). The global convergence properties of sLCL do not require independent constraint gradients or bounded multipliers for every subproblem. (These are required only at limit points of the sequence generated by the method.)

4. The stabilized LCL algorithm. Algorithm 2 outlines the sLCL algorithm. Its structure closely parallels the BCL algorithm described in Conn, Gould, and Toint [10]. Based on the current primal infeasibility, each iteration of the sLCL algorithm is regarded as either “successful” or “unsuccessful.” In the “successful” case, the solution estimates are updated by using information from the current subproblem solution. Otherwise, the subproblem solutions are discarded, the current solution estimates are retained, and the penalty parameter ρ_k is increased in an effort to reduce the primal infeasibility in the next iteration. To prevent the linearized constraints from interfering with the penalty parameter’s ability to reduce the primal infeasibility, the algorithm may reduce the elastic penalty parameter σ_k . (A small value of σ_k encourages deviation from the current constraint linearizations, which may be poor approximations to the true constraints.)

The parameter $\bar{\sigma}$ ensures that $\|\Delta y_k^*\|$ remains uniformly bounded. In particular, $\rho_k > 0$ and $\tau_\sigma > 1$ in steps 8 and 12 ensure that σ_k is uniformly bounded, and step 1 ensures that ω_k is uniformly bounded; therefore $\|\Delta y_k^*\|$ is uniformly bounded by virtue of (3.4). The division by $1 + \rho_k$ is a safeguard if the algorithm continues to oscillate between successful and unsuccessful iterations; in that event, $\rho_k \rightarrow \infty$ and $\sigma_k \rightarrow 0$, and the algorithm becomes increasingly similar to a quadratic penalty function method.

The two salient features of this algorithm are that it is globally convergent and asymptotically equivalent to the canonical LCL method. In section 5 we demonstrate the global convergence properties of the algorithm by proving results analogous to Lemma 4.3 and Theorem 4.4 in [10]. In section 6 we demonstrate that the algorithm eventually reduces to the canonical LCL method and hence inherits that method’s asymptotic convergence properties.

5. Global convergence. Let $(x_*, \mathcal{K})$ be a limit point of the sequence $\{x_k^*\}$. Assumptions 1.2 and 1.4 guarantee that $\hat{J}(x)$ has full row rank at x_* and at all neighboring points. At such points x we define the *least-squares multiplier estimate* $\hat{y}(x)$ as the unique solution of the least-squares problem

$$(5.1) \quad \hat{y}(x) \equiv \arg \min_y \|\hat{g}(x) - \hat{J}(x)^T y\|^2.$$

Note that the definitions of $\hat{g}$, $\hat{J}$, and hence $\hat{y}$ require a priori knowledge of the bounds active at x_* (see (1.7)). We emphasize that $\hat{y}$ is used only as an analytical device and its computation is never required.

For the global analysis we assume that $\omega_* = \eta_* = 0$.

5.1. Preliminaries. We need the following lemma to bound the errors in the least-squares multiplier estimates relative to $\|x_k^* - x_*\|$, the error in x_k^* . The lemma follows from Lemmas 2.1 and 4.4 of Conn et al. [9]. It simply demonstrates that $\hat{y}(x)$ is Lipschitz continuous in a neighborhood of x_* .

Algorithm 2: Stabilized LCL.

Input: x_0, y_0, z_0
Output: x_*, y_*, z_*

Set $\bar{\sigma} \gg 1$ and scale factors $\tau_\rho, \tau_\sigma > 1$.
Set penalty parameters $\rho_0 > 0$ and $\sigma_0 \in [1, \bar{\sigma}]$.
Set positive convergence tolerances $\omega_*, \eta_* \ll 1$ and infeasibility tolerance $\eta_0 > \eta_*$.
Set constants $\alpha, \beta > 0$ with $\alpha < 1$.
 $k \leftarrow 0$
converged $\leftarrow$ false

repeat

- 1 Choose $\omega_k > 0$ such that $\lim_{k \rightarrow \infty} \omega_k \leq \omega_*$.
- 2 Find a point $(x_k^*, v_k^*, w_k^*, \Delta y_k^*, z_k^*)$ that solves (ELC_k) approximately as in (3.3).
If there is more than one such point, choose $(x_k^*, \Delta y_k^*, z_k^*)$ closest to $(x_k, 0, z_k)$.
- 3 **if** $\|c(x_k^*)\| \leq \max(\eta_*, \eta_k)$ **then**
- 4 $x_{k+1} \leftarrow x_k^*$ [update solution estimates]
- 5 $y_{k+1} \leftarrow y_k + \Delta y_k^* - \rho_k c(x_k^*) \equiv \tilde{y}(x_k^*, y_k + \Delta y_k^*, \rho_k)$
- 6 $z_{k+1} \leftarrow z_k^*$
- 7 **if** $(x_{k+1}, y_{k+1}, z_{k+1})$ satisfies (1.6) **then** converged $\leftarrow$ true
- 8 $\rho_{k+1} \leftarrow \rho_k$ [keep ρ_k]
- 9 $\sigma_{k+1} \leftarrow \frac{1}{1+\rho_k} \min(1 + \|\Delta y_k^*\|_\infty, \bar{\sigma})$ [reset σ_k]
- 9 $\eta_{k+1} \leftarrow \eta_k / (1 + \rho_{k+1}^\beta)$ [decrease η_k]
- else**
- 10 $x_{k+1} \leftarrow x_k, y_{k+1} \leftarrow y_k, z_{k+1} \leftarrow z_k$ [keep solution estimates]
- 11 $\rho_{k+1} \leftarrow \tau_\rho \rho_k$ [increase ρ_k]
- 12 $\sigma_{k+1} \leftarrow \sigma_k / \tau_\sigma$ [decrease σ_k]
- 13 $\eta_{k+1} \leftarrow \eta_0 / (1 + \rho_{k+1}^\alpha)$ [may increase or decrease η_k]
- end**
- $k \leftarrow k + 1$

until converged

$x_* \leftarrow x_k, y_* \leftarrow y_k, z_* \leftarrow z_k$

LEMMA 5.1. *Suppose Assumption 1.4 holds, and let $(x_*, \mathcal{K})$ be a limit point of $\{x_k^*\}$. Then there exists a positive constant α such that $\|\tilde{y}(x_k^*) - \tilde{y}(x_*)\| \leq \alpha \|x_k^* - x_*\|$ for all $k \in \mathcal{K}$ sufficiently large.*

To prove the global convergence properties of Algorithm 2, we first describe the properties of any limit point that the algorithm generates. We are not claiming (yet!) that the algorithm is globally convergent, only that if it *does* converge, then the set of limit points generated must satisfy some desirable properties. The following lemma is adapted from Lemma 4.4 of [9].

LEMMA 5.2. *Suppose Assumptions 1.3 and 1.4 hold. Let $\{\omega_k\}$ and $\{\rho_k\}$ be sequences of positive scalars, where $\omega_k \rightarrow 0$. Let $\{x_k\}$ be any sequence of n -vectors in $\mathcal{B}$ and let $\{y_k\}$ be any sequence of m -vectors. Let $\{(x_k^*, \Delta y_k^*, z_k^*)\}$ be a sequence of vectors satisfying (3.3a), (3.3c), and (3.3d). Let $(x_*, \mathcal{K})$ be a limit point of the sequence $\{x_k^*\}$. Set $\tilde{y}_k = \tilde{y}(x_k^*, y_k + \Delta y_k^*, \rho_k)$ and $y_* = \tilde{y}(x_*)$. The following properties then hold:*

1. *There are positive constants α_1, α_2 , and M such that*

$$(5.2) \quad \|\tilde{y}_k - y_*\| \leq \beta_1 \equiv \alpha_1 \omega_k + M \|x_k^* - x_k\| \|\Delta y_k^*\| + \alpha_2 \|x_k^* - x_*\|,$$

$$(5.3) \quad \rho_k \|c(x_k^*)\| \leq \beta_2 \equiv \beta_1 + \|\Delta y_k^*\| + \|y_k - y_*\|,$$

for all $k \in \mathcal{K}$ sufficiently large.

2. If $\|\Delta y_k^*\| \rightarrow 0$ as $k \in \mathcal{K} \rightarrow \infty$, or $\|\Delta y_k^*\|$ is bounded above and $\|x_k^* - x_k\| \rightarrow 0$, then $\tilde{y}_k \rightarrow y_*$ and $z_k^* \rightarrow z_* \stackrel{\text{def}}{=} \nabla_x \mathcal{L}(x_*, y_*, 0)$.
3. If $c_* = 0$ also, then (x_*, y_*, z_*) is a first-order KKT point for (GNP).

Proof. From Assumption 1.4, $\hat{J}(x_k^*)$ has full row rank for all $k \in \mathcal{K}$ large enough. The least-squares multiplier estimate $\hat{y}(x_k^*)$ (5.1) therefore exists. For full-rank least-squares problems in general, it is straightforward to show that the error in an approximate solution is bounded by a finite multiple of the associated residual. Hence with $x = x_k^*$ in problem (5.1) and with $y = \tilde{y}_k$ as an approximate solution, we may write

$$(5.4) \quad \|\hat{y}(x_k^*) - \tilde{y}_k\| \leq \frac{\alpha_1}{\sqrt{n}} \|\hat{g}(x_k^*) - \hat{J}(x_k^*)^T \tilde{y}_k\|$$

for some positive scalar α_1 .

We now show that $\|\hat{g}(x_k^*) - \hat{J}(x_k^*)^T \tilde{y}_k\|$ is bounded. By hypothesis, $(x_k^*, \Delta y_k^*, z_k^*)$ satisfies (3.3c). Together with (1.2),

$$(5.5) \quad \begin{aligned} z_k^* &= \nabla \mathcal{L}_k(x_k^*) - J_k^T \Delta y_k^* \\ &= g(x_k^*) - J(x_k^*)^T (y_k - \rho_k c(x_k^*)) - J_k^T \Delta y_k^* \\ &= g(x_k^*) - J(x_k^*)^T (y_k + \Delta y_k^* - \rho_k c(x_k^*)) + (J(x_k^*) - J_k)^T \Delta y_k^* \\ &= g(x_k^*) - J(x_k^*)^T \tilde{y}_k + (J(x_k^*) - J_k)^T \Delta y_k^*, \end{aligned}$$

where $\tilde{y}_k \stackrel{\text{def}}{=} \tilde{y}(x_k^*, y_k + \Delta y_k^*, \rho_k) = y_k + \Delta y_k^* - \rho_k c(x_k^*)$. For $k \in \mathcal{K}$ large enough, x_k^* is sufficiently close to x_* so that $[x_k^*]_j > 0$ if $[x_*]_j > 0$. Therefore, (3.3d) and $\omega_k \rightarrow 0$ imply that $\min([x_k^*]_j, [z_k^*]_j) = [z_k^*]_j$, so that

$$(5.6) \quad \|[z_k^*]_{\mathcal{I}}\| \leq \|\min(x_k^*, z_k^*)\|,$$

where $\mathcal{I}$ is the index set of inactive bounds at x_* as defined in section 1.4. Because x_k^* and z_k^* both satisfy (3.3d), (5.6) implies that

$$(5.7) \quad \|[z_k^*]_{\mathcal{I}}\| \leq \sqrt{n} \omega_k.$$

Combining (5.5) and (5.7), we obtain

$$\|\hat{g}(x_k^*) - \hat{J}(x_k^*)^T \tilde{y}_k + (\hat{J}(x_k^*) - \hat{J}_k)^T \Delta y_k^*\| \leq \sqrt{n} \omega_k$$

(see (1.7)). Also, because c is twice-continuously differentiable, J is Lipschitz continuous over $\mathcal{B}$ and there exists a positive constant M such that $\|\hat{J}(x_k^*) - \hat{J}_k\| \leq M \frac{\sqrt{n}}{\alpha_1} \|x_k^* - x_k\|$. Hence, from the triangle and Cauchy-Schwartz inequalities, we have

$$(5.8) \quad \begin{aligned} \|\hat{g}(x_k^*) - \hat{J}(x_k^*)^T \tilde{y}_k\| &\leq \|\hat{g}(x_k^*) - \hat{J}(x_k^*)^T \tilde{y}_k + (\hat{J}(x_k^*) - \hat{J}_k)^T \Delta y_k^*\| \\ &\quad + \|\hat{J}(x_k^*) - \hat{J}_k\| \|\Delta y_k^*\| \\ &\leq \sqrt{n} \omega_k + M \frac{\sqrt{n}}{\alpha_1} \|x_k^* - x_k\| \|\Delta y_k^*\|, \end{aligned}$$

and so we have bounded the left-hand side as required.

We now derive (5.2). From the triangle inequality,

$$(5.9) \quad \|\tilde{y}_k - y_*\| \leq \|\hat{y}(x_k^*) - \tilde{y}_k\| + \|\hat{y}(x_k^*) - y_*\|.$$

Using inequality (5.8) in (5.4), we deduce that

$$(5.10) \quad \|\widehat{y}(x_k^*) - \widetilde{y}_k\| \leq \alpha_1 \omega_k + M \|x_k^* - x_k\| \|\Delta y_k^*\|,$$

and Lemma 5.1 implies that there exists a constant α_2 such that

$$(5.11) \quad \|\widehat{y}(x_k^*) - y_*\| \leq \alpha_2 \|x_k^* - x_*\|,$$

for all $k \in \mathcal{K}$ large enough (recall that $y_* \equiv \widehat{y}(x_*)$). Substituting (5.10) and (5.11) into (5.9), we obtain $\|\widetilde{y}_k - y_*\| \leq \beta_1$ as stated in (5.2).

We now prove (5.3). From the definition of $\widetilde{y}_k$, rearranging terms yields

$$(5.12) \quad \rho_k c(x_k^*) = y_k + \Delta y_k^* - \widetilde{y}_k.$$

Taking norms of both sides of (5.12) and using (5.2) yields

$$\begin{aligned} \rho_k \|c(x_k^*)\| &= \|y_k + \Delta y_k^* - \widetilde{y}_k\| = \|y_k - y_* + y_* - \widetilde{y}_k + \Delta y_k^*\| \\ &\leq \|\widetilde{y}_k - y_*\| + \|y_k - y_*\| + \|\Delta y_k^*\| \\ &\leq \beta_1 + \|y_k - y_*\| + \|\Delta y_k^*\| \\ &\equiv \beta_2, \end{aligned}$$

and so part 1 of Lemma 5.2 is proved.

Now suppose $\|\Delta y_k^*\| \rightarrow 0$ as $k \in \mathcal{K}$ goes to infinity. Because $\{x_k^*\}$ and $\{x_k\}$ are in the compact set $\mathcal{B}$, $\|x_k^* - x_k\|$ is bounded. From (5.2) and the fact that $x_k^* \rightarrow x_*$ and $\omega_k \rightarrow 0$, we conclude that $\widetilde{y}_k \rightarrow y_*$ as $k \in \mathcal{K}$ goes to infinity. We also conclude from the continuity of J on the compact set $\mathcal{B}$ that $\|J(x_k^*) - J_k\|$ is bounded, so that

$$(5.13) \quad \lim_{k \in \mathcal{K}} \|(J(x_k^*) - J_k)^T \Delta y_k^*\| = 0.$$

On the other hand, suppose $\|\Delta y_k^*\|$ is uniformly bounded and $\lim_{k \in \mathcal{K}} \|x_k^* - x_k\| = 0$. We then conclude from (5.2) that $\widetilde{y}_k \rightarrow y_*$ as $k \in \mathcal{K}$ goes to infinity and again (5.13) holds. Because $\lim_{k \in \mathcal{K}} (x_k^*, \widetilde{y}_k) = (x_*, y_*)$, we have

$$\lim_{k \in \mathcal{K}} g(x_k^*) - J(x_k^*)^T \widetilde{y}_k = g_* - J_*^T y_*,$$

and so (5.5) and (5.13) together imply that $\lim_{k \in \mathcal{K}} z_k^* = z_* \equiv \nabla_x \mathcal{L}(x_*, y_*, 0)$. Thus we have proved part 2 of Lemma 5.2.

Now note that each (x_k^*, z_k^*) satisfies (3.3d), so that $\lim_{k \in \mathcal{K}} (x_k^*, z_k^*) = (x_*, z_*)$ and $\omega_k \rightarrow 0$ together imply $\min(x_*, z_*) = 0$. Hence if $c_* = 0$ also, (x_*, y_*, z_*) satisfies (1.4) and so is a first-order KKT point for (GNP). Part 3 is thus proved. $\square$

The conclusions of Lemma 5.2 pertain to any sequence $\{(x_k^*, \Delta y_k^*, z_k^*)\}$ satisfying the approximate first-order conditions (3.3). Algorithm 2 generates such a sequence and also generates auxiliary sequences of scalars $\{\omega_k\}$ and $\{\rho_k\}$ in such a way as to guarantee that the hypotheses of Lemma 5.2 hold. We demonstrate in Theorem 5.4 below that the condition of part 3 of Lemma 5.2 holds. Therefore, every limit point of $\{(x_k^*, \widetilde{y}_k, z_k^*)\}$ is a first-order KKT point for (GNP).

5.2. Convergence of $\|y_k\|/\rho_k$. Before stating and proving the global convergence properties of the sLCL algorithm, we need to show that if $\rho_k \rightarrow \infty$, then the quotient $\|y_k\|/\rho_k$ converges to 0. This property is required (and used by Conn et al. [9] and by Conn, Gould, and Toint [10]) in lieu of assuming that $\|y_k\|$ remains bounded.

LEMMA 5.3. Suppose $\rho_k \rightarrow \infty$ as k increases when Algorithm 2 is executed. Then $\|y_k\|/\rho_k \rightarrow 0$.

Proof. Define the sequence $\{k_1, k_2, k_3, \dots\}$ as the set of iterations that execute step 11 of Algorithm 2, and let $\{l_1, l_2, l_3, \dots\}$ be the associated sequence of iterations for which step 5 was *last* executed. With these definitions,

$$(5.14) \quad y_{k_i+1} = \tilde{y}_{l_i} \equiv \tilde{y}(x_{l_i}^*, y_{l_i} + \Delta y_{l_i}^*, \rho_{l_i}) \quad \text{and} \quad \rho_{k_i+1} = \tau_\rho^{k_i - l_i} \rho_{l_i}.$$

Because the parameter ρ_k increases if and only if step 11 is executed, it is sufficient to show that $\|y_{k_i+1}\|/\rho_{k_i+1} \rightarrow 0$ as $i \rightarrow \infty$.

Suppose there exists a maximum $\bar{l}$ over $\{l_1, l_2, l_3, \dots\}$. (Set $\bar{l} = 0$ if the set is empty.) Then step 5 is executed only finitely many (or zero) times, and $y_k \equiv y_{\bar{l}}$ for all k large enough. Therefore, $\|y_k\|/\rho_k \rightarrow 0$, as required. Assume for the remainder that $\{l_1, l_2, l_3, \dots\}$ is not empty and $l_i \rightarrow \infty$ as $i \rightarrow \infty$, implying that step 5 is executed infinitely many times.

The multiplier update used in step 5, together with the triangle inequality, (5.14), and the fact that $\tau_\rho^{(k_i - l_i)} \geq 1$, implies that for each k_i ,

$$(5.15) \quad \begin{aligned} \frac{\|y_{k_i+1}\|}{\rho_{k_i+1}} &= \frac{\|\tilde{y}_{l_i}\|}{\tau_\rho^{(k_i - l_i)} \rho_{l_i}} \leq \frac{\|\tilde{y}_{l_i}\|}{\rho_{l_i}} = \frac{\|y_{l_i} + \Delta y_{l_i}^* - \rho_{l_i} c(x_{l_i}^*)\|}{\rho_{l_i}} \\ &\leq \underbrace{\frac{\|\Delta y_{l_i}^*\|}{\rho_{l_i}}}_{(a)} + \underbrace{\frac{\|y_{l_i} - \rho_{l_i} c(x_{l_i}^*)\|}{\rho_{l_i}}}_{(b)}. \end{aligned}$$

As discussed in section 4, $\|\Delta y_k^*\|$ is uniformly bounded, so that (a) goes to zero as $\rho_{l_i} \rightarrow \infty$. Also, the definition of the forcing sequence η_k (steps 9 and 13) is analogous to that in [10, Algorithm 1] so that [10, Lemma 4.1] applies and shows that (b) goes to zero as $\rho_{l_i} \rightarrow \infty$. Thus, (5.15) implies that $\|y_{k_i+1}\|/\rho_{k_i+1} \rightarrow 0$. $\square$

5.3. Convergence of LC subproblem solutions. With Lemmas 5.2 and 5.3 in hand, we are now able to prove global convergence of the sLCL method under the assumption that *only a single limit point exists*. At the end of this section we discuss a modification of Algorithm 2 that guarantees global convergence even when multiple limit points exist.

THEOREM 5.4 (global convergence of subproblem solutions). Suppose Assumptions 1.3 and 1.4 hold. Let $\{(x_k^*, \Delta y_k^*, z_k^*)\}$ be the sequence of vectors generated by Algorithm 2 with tolerances $\omega_* = 0$ and $\eta_* = 0$. Let x_* be the single limit point of $\{x_k^*\}$. Then all parts of Lemma 5.2 hold. In other words, (x_*, y_*, z_*) as defined in Lemma 5.2 is a first-order KKT point.

Proof. Algorithm 2 generates positive scalars ρ_k and, by step 1, generates positive scalars $\omega_k \rightarrow 0$. Step 2 of the algorithm generates a sequence $\{(x_k^*, \Delta y_k^*, z_k^*)\}$ that satisfies (3.3) for each k . Therefore, the hypotheses of Lemma 5.2 hold, and part 1 of the lemma follows immediately.

Note that each x_k^* satisfies (3.3a), and so all $x_k^* \geq 0$. Thus, $x_* \geq 0$. Moreover, σ_k is uniformly bounded. We then need to consider the four possible cases:

1. ρ_k is uniformly bounded, and $\sigma_k \rightarrow 0$;
2. ρ_k is uniformly bounded, and $\sigma_k \nrightarrow 0$;
3. $\rho_k \rightarrow \infty$ and $\sigma_k \rightarrow 0$;
4. $\rho_k \rightarrow \infty$ and $\sigma_k \nrightarrow 0$.

We dismiss cases 1 and 4 because they cannot be generated by the algorithm. (As k gets large, $\sigma_k \rightarrow 0$ only if step 11 is executed infinitely many times, contradicting the finiteness of ρ_k . Conversely, if $\rho_k \rightarrow \infty$, then steps 8 and 12 ensure that $\sigma_k \rightarrow 0$.)

Case 2 implies that step 4 of Algorithm 2 is executed for all k large enough. Thus, $x_{k+1} = x_k^*$ for all large k , and hence $x_k^* \rightarrow x_*$ implies $x_{k+1} \rightarrow x_*$. The “single limit point” assumption implies $x_k \rightarrow x_*$. Therefore, $\|x_k^* - x_k\| \rightarrow 0$. Because $\|\Delta y_k^*\|$ is bounded (see section 4), part 2 of Lemma 5.2 holds. In addition, $\eta_k \rightarrow 0$ because $\rho_k > 0$ and step 9 is executed for all k large enough. Therefore, the condition $\|c(x_k^*)\| \leq \eta_k$ for all k large enough implies that $c(x_k^*) \rightarrow 0$. By continuity of c , $c_* = 0$. Thus, part 3 of Lemma 5.2 holds.

Now consider case 3. Because $\sigma_k \rightarrow 0$ and $\omega_k \rightarrow 0$, (3.4) implies that $\|\Delta y_k^*\| \rightarrow 0$ as k increases. Then part 2 of the lemma holds. To show that $c(x_k^*) \rightarrow 0$, divide both sides of (5.3) by ρ_k to obtain

$$\|c(x_k^*)\| \leq \underbrace{\frac{\alpha_1 \omega_k}{\rho_k}}_{(a)} + \underbrace{\frac{1}{\rho_k} \|\Delta y_k^*\| (M \|x_k^* - x_k\| + 1)}_{(b)} + \underbrace{\frac{\alpha_2 \|x_k^* - x_*\|}{\rho_k}}_{(c)} + \underbrace{\frac{1}{\rho_k} \|y_k - y_*\|}_{(d)}.$$

Term (a) clearly goes to zero as ρ_k increases. Because Δy_k^* satisfies (3.4), and because x_k^* and x_k belong to the compact set $\mathcal{B}$, (b) and (c) go to zero as ρ_k increases. By Lemma 5.3, $\|y_k\|/\rho_k \rightarrow 0$, and so (d) goes to 0. We conclude that $\|c(x_k^*)\| \rightarrow 0$ as k increases, as required. $\square$

Note that the “single limit point” assumption in Theorem 5.4 is needed only in case 2 to ensure that $\|x_k^* - x_k\| \rightarrow 0$, as required by part 2 of Lemma 5.2. (If multiple limit points exist, we can only ensure that $\|x_k^* - x_{k+1}\| \rightarrow 0$.) To allow for multiple limit points, step 8 of Algorithm 2 could be replaced by the following update:

```

8a if  $\|\Delta y_k^*\|_\infty \leq \delta(\frac{1}{2})^j$  then
     $\sigma_{k+1} \leftarrow \frac{1}{1+\rho_k} \min(1 + \|\Delta y_k^*\|_\infty, \bar{\sigma})$  [reset  $\sigma_k$ ]
else
8b  $\sigma_{k+1} \leftarrow \sigma_k / \tau_\sigma$  [decrease  $\sigma_k$ ]
end

```

(The counter j is the number of consecutive successful major iterations, and δ is a positive parameter.) We now consider case 2 under the alternative update. If ρ_k is uniformly bounded, then all iterations are successful for k large enough, and $j \rightarrow \infty$. Then if Δy_k^* does not satisfy step 8a infinitely often, $\sigma_k \rightarrow 0$, and (3.4) implies that $\|\Delta y_k^*\| \rightarrow 0$. Then part 2 of Lemma 5.2 holds, as required.

Any forcing sequence converging to zero could be used in step 8a. However, requiring only a gentle decrease in $\|\Delta y_k^*\|_\infty$ at each iteration should interfere less with the fast local convergence of the sLCL method, because $\|\Delta y_k^*\|$ may be expected to decrease at a linear rate, as noted in Theorem 6.2, (6.6).

5.4. Termination in the limit. Note that the convergence test takes place only if step 3 of Algorithm 2 tests true; i.e., if $\|c(x_k^*)\| \leq \eta_k$ (because $\eta_* = 0$). To guarantee that the algorithm will eventually terminate as the iterates x_k , y_k , and z_k converge, we need to guarantee that step 7 executes infinitely often. The forcing sequence η_k is intimately tied to this occurrence. For example, if $\eta_k \equiv 0$, then we would not normally expect step 3 to evaluate true (except in rare occasions when $c(x_k^*) = 0$). The forcing sequence defined by steps 9 and 13 of Algorithm 2 is suggested by Conn et al. [9]

and Conn, Gould, and Toint [10]. The following corollaries show that this forcing sequence has the desired property and summarize the global convergence properties of Algorithm 2.

COROLLARY 5.5 (global convergence with a single limit point). *Let $\{(x_k, y_k, z_k)\}$ be the sequence of vectors generated by Algorithm 2. Let x_* be the single limit point of $\{x_k^*\}$. Suppose Assumptions 1.3 and 1.4 hold. Then*

$$\lim_{k \rightarrow \infty} (x_k, y_k, z_k) = (x_*, y_*, z_*),$$

and (x_*, y_*, z_*) is a first-order KKT point for (GNP).

Proof. Let $\{(x_k^*, \Delta y_k^*, z_k^*)\}$ be the sequence of vectors generated by step 2 of Algorithm 2 and set $\tilde{y}_k = \tilde{y}(x_k^*, y_k + \Delta y_k^*, \rho_k)$. By Lemma 5.2 and Theorem 5.4,

$$\lim_{k \rightarrow \infty} \tilde{y}_k = y_* \quad \text{and} \quad \lim_{k \rightarrow \infty} z_k^* = z_*.$$

Moreover, (x_*, y_*, z_*) is a first-order KKT point for (GNP). Suppose steps 4–6 are executed infinitely often. The result then follows immediately because x_k , y_k , and z_k are updated infinitely often and form a convergent sequence from x_k^* , $\tilde{y}_k$, and z_k^* .

We now show by contradiction that steps 4–6 do occur infinitely often. Suppose instead that they do not. Then there exists a k_1 large enough so that steps 10 and 11 are executed for all $k > k_1$. Consider only iterations $k > k_1$. Then $y_k \equiv \bar{y}$ and $\rho_k \rightarrow \infty$. As in (5.12), the definition of $\tilde{y}_k$ gives

$$(5.16) \quad \begin{aligned} \rho_k \|c(x_k^*)\| &= \|\bar{y} + \Delta y_k^* - \tilde{y}_k\| \\ &\leq \|\Delta y_k^*\| + \|\bar{y}\| + \|\tilde{y}_k\|. \end{aligned}$$

Each Δy_k^* satisfies $\|\Delta y_k^*\|_\infty \leq \sigma_k + \omega_k$ (3.4) with $\sigma_k + \omega_k$ uniformly bounded. Moreover, $\lim_{k \rightarrow \infty} \tilde{y}_k = y_*$ and y_* is bounded. (Assumption 1.4 ensures that the least-squares solution of (5.1) exists and is unique.) Then from (5.16) there exists some constant $L > 0$, independent of k , such that

$$(5.17) \quad \rho_k \|c(x_k^*)\| \leq L$$

for all k . But the test at step 3 fails at every iteration so that

$$(5.18) \quad \eta_k < \|c(x_k^*)\|.$$

Combining (5.17) and (5.18), we find that

$$(5.19) \quad \rho_k \eta_k < \rho_k \|c(x_k^*)\| \leq L.$$

From step 13, $\eta_{k+1} = \eta_0 / \rho_{k+1}^\alpha$, so

$$(5.20) \quad \rho_k \eta_k = \rho_k \frac{\eta_0}{\rho_k^\alpha} = \eta_0 \rho_k^{1-\alpha}.$$

Substituting (5.20) into (5.19), we find that $\eta_0 \rho_k^{1-\alpha} < L$ for all k . This is a contradiction under the hypothesis that $\alpha < 1$ and $\rho_k \rightarrow \infty$. Therefore, steps 4–6 must occur infinitely often. $\square$

5.5. Finite termination. The following result simply asserts that Algorithm 2 will eventually exit when ω_* and η_* are positive, as they are in practice.

COROLLARY 5.6 (finite termination). *Suppose the convergence tolerances ω_* and η_* are strictly positive. Then, under the conditions of Corollary 5.5, Algorithm 2 terminates after a finite number of iterations.*

Proof. Let $\{(x_k^*, \Delta y_k^*, z_k^*)\}$ and x_* be as in Theorem 5.4. Set $\tilde{y}_k = \tilde{y}(x_k^*, y_k + \Delta y_k^*, \rho_k)$. (This is y_{k+1} in step 5.) By Theorem 5.4,

$$\begin{aligned} \lim_{k \rightarrow \infty} \tilde{y}_k &= y_*, \\ \lim_{k \rightarrow \infty} z_k^* &= z_* \stackrel{\text{def}}{=} \nabla_x \mathcal{L}(x_*, y_*, 0), \end{aligned}$$

and (x_*, y_*, z_*) is a first-order KKT point for (GNP). By continuity, $\lim_{k \rightarrow \infty} \|c(x_k^*)\| = c_* = 0$, and hence $\|c(x_k^*)\| < \eta_* \leq \max(\eta_k, \eta_*)$ for all k large enough. Consequently, steps 4–6 are executed infinitely often and

$$\lim_{k \rightarrow \infty} (x_k, y_k, z_k) \rightarrow (x_*, y_*, z_*).$$

Therefore, (x_k, y_k, z_k) satisfies conditions (1.6) for some k large enough. $\square$

5.6. Infeasible problems. Not all optimization problems are well defined. The user of an optimization algorithm may formulate a set of nonlinear constraints $c(x) = 0$ for which no nonnegative solution exists. Detecting infeasibility of a system $c(x) = 0$, $x \geq 0$, is equivalent to verifying that the *global* minimizer of

$$(5.21) \quad \begin{aligned} &\underset{x}{\text{minimize}} && \frac{1}{2} \|c(x)\|^2 \\ &\text{subject to} && x \geq 0 \end{aligned}$$

yields a positive objective value. Detecting such infeasibility is a useful feature, but it is a very difficult problem and is beyond the purview of this paper.

We analyze the properties of the sLCL algorithm when it is applied to an infeasible problem with convergence tolerances $\omega_* = \eta_* = 0$. We show that Algorithm 2 converges to a point that satisfies the first-order optimality conditions of the minimum-norm problem (5.21).

THEOREM 5.7 (convergence for infeasible problems). *Suppose Assumption 1.3 holds. Let $(x_*, \mathcal{K})$ be a limit point of the sequence $\{x_k^*\}$ generated by Algorithm 2. Suppose (GNP) is infeasible. Then*

$$\lim_{k \in \mathcal{K}} J(x_k^*)^T c(x_k^*) = z_* \stackrel{\text{def}}{=} J_*^T c_*,$$

and (x_*, z_*) is a first-order KKT point for (5.21).

Proof. The pair (x_*, z_*) satisfies the first-order KKT conditions of (5.21) if

$$(5.22) \quad J_*^T c_* = z_* \quad \text{and} \quad \min(x_*, z_*) = 0.$$

Because (GNP) is infeasible, there exists a constant $\delta > 0$ such that $\delta < \|c(x)\|$ for all $x \geq 0$. Moreover, steps 9 and 13 of Algorithm 2 generate a sequence $\{\eta_k\}$ converging to 0, and so $\eta_k < \delta$ for all k large enough. Consider only such k . Then, $\eta_k < \delta < \|c(x_k^*)\|$, and step 11 is executed at every k so that $\rho_k \rightarrow \infty$ and $\sigma_k \rightarrow 0$. Moreover, x_k and y_k are not updated so that for some n -vector $\bar{x}$ and m -vector $\bar{y}$,

$$(5.23) \quad x_k \equiv \bar{x} \quad \text{and} \quad y_k \equiv \bar{y}.$$

Note that Algorithm 2 generates x_k^* satisfying (3.3). Therefore, $x_k^* \geq 0$ for all k , and so $\lim_{k \in \mathcal{K}} x_k^* = x_*$ implies

$$(5.24) \quad x_* \geq 0.$$

As in (5.5), and using (3.3d) and (5.23), we have

$$g(x_k^*) - J(x_k^*)^T(\bar{y} - \rho_k c(x_k^*)) - J(\bar{x})^T \Delta y_k^* \geq -\omega_k e,$$

or, after rearranging terms,

$$(5.25) \quad \underbrace{g(x_k^*) - J(x_k^*)^T \bar{y}}_{(a)} - \underbrace{J(\bar{x})^T \Delta y_k^*}_{(b)} + \rho_k J(x_k^*)^T c(x_k^*) \geq -\omega_k e.$$

By hypothesis, all iterates x_k^* lie in a compact set, and so (a) is bounded because g and J are continuous and $\bar{y}$ is constant. Also, (b) is bounded because $\bar{x}$ is constant, and from (3.4) we have $\|\Delta y_k^*\|_\infty \leq \sigma_k + \omega_k$. Then, because $\omega_k \rightarrow 0$ and $\rho_k \rightarrow \infty$, (5.25) implies that $J(x_k^*)^T c(x_k^*) \geq 0$ for all k large enough. Otherwise, (5.25) would eventually be violated as ρ_k grew large. Then,

$$(5.26) \quad z_* \stackrel{\text{def}}{=} \lim_{k \in \mathcal{K}} J(x_k^*)^T c(x_k^*) = J_*^T c_* \geq 0.$$

Because all x_k^* lie in a compact set, there exists some constant $L > 0$ such that

$$(5.27) \quad \|x_k^* - \bar{x}\| \leq \frac{L\alpha_1}{\sqrt{n}M},$$

where M and α_1 are as defined in Lemma 5.2. Expression (5.8) can be derived again under the assumptions of Theorem 5.7 (note that Assumption 1.4 is not required). Substituting (5.27) into (5.8) and using (3.4) and $\tilde{y}_k \equiv y_k + \Delta y_k^* - \rho_k c(x_k^*)$, we have

$$(5.28) \quad \|\hat{g}(x_k^*) - \hat{J}(x_k^*)^T(\bar{y} + \Delta y_k^*) + \rho_k \hat{J}(x_k^*)^T c(x_k^*)\| \leq \sqrt{n}\{\omega_k + \hat{L}(\sigma_k + \omega_k)\},$$

where $\hat{L} > 0$ accounts for the ∞ -norm in (3.4). Dividing (5.28) through by ρ_k , we obtain

$$(5.29) \quad \left\| \frac{1}{\rho_k} (\hat{g}(x_k^*) - \hat{J}(x_k^*)^T(\bar{y} + \Delta y_k^*)) + \hat{J}(x_k^*)^T c(x_k^*) \right\| \leq \frac{\sqrt{n}\{\omega_k + \hat{L}(\sigma_k + \omega_k)\}}{\rho_k}.$$

The quantity $\hat{g}(x_k^*) - \hat{J}(x_k^*)^T(y_k + \Delta y_k^*)$ is bounded for the same reasons that (a) and (b) above are bounded. Taking limits on both sides of (5.29), we see that $\rho_k \rightarrow \infty$ and $\omega_k, \sigma_k \rightarrow 0$ imply $\hat{J}(x_k^*)^T c(x_k^*) \rightarrow 0$. By continuity of J and c , $\hat{J}_*^T c_* = 0$. Equivalently, we may write

$$(5.30) \quad [J_*^T c_*]_j = 0 \quad \text{if} \quad [x_*]_j > 0, \quad j = 1, \dots, n.$$

Together, (5.24), (5.26), and (5.30) imply that (x_*, z_*) satisfies conditions (5.22), as required. $\square$

Theorem 5.7 describes a useful feature of Algorithm 2. When applied to an infeasible problem, the algorithm converges to a solution of (5.21), or at least to a first-order point. One important caveat deserves mention: if the convergence tolerance η_* is small (it usually will be), Algorithm 2 may never terminate. We need to insert an additional test to provide for the possibility that (GNP) is infeasible. For example, the test could force the algorithm to exit if ρ_k is above a certain threshold value and $\|c(x_k^*)\|$ is no longer decreasing. Any test we devise is necessarily heuristic, however; it is impossible to know for certain whether a larger value of ρ_k would force $\|c(x_k^*)\|$ to be less than η_* . We discuss this point further in section 7.6.

5.7. Second-order optimality. The sLCL method imposes few requirements on the manner in which the LC subproblems are solved. Our implementation (see section 7) uses MINOS or SNOPT to solve the LC subproblems. These are active-set solvers suitable for optimization problems with few expected degrees of freedom at the solution and in which only first derivatives are available. However, second derivatives might be readily available for some problems. Also, some problems are expected to have many degrees of freedom at the solution. In either case, an interior-point solver (requiring second derivatives) may be more appropriate for the solution of the subproblems.

Lemma 5.2 and Theorem 5.4 assert that the sLCL iterates converge to first-order KKT points. A subproblem solver that uses second derivatives may be able to guarantee convergence to second-order points. If we augment the convergence criteria for the solution of each subproblem to include second-order conditions, we can show that Algorithm 2 generates iterates converging to points satisfying the second-order sufficiency conditions for (GNP).

DEFINITION 5.8 (strict complementarity). *The first-order point (x_*, y_*, z_*) satisfies strict complementarity if $\max(x_*, z_*) > 0$.*

DEFINITION 5.9 (second-order sufficiency). *The first-order point (x_*, y_*, z_*) satisfies the second-order sufficiency conditions for (GNP) if it satisfies strict complementarity and if for $\rho = 0$,*

$$(5.31) \quad p^T \nabla_{xx}^2 \mathcal{L}(x_*, y_*, \rho) p > 0$$

for all $p \neq 0$ satisfying

$$(5.32) \quad J(x_*)p = 0 \text{ and } [p]_j = 0 \text{ for all } j \text{ such that } [x_*]_j = 0 \text{ (and } [z_*]_j > 0).$$

Again, second-order sufficiency is normally defined in terms of the Lagrangian, not the augmented Lagrangian used in Definition 5.9, but if (5.31) holds for $\rho = 0$, then it certainly holds for all $\rho \geq 0$.

The following assumption strengthens the first-order conditions (3.1).

ASSUMPTION 5.10. *Let $\{(x_k^*, \Delta y_k^*, z_k^*)\}$ be a sequence of local solutions to subproblems (ELC_k), and let $(x_*, \mathcal{K})$ be a limit point of $\{x_k^*\}$. For all $k \in \mathcal{K}$ large enough, the following conditions hold at each solution for some $\delta > 0$ independent of k :*

1. (Strict complementarity)

$$\max(x_k^*, z_k^*) > \delta e;$$

2. (Second-order condition)

$$(5.33) \quad p^T \nabla_{xx}^2 \mathcal{L}(x_k^*, y_k + \Delta y_k^*, \rho_k) p > \delta \|p\|^2$$

for all $p \neq 0$ satisfying

$$(5.34) \quad J(x_k^*)p = 0 \text{ and } [p]_j = 0 \text{ for all } j \text{ such that } [x_k^*]_j = 0.$$

Condition (5.33) implies that the reduced Hessian of $\mathcal{L}$ is uniformly positive definite at all x_k^* .

The following result extends the global analysis of section 5.3 to the case in which the iterates generated by Algorithm 2 satisfy Assumption 5.10. Conn, Gould, and Toint [10] show a similar result for their BCL method.

THEOREM 5.11 (convergence to a local minimizer). *Let $\{(x_k^*, \Delta y_k^*, z_k^*)\}$ be a sequence satisfying Assumption 5.10. Let $(x_*, \mathcal{K})$ be a limit point of $\{x_k^*\}$, set $\tilde{y}_k = \tilde{y}(x_k^*, y_k + \Delta y_k^*, \rho_k)$, and let*

$$(5.35) \quad (x_*, y_*, z_*) = \lim_{k \in \mathcal{K}} (x_k^*, \tilde{y}_k, z_k^*)$$

be a first-order KKT point of (GNP). Then (x_*, y_*, z_*) is an isolated local minimizer.

Proof. By hypothesis, x_k^* and z_k^* satisfy part 1 of Assumption 5.10 for all $k \in \mathcal{K}$. Therefore, their limit points satisfy $\max(x_*, z_*) \geq \delta e > 0$, and so x_* and z_* satisfy strict complementarity (Definition 5.8). Now for each $k \in \mathcal{K}$ large enough, let p be a nonzero vector satisfying (5.34). Then

$$(5.36) \quad p^T \nabla_{xx}^2 \mathcal{L}(x_k^*, y_k + \Delta y_k^*, \rho_k) p = p^T \left(H(x_k^*) - \sum_{i=1}^m [\tilde{y}_k]_i H_i(x_k^*) \right) p.$$

Part 2 of Assumption 5.10 and (5.36) imply that

$$(5.37) \quad p^T \left(H(x_k^*) - \sum_{i=1}^m [\tilde{y}_k]_i H_i(x_k^*) \right) p > \delta \|p\|^2,$$

where δ is some positive constant. If we take the limit of (5.37), the continuity of H and H_i (Assumption 1.2) and (5.35) imply that for $\rho = 0$,

$$p^T \nabla_{xx}^2 \mathcal{L}(x_*, y_*, \rho) p \equiv p^T \left(H(x_*) - \sum_{i=1}^m [y_*]_i H_i(x_*) \right) p \geq \delta \|p\|^2 > 0$$

for all $p \neq 0$ satisfying (5.32). Therefore, (x_*, y_*, z_*) satisfies the second-order sufficiency conditions for (GNP), as required. $\square$

Note that section 5.3 (with two options for step 8) guarantees that the limit (5.35) is a first-order KKT point whether or not there are multiple limit points.

6. Local convergence. In this section we show that the sLCL algorithm preserves the local convergence characteristics of Robinson's original LCL algorithm. Moreover, it can retain fast local convergence under inexact solutions to the subproblems.

Bertsekas [1], Conn et al. [9], and Conn, Gould, and Toint [10] show how to construct a forcing sequence $\{\eta_k\}$ to guarantee that $\|c(x_k^*)\| \leq \eta_k$ will eventually always be true so that the iterates x_k , y_k , and z_k are updated (see steps 4–6 of Algorithm 2) for all iterations after some k large enough. The penalty parameter ρ_k then remains uniformly bounded—an important property. These results rely on a relationship between $\|c(x_k^*)\|$ and ρ_k , namely, (5.3). We know from BCL convergence theory that the convergence rate is superlinear if $\rho_k \rightarrow \infty$ and linear otherwise (cf. [1, 9, 10]). Because η_k is reduced at a *sublinear* rate, $\|c(x_k^*)\|$ will eventually go to zero faster than η_k , at which point it is no longer necessary to increase ρ_k . Thus, we can be assured that Algorithm 2 does not increase ρ_k without bound.

Bertsekas [1] and Powell [39] suggest constructing the sequence η_k as

$$(6.1) \quad \eta_{k+1} = \zeta \|c(x_k^*)\|$$

for some $\zeta < 1$. Within Algorithm 2, this would lead to the following update rule:

$$(6.2) \quad \rho_{k+1} = \begin{cases} \rho_k & \text{if } \|c(x_k^*)\| \leq \zeta \|c(x_k)\|, \\ \tau_\rho \rho_k & \text{if } \|c(x_k^*)\| > \zeta \|c(x_k)\|. \end{cases}$$

As ρ_k gets larger, the convergence rate becomes arbitrarily close to superlinear, so that the first case of (6.2) is always satisfied, and ρ_k becomes constant for all k large enough. We prefer not to use rule (6.1) because it may be too strict. Any intermediate (and nonoptimal) iterate x_k^* could be feasible or nearly feasible for (GNP), so that $\|c(x_k^*)\|$ could be very small. Then η_{k+1} would be smaller than warranted on the following iteration. The forcing sequence suggested by Conn et al. [9] and Conn, Gould, and Toint [10] does not suffer from this defect and has been proved by them to keep ρ_k bounded. We have used this update in Algorithm 2 (see steps 9 and 13).

For this analysis and the remainder of this section, we assume that ρ_k is uniformly bounded so that $\rho_k = \bar{\rho}$ for all k greater than some $\bar{k}$. Hence, we drop the subscript on ρ_k and simply write $\bar{\rho}$. We consider only iterations $k > \bar{k}$.

The local convergence analysis requires the following second-order condition.

ASSUMPTION 6.1. *The point (x_*, y_*, z_*) satisfies the second-order sufficiency conditions for (GNP).*

We begin by discussing the local convergence rates of Algorithm 2 under the assumption that the elastic variables are always zero—that is, the linearized constraints are always satisfied. Next, we show that after finitely many iterations the elastic penalty parameter σ_k will always be large enough to guarantee that this assumption holds. In this way, we demonstrate that sLCL becomes equivalent to canonical LCL (and to MINOS) as it approaches the solution.

6.1. Convergence rate. Robinson's [40] local convergence analysis applies to the canonical LCL algorithm under the special case in which $\rho_k \equiv 0$ and each subproblem is solved to full accuracy (i.e., $\omega_k \equiv 0$). He proved that one can expect fast convergence from a good enough starting point. In particular, under Assumptions 1.2, 6.1, and 1.4, we can expect an R-quadratic rate of convergence. (See Ortega and Rheinboldt [38] for an in-depth discussion of root-convergence rates.) For a sufficiently good starting point, Robinson [41] proves that the subproblems (LC_k) are always feasible. He also shows that near a solution, the solutions to the linearly constrained subproblems, if parameterized appropriately, form a continuous path converging to (x_*, y_*, z_*) .

In a later paper, Bräuninger [6] shows how the fast local convergence rate can be preserved with only *approximate* solutions of the subproblems (again, with $\rho_k \equiv 0$). The subproblems are solved to a tolerance that is tightened at a rate that matches the decrease in the square of the primal and dual infeasibilities. Our proposed LCL algorithm uses a similar strategy.

Robinson's local convergence analysis also applies to the canonical LCL algorithm when $\rho_k \equiv \bar{\rho} > 0$. One can see this by considering the optimization problem

$$(6.3) \quad \begin{aligned} & \underset{x}{\text{minimize}} && f(x) + \frac{1}{2}\bar{\rho}\|c(x)\|^2 \\ & \text{subject to} && c(x) = 0, \quad x \geq 0, \end{aligned}$$

and applying Robinson's method. The solutions of (6.3) are identical to the solutions of (GNP). The Robinson LCL subproblem objective for problem (6.3) is

$$R_k(x) \equiv f(x) + \frac{1}{2}\bar{\rho}\|c(x)\|^2 - y_k^T c(x).$$

The canonical LCL subproblem objective is $\mathcal{L}_k(x) \equiv \mathcal{L}(x, y_k, \rho_k)$, and so $\mathcal{L}_k(x) \equiv R_k(x)$ for all k because $\rho_k \equiv \bar{\rho}$. We then observe that the canonical LCL subproblem corresponding to (GNP), with a penalty parameter $\rho_k \equiv \bar{\rho}$, is *equivalent* to the

Robinson LCL subproblem corresponding to problem (6.3), with $\rho_k \equiv 0$. The convergence characteristics of the canonical LCL algorithm are therefore the same as those demonstrated by Robinson [40]. (However, while the asymptotic convergence rate remains R-quadratic, we expect a different asymptotic error constant.)

Under the assumption that the elastic variables are always equal to 0 and that $\bar{\rho}$ is finite, the steps executed by Algorithms 1 and 2 are identical, and the subproblems (ELC_k) and (LC_k) are also identical. The only difference is the multiplier update formulas

$$(6.4a) \quad \text{Canonical LCL update} \quad y_{k+1} = y_k + \Delta y_k^*,$$

$$(6.4b) \quad \text{Stabilized LCL update} \quad y_{k+1} = y_k + \Delta y_k^* - \bar{\rho}c(x_k^*),$$

which differ by the vector $\bar{\rho}c(x_k^*)$. We may think of this vector as a perturbation of the LCL multiplier update (6.4a). Moreover, Robinson [40] shows that this perturbation converges to 0 at the same rate as $\{x_k^*\}$ converges to x_* . Therefore, it does not interfere with the convergence rate of the sLCL iterates. Robinson's local convergence analysis then applies to the sLCL method.

We summarize the convergence results in Theorem 6.2. Note that the function

$$F(x, y, z) = \begin{bmatrix} c(x) \\ \nabla_x \mathcal{L}(x, y, \rho) - z \\ \min(x, z) \end{bmatrix}$$

captures the first-order optimality conditions of (GNP), in the sense that the vector $F(x_*, y_*, z_*) = 0$ if and only if (x_*, y_*, z_*) is a first-order KKT point for (GNP). Thus, $\|F(x, y, z)\|$ is a measure of the deviation from optimality. For the next theorem only, define

$$r = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \quad \text{and} \quad F(r) = F(x, y, z).$$

THEOREM 6.2 (Robinson [40]; Bräuninger [6]). *Suppose Assumptions 1.4 and 6.1 hold at r_* . Moreover, suppose $\omega_k = O(\|F(r_k)\|^2)$ for all $k \geq 0$. Then there is a positive constant δ such that if $\|r_0 - r_*\| < \delta$, the sequence $\{r_k\}$ generated by Algorithm 2 converges to r_* . Moreover, the sequence converges R-quadratically, so that for all $k \geq 0$,*

$$(6.5) \quad \|r_k - r_*\| \leq L\left(\frac{1}{2}\right)^{2^k}$$

for some positive constant L . Also, the difference between consecutive iterates converges at an R-linear rate:

$$(6.6) \quad \|r_{k+1} - r_k\| \leq \delta\left(\frac{1}{2}\right)^k.$$

Robinson does not state (6.6) as part of a theorem, but it is found in the proof of (6.5).

6.2. Asymptotic equivalence to canonical LCL. Much of the efficiency of LCL methods, including MINOS, derives from the fact that they eventually identify the correct active set, and each subproblem restricts its search to the subspace defined by a linear approximation of the constraints. This approximation can be very accurate

near the solution. The sLCL subproblems do *not* restrict themselves to this subspace. In early iterations we do not expect, nor do we wish, the method to honor these linearizations. The elastic variables give the subproblems an opportunity to deviate from this subspace. In order to recover LCL's fast convergence rate, however, it is not desirable to allow deviation near the solution.

As we show in the next theorem, the elastic variables v_k^* and w_k^* will be bounded by the parameter ω_k that controls the termination of the subproblems. In practice, we might choose to set $\omega_k \equiv \omega_*$ after some number of iterations, in effect asking for accurate subproblem solutions for all remaining iterates. Hence, x_k^* will eventually always satisfy the linearized constraints to within the specified tolerance ω_* .

THEOREM 6.3 (convergence to inelastic subproblem solutions). *Let $\{(x_k, y_k, z_k)\}$ and $\{(v_k^*, w_k^*)\}$ be the sequences generated by Algorithm 2. Let x_* be the single limit point of $\{x_k^*\}$. Suppose Assumption 1.4 holds and the sequence $\{\rho_k\}$ remains bounded. Then for all k large enough, $\|v_k^*\|_\infty \leq \omega_k$ and $\|w_k^*\|_\infty \leq \omega_k$.*

Proof. The assumed boundedness of ρ_k implies that $\rho_k \equiv \bar{\rho}$ for all k large enough. Consider only such k for the remainder of the proof. Set $\underline{\sigma} = 1/(1 + \bar{\rho})$. Because $\bar{\sigma} > 1$ and $\min(1 + \|\Delta y_k^*\|_\infty, \bar{\sigma}) > 1$, step 8 of Algorithm 2 guarantees that $\underline{\sigma} \leq \sigma_k$ for all k .

The hypotheses of the theorem are the same as those for Corollary 5.5. Therefore, Corollary 5.5 applies and $x_k^* \rightarrow x_*$ with $c(x_*) = 0$ and $y_k \rightarrow y_*$. Because c is continuous, the sequence $\{y_k\}$ is Cauchy, and $\omega_k \rightarrow 0$, there exists some $\bar{k}$ large enough so that

$$(6.7a) \quad \bar{\rho} \|c(x_k^*)\|_\infty < \frac{1}{2} \underline{\sigma},$$

$$(6.7b) \quad \|y_{k+1} - y_k\|_\infty < \frac{1}{2} \underline{\sigma} - \omega_k$$

for all $k > \bar{k}$. Step 5 is executed for all k large enough so that $y_{k+1} = y_k + \Delta y_k^* - \bar{\rho} c(x_k^*)$. From the triangle inequality and (6.7), we see that

$$\begin{aligned} \|\Delta y_k^*\|_\infty &\leq \|\Delta y_k^* - \bar{\rho} c(x_k^*)\|_\infty + \bar{\rho} \|c(x_k^*)\|_\infty \\ &= \|y_k + \Delta y_k^* - \bar{\rho} c(x_k^*) - y_k\|_\infty + \bar{\rho} \|c(x_k^*)\|_\infty \\ &= \|y_{k+1} - y_k\|_\infty + \bar{\rho} \|c(x_k^*)\|_\infty \\ &< \underline{\sigma} - \omega_k \\ &\leq \sigma_k - \omega_k \end{aligned}$$

for all $k > \bar{k}$. Therefore, $\sigma_k e - \Delta y_k^* > \omega_k e$ for all $k > \bar{k}$. Because v_k^* and Δy_k^* both satisfy (3.3e), it follows that $\|v_k^*\|_\infty \leq \omega_k$ for all k large enough. By a similar argument, $\sigma_k e + \Delta y_k^* > \omega_k e$, implying that $\|w_k^*\|_\infty \leq \omega_k$ for all k large enough. $\square$

7. Implementation. The practical implementation of an algorithm invariably requires many features that are not made explicit by its theory. In this section we discuss some important details of our sLCL implementation. The algorithm has been implemented in MATLAB 6 [33] and is called LCLOPT. There are two versions. LCLOPT/MINOS uses MINOS 5.5 [35, 36] to solve the sequence of linearly constrained subproblems, while LCLOPT/SNOPT uses SNOPT 6.1-1(5) [24]. We now turn our attention to an optimization problem with more general constraints and leave (GNP) behind.

7.1. Problem formulation. LCLOPT solves problems of the form

$$\begin{array}{ll} \text{(NPI)} & \underset{x,s}{\text{minimize}} \quad f(x) \\ & \text{subject to} \quad \begin{pmatrix} c(x) \\ Ax \end{pmatrix} - s = 0, \quad l \leq \begin{pmatrix} x \\ s \end{pmatrix} \leq u. \end{array}$$

This matches the problem formulation used by SNOPT and is closely related to that used by MINOS. As in those methods, our implementation distinguishes between variables in the vector x that appear and do not appear nonlinearly in the objective or the constraints; variables that appear only linearly are treated specially. The following discussion ignores this detail in order to keep the notation concise.

The linearly constrained subproblems corresponding to (NPI) take the form

$$\begin{array}{ll} \text{(ELCi}_k\text{)} & \underset{x,s,v,w}{\text{minimize}} \quad \mathcal{L}_k(x) + \sigma_k e^T(v + w) \\ & \text{subject to} \quad \begin{pmatrix} c_k + J_k(x - x_k) + v - w \\ Ax \end{pmatrix} - s = 0, \quad l \leq \begin{pmatrix} x \\ s \end{pmatrix} \leq u, \\ & \quad \quad \quad 0 \leq v, w. \end{array}$$

7.2. The main algorithm. The computational kernel of LCLOPT resides in the solution of each LC subproblem, and the efficiency of the implementation ultimately relies on the efficiency of the subproblem solver. The main tasks of the outer level are to form the subproblems, update solution estimates, update parameters, and test for convergence or errors.

7.3. Solving the LC subproblems. For the linearly constrained subproblems, MINOS uses a reduced-gradient method, coupled with a quasi-Newton approximation of the reduced Hessian of the problem objective. SNOPT implements a sparse SQP method and maintains a limited-memory, quasi-Newton approximation of the Hessian of the problem objective. (In both cases, the problem objective is the objective of (ELCi_k).) For linearly constrained problems, SNOPT avoids performing an expensive Cholesky factorization of the reduced Hessian for the quadratic programming subproblem in each of its own major iterations, and thus realizes considerable computational savings over problems with nonlinear constraints [25].

Both MINOS and SNOPT are available as libraries of Fortran 77 routines. We implemented MEX interfaces [34] written in C to make each of the routines from the MINOS and SNOPT libraries accessible from within MATLAB. The subproblem solvers evaluate the nonlinear objective function (there are no nonlinear constraints in (ELCi_k)) through a generic MEX interface, `funObj.c`. This routine makes calls to a MATLAB routine to evaluate the nonlinear objective $\mathcal{L}_k$. In turn, the routine for $\mathcal{L}_k$ makes calls to routines (available as MATLAB or MEX routines) to evaluate the original nonlinear functions f and c .

7.4. Computing an initial point. MINOS and SNOPT both ensure that all iterates remain feasible (to within a small tolerance) with respect to the constraints in (ELCi_k), which include the bounds and linear constraints from the original problem (NPI). A user of LCLOPT may thus introduce bounds and linear constraints into (NPI) to help guard against evaluation of the nonlinear functions at points where they are not defined.

Before entering the first sLCL iteration, LCLOPT uses SNOPT to solve one of the following *proximal-point* (PP) problems:

(PP1) or (PP2)	$\begin{aligned} &\underset{x}{\text{minimize}} \quad \ x - \tilde{x}\ _1 \quad \text{or} \quad \frac{1}{2} \ x - \tilde{x}\ _2^2 \\ &\text{subject to} \quad l \leq \begin{pmatrix} x \\ Ax \end{pmatrix} \leq u, \end{aligned}$

where $\tilde{x}$ is a vector provided by the LCLOPT user. The solution is used as initial point x_0 for the sLCL algorithm. The PP objective functions help find an x_0 reasonably close to $\tilde{x}$, while the constraints ensure that x_0 is feasible with respect to the bounds and linear constraints of (NPi). If the PP problem proves infeasible, (NPi) is declared infeasible and LCLOPT exits immediately with an error message.

The computational results presented in section 8 were derived by using (PP2) to compute x_0 . As suggested by Gill, Murray, and Saunders [25], a loose optimality tolerance on (PP2) is used to limit the computational expense of its solution: reducing the number of iterations and (typically) the number of superbasic variables.

7.5. Early termination of the LC subproblems. The global convergence results for the sLCL algorithm (cf. Lemma 5.2 and Theorem 5.4) assume that the optimality tolerances ω_k for the subproblems converge to 0. This loose requirement allows much flexibility in constructing the sequence $\{\omega_k\}$.

The solution estimates may be quite poor during early iterations. We expect slow progress during those iterations, even if they are solved to tight optimality tolerances. A loose tolerance may help limit the computational work performed by the subproblem solver during these early iterations. Near a solution, however, we wish to reduce the optimality tolerance quickly in order to take advantage of the fast local convergence rate predicted by Theorem 6.2.

To construct $\{\omega_k\}$, we replace step 1 of Algorithm 2 by

$$(7.1) \quad \begin{aligned} \omega &\leftarrow \min(\omega_k, \|F(x_k, y_k, z_k)\|_\infty^2), \\ \omega_{k+1} &\leftarrow \max\left(\frac{1}{2}\omega, \omega_*\right), \end{aligned}$$

where ω_0 can be set by a user to any value between $\frac{1}{2}$ and ω_* . The update (7.1) guarantees that $\omega_k \rightarrow \omega_*$, as required. A gentler reduction is discussed in section 8.5.

Following the prescription outlined in section 3.2, we fix at a small value the feasibility tolerance for satisfying the linearized constraints. The feasibility and optimality tolerances for each major iteration are passed to the subproblem solver as run-time parameters.

7.6. Detecting infeasibility and unboundedness. As discussed in section 5.6, Algorithm 2 will *not* exit if the optimization problem is infeasible and the infeasibility tolerance η_* is small. We declare (NPi) infeasible if at any given iteration k , x_k is infeasible with respect to the nonlinear constraints and the penalty parameter is greater than some threshold value $\bar{\rho}$. In particular, at step 11, Algorithm 2 exits and (NPi) is declared infeasible if

$$\max(\|l_c - c_k\|_\infty^+, \|c_k - u_c\|_\infty^+) > \eta_* \quad \text{and} \quad \rho_k > \bar{\rho},$$

where l_c and u_c are the lower and upper bounds for the nonlinear constraints and $[\cdot]^+$ is the positive part of a vector. For the computational results in section 8 the threshold value was set at $\bar{\rho} = 10^8$.

We also need to consider the possibility that (NPi) is unbounded—i.e., that the objective $f(x)$ is unbounded below in the feasible region or that $\|x\| \rightarrow \infty$. As with

tests for infeasibility, any test for unboundedness must be ad hoc. We rely on the LC solver to help detect infeasibility. Problem (NPi) is declared unbounded and LCLOPT exits if the point x_k is feasible and the LC solver reports (ELCi_k) as unbounded.

7.7. Summary of the stabilized LCL method. The following is a summary of the sLCL algorithm implemented in LCLOPT. We assume that $\tilde{x}$ is given and the starting tolerances ω_0, η_0 and the parameters ρ_0, σ_0 are set.

1. Apply the LC solver to (PP1) or (PP2) to obtain a starting point x_0 that is feasible with respect to the bounds and linear constraints and is reasonably close to $\tilde{x}$. If the PP problem is infeasible, declare (NPi) infeasible and exit. Otherwise, set $k = 0$.
2. Evaluate the functions and gradients at x_k . Linearize the constraints and form (ELCi_k).
3. Apply the LC solver to (ELCi_k) with optimality tolerance ω_k to obtain $(x_k^*, \Delta y_k^*, z_k^*)$. Set $y_k^* = y_k + \Delta y_k^*$.
4. If (ELCi_k) is unbounded and x_k^* is *feasible*, declare (NPi) unbounded and exit. If (ELCi_k) is unbounded and x_k^* is *infeasible*, go to step 8. Otherwise, continue.
5. If x_k^* meets the current nonlinear feasibility threshold η_k , continue. Otherwise, go to step 8.
6. Update the solution estimate: $(x_{k+1}, y_{k+1}, z_{k+1}) \leftarrow (x_k^*, y_k^* - \rho_k c(x_k^*), z_k^*)$. Keep the penalty parameter ρ_k fixed and reset the elastic weight σ_k .
7. Test convergence: If $(x_{k+1}, y_{k+1}, z_{k+1})$ satisfies the optimality conditions for (NPi), declare optimality, return $(x_{k+1}, y_{k+1}, z_{k+1})$, and exit. Otherwise, go to step 9.
8. If $\rho_k > \bar{\rho}$, declare (NPi) infeasible, return (x_k^*, y_k^*, z_k^*) , and exit. Otherwise, discard the subproblem solution (i.e., $(x_{k+1}, y_{k+1}, z_{k+1}) \leftarrow (x_k, y_k, z_k)$), increase the penalty parameter ρ_k , and reduce the elastic weight σ_k .
9. Set the next nonlinear feasibility threshold η_{k+1} and LC subproblem optimality tolerance ω_{k+1} so that $\{(\omega_k, \eta_k)\} \rightarrow (\omega_*, \eta_*)$.
10. Set $k \leftarrow k + 1$. Return to step 2.

7.8. Keeping the penalty parameter small. Preliminary experimentation reveals that a small penalty parameter ρ_k can significantly reduce the difficulty of each subproblem solve. BCL methods require that ρ_k be larger than some threshold value $\bar{\rho}$. In contrast, LCL methods can converge when $\rho_k \equiv 0$ if they are started near a solution (see section 6.2).

The challenge here is to find a strategy that can keep ρ_k small or reduce it without destabilizing the method. A tentative strategy might be to reduce ρ_k only finitely many times. This approach does not violate the hypotheses of Lemma 5.2, and may be effective in practice. A form of this strategy was used for the runs discussed next.

8. Numerical results. This section summarizes the performance of the two LCLOPT implementations on a set of nonlinearly constrained test problems from the COPS 2.0 [15], Hock–Schittkowski [30], and CUTE [5] test suites. As a benchmark, we also give results for AMPL/MINOS 19981015 on all the test problems.

We used the AMPL versions of all problems, as formulated by Vanderbei [45]. A MEX interface to the AMPL library makes functions and gradients available in MATLAB; see Gay [22] for details. (The CUTE versions of the problems could also have been used from MATLAB.) All runs were conducted on an AMD Athlon 1700XP using 384 MB of RAM, running Linux 2.4.18.

Figure 8.1 shows *performance profiles*, as described by Dolan and Moré [16], for LCLOPT/MINOS (dotted line), LCLOPT/SNOPT (dashed line), and MINOS (solid line). The statistic profiled in the top chart is the total number of function and gradient evaluations. In the bottom chart it is the total minor iterations. (Because the nonlinear objective, constraint, and gradient evaluations always occur together, each evaluation of (f, c, g, J) is counted once.) All 135 problems selected from the COPS, Hock–Schittkowski, and CUTE test suites are included in each profile. For each solver and each τ on the horizontal axis, a profile shows the percentage of problems for which the statistic in question (following a successful solve) is within a factor τ of the best.

We see that LCLOPT/MINOS solved the largest proportion of problems and may therefore be regarded as the most reliable method. Compared with MINOS, both versions of LCLOPT tend to require more minor iterations—a measure of total computational work. We comment further in section 8.5.

8.1. Default parameters. Figure 8.2 shows the options files used by LCLOPT for the LC solvers on all subproblems. Separately, LCLOPT sets the **Optimality Tolerance** for MINOS and the **Major Optimality Tolerance** for SNOPT at each major iteration. These are the subproblem optimality tolerances ω_k (see section 7.5).

Each test problem supplies a default starting point. It is used as $\tilde{x}$ in the proximal-point problem (see section 7.4). The initial multiplier vector y_0 is set to zero.

Both MINOS and SNOPT provide the option to reuse a quasi-Newton approximation of a Hessian from a previous solve: MINOS approximates the reduced Hessian; SNOPT approximates the full Hessian. We take advantage of this feature for all iterations $k = 2, 3, 4, \dots$ by setting the MINOS and SNOPT options **Start** = 'Hot'.

The parameters used by Algorithm 2 are set as follows. The upper bound on the elastic penalty parameters is $\bar{\sigma} = 10^4$. The initial elastic weight is $\sigma_0 = 10^2$. (Normally, LCLOPT scales this quantity by $1 + \|y_0\|_\infty$, but the scaling has no effect for these test runs because $y_0 \equiv 0$.) The penalty scaling factors are $\tau_\rho = \tau_\sigma = 10$. As suggested in [10], we set $\alpha = 0.1$ and $\beta = 0.9$. The initial penalty parameter is $\rho_0 = 10^{5/2}/m_c$, where m_c is the number of nonlinear constraints. The final optimality and feasibility tolerances are $\omega_* = \eta_* = 10^{-6}$. The initial optimality and feasibility tolerances are $\omega_0 = 10^{-3}$ ($= \sqrt{\omega_*}$) and $\eta_0 = 1$.

Default options are used for the AMPL/MINOS benchmarks, except for the special values **Major Iterations** 500 and **Superbasics Limit** 2000.

8.2. The COPS test problems. The COPS 2.0 collection [15] comprises 17 problems. Five problems are excluded for the following reasons:

- *bearing*, *minsurf*, and *torsion* are unconstrained or bound constrained;
- *glider* and *marine* cause system errors when called from the AMPL MEX interface.

The dimensions of the COPS test problems can be adjusted. In all cases, the solvers were applied to the largest version of the problem that would not cause the system to page memory to disk. Tables 8.1 and 8.2 specify the chosen dimensions.

As shown in Table 8.3, LCLOPT/MINOS solved all 12 problems to first-order optimality. LCLOPT/SNOPT solved 11 problems to first-order optimality; the exception was *robot*, which it declared as having infeasible nonlinear constraints. MINOS solved 10 of the 12 problems to optimality; it declared *steering* an infeasible problem, and it terminated the solution of *elec* because of excessive iterations. Feasible points exist for all of the test problems chosen, so we consider all declarations of infeasibility to be errors.

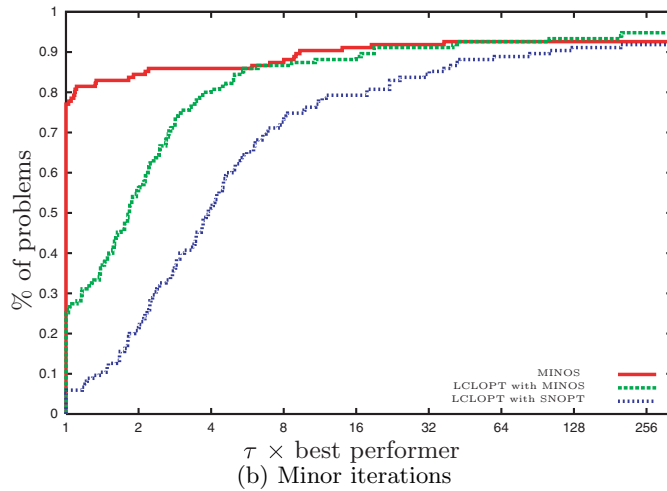
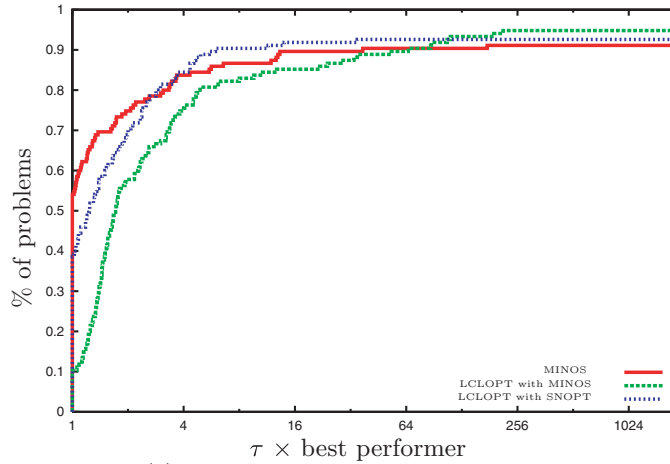


FIG. 8.1. Performance profiles for the number of function and gradient evaluations and the total minor iterations. For each solver, the vertical axes represent the percentage of problems for which the relevant statistic is within a factor τ of the best (among all three solvers). The horizontal axes are based on a log scale. The profiles include results for all 135 test problems.

<pre> BEGIN LCL SUBPROBLEM Scale option 0 Superbasics limit 2000 Iterations 5000 Feasibility tol 1.0e-6 END LCL SUBPROBLEM </pre>	<pre> BEGIN LCL SUBPROBLEM Scale option 0 Superbasics limit 2000 Iterations 5000 Major iterations 1000 Minor iterations 500 Minor feasibility tol 1.0e-6 Minor optimality tol 2.5e-7 END LCL SUBPROBLEM </pre>
The MINOS options file	The SNOPT options file

FIG. 8.2. The fixed optional parameters for every subproblem solve.

TABLE 8.1
The heads used in Tables 8.2 and 8.5.

m	Constraints (linear and nonlinear)
m_c	Nonlinear constraints
n	Variables
n_c	Variables appearing nonlinearly in c
n_f	Variables appearing nonlinearly in f

TABLE 8.2
Dimensions of the 12 selected COPS test problems.

Problem	m	m_c	n	n_c	n_f
<i>camshape</i>	1604	801	800	800	0
<i>catmix</i>	1603	1600	2403	2403	0
<i>chain</i>	204	1	402	201	402
<i>channel</i>	800	400	800	800	0
<i>elec</i>	201	200	600	600	600
<i>gasoil400</i>	4004	3200	4003	4003	202
<i>marine</i>	1208	800	1215	1215	344
<i>methanol</i>	2406	1800	2405	1605	1670
<i>pinene</i>	4006	3000	4005	2405	2469
<i>polygon</i>	1377	1225	100	100	100
<i>robot</i>	2414	2400	3611	3209	0
<i>rocket</i>	2409	1200	1605	1605	0
<i>steering</i>	2011	1600	2007	1204	0

TABLE 8.3
Summary of results for the 12 selected COPS test problems.

	LCLOPT/MINOS	LCLOPT/SNOPT	MINOS
Optimal	12	11	10
False infeasibility		1	1
Terminated by iteration limit			1
Major iterations	118	179	380
Minor iterations	53950	147518	61388
Function evaluations	53081	11014	63701

We note that different local optima appear to have been found for problems *camshape*, *methanol*, *polygon*, and *rocket*. Also, many minor iterations were required on *catmix*, *elec*, and *robot* with LCLOPT/SNOPT. Especially during early major iterations, SNOPT was unable to solve the subproblems to the required optimality tolerance within the 5000 iteration limit. Rather than using a higher limit, LCLOPT restarts SNOPT several times on the same subproblem—a strategy that sometimes reduces the total minor iterations [23].

8.3. The Hock–Schittkowski (HS) test problems. The HS test suite contains 86 nonlinearly constrained problems [30]. These are generally small and dense problems. We exclude five problems from this set for the following reasons:

- *hs67*, *hs85*, and *hs87* are not smooth;
- *hs68* and *hs69* require external functions.

Both versions of LCLOPT solved the same 80 problems to first-order optimality, but both declared *hs109* infeasible. MINOS solved 80 problems to first-order optimality but declared *hs93* infeasible. Table 8.4 summarizes these results.

TABLE 8.4
Summary of results for the 81 selected Hock–Schittkowski test problems.

	LCLOPT/MINOS	LCLOPT/SNOPT	MINOS
Optimal	80	80	80
False infeasibility	1	1	1
Major iterations	654	648	1160
Minor iterations	7415	25290	10111
Function evaluations	12269	14712	27127

Objective function type	: *
Constraint type	: Q 0 (quadratic, general nonlinear)
Regularity	: R (smooth)
Degree of available derivatives	: 1 (first derivatives, at least)
Problem interest	: M R (modeling, real applications)
Explicit internal variables	: *
Number of variables	: *
Number of constraints	: *

FIG. 8.3. Selection criteria for the CUTE test set (where * is a wild-card character).

On *hs13*, all the solvers reached different solutions. However, the linear independence constraint qualification does not hold at the solution, and this violates the assumptions made for both LCLOPT and MINOS.

Recall that LCLOPT and MINOS use only first derivatives and hence may not converge to local minimizers of a problem. For example, LCLOPT (in both versions) converged to a known local solution of *hs16*, but MINOS converged to some other first-order point. In contrast, MINOS converged to the known local solutions of *hs97* and *hs98*, while LCLOPT (in both versions) converged to other first-order points. Similar differences exist for problems *hs47* and *hs77*.

8.4. A selection of CUTE test problems. With the `select` utility [5], we extracted from the CUTE test suite dated September 7, 2000, problems with the characteristics shown in Figure 8.3. These criteria yield 108 problems, but we excluded 66 for the following reasons:

- 33 problems do not have AMPL versions: *car2*, *c-reload*, *dembo7*, *drugdis*, *durgdise*, *errinbar*, *junkturn*, *leaknet*, *lubrif*, *mrribasis*, *nystrom5*, *orbit2*, *reading4*, *reading5*, *reading6*, *reading7*, *reading8*, *reading9*, *rotodisc*, *saromm*, *saro*, *tenbars1*, *tenbars2*, *tenbars3*, *tenbars4*, *trigger*, *truspyr1*, *truspyr2*, *zamb2*, *zamb2-8*, *zamb2-9*, *zamb2-10*, *zamb2-11*;
- 21 problems cause system errors when evaluated either by the AMPL MEX interface or by MINOS (when invoked from AMPL): *brainpc2*, *brainpc3*, *brainpc4*, *brainpc5*, *brainpc6*, *brainpc7*, *brainpc8*, *brainpc9*, *bratu2dt*, *cresc132*, *csfi1*, *csfi2*, *drcav1lq*, *drcav2lq*, *drcav3lq*, *kissing*, *lakes*, *porous1*, *porous2*, *trainf*, *trainh*;
- The AMPL versions of 12 problems are formulated with no nonlinear constraints: *drcavty1*, *drcavty2*, *drcavty3*, *flosp2hh*, *flosp2hl*, *flosp2hm*, *flosp2th*, *flosp2tl*, *flosp2tm*, *methanb8*, *methanl8*, *res*.

(To avoid such an exclusion rate, future experiments will work directly with the CUTER interface [27].) Of the remaining 42 problems, 17 can be adjusted in size. The solvers were again applied to the largest versions that would not cause memory paging. Table 8.5 gives the dimensions.

TABLE 8.5
Dimensions of the variable-size CUTE test problems.

Problem	m	m_c	n	n_c	n_f
<i>bdvalue</i>	1000	1000	1000	1000	0
<i>bratu2d</i>	4900	4900	4900	4900	0
<i>bratu3d</i>	512	512	512	512	0
<i>cbratu2d</i>	882	882	882	882	0
<i>cbratu3d</i>	1024	1024	1024	1024	0
<i>chandheq</i>	100	100	100	100	0
<i>chemrcta</i>	2000	1996	2000	1996	0
<i>chemrctb</i>	1000	998	1000	998	0
<i>clnlbeam</i>	1001	500	1499	499	1000
<i>hadamard</i>	257	128	65	64	65
<i>manne</i>	731	364	1094	364	729
<i>reading1</i>	5001	5000	10001	10000	10000
<i>reading3</i>	103	101	202	202	202
<i>sreading3</i>	5001	5000	10000	9998	9998
<i>ssnlbeam</i>	21	10	31	11	22
<i>svanberg</i>	1001	1000	1000	1000	1000
<i>ubh5</i>	14001	2000	19997	6003	0

TABLE 8.6
Summary of results for the 42 selected CUTE test problems.

	LCLOPT/MINOS	LCLOPT/SNOPT	MINOS
Optimal	36	34	34
False infeasibility	4	3	3
Terminated by iteration limit	1	3	1
Terminated by superbasics limit		1	
Unbounded/badly scaled			3
Final point cannot be improved	1	1	1
Major iterations	400	368	1149
Minor iterations	70476	278162	29021
Function evaluations	59216	57732	53069

As Table 8.6 shows, LCLOPT/MINOS solved 36 of 42 problems to first-order optimality, while LCLOPT/SNOPT and MINOS solved 34. In terms of major iterations, the sLCL method was largely independent of the subproblem solver if the latter did not fail. The high *minor* iterations count for LCLOPT/SNOPT is due to an inordinate number of iterations (about 83,500) to solve *cresc50*, *svanberg*, and *ubh5*, and to another 102,500 minor iterations for failed subproblem solves terminated by the iteration limit.

The lower minor iterations total for MINOS is largely due to the “basis repair” strategy invoked at the start of each major iteration, as described in [25, section 4.5]. This feature is lost when the current MINOS is used to solve separate LC subproblems, but it is a desirable option for future LC solvers.

8.5. Importance of early termination. Tables 8.3, 8.4, and 8.6 show that MINOS solved many of the test problems using a reasonable number of minor iterations but rather many *major* iterations (i.e., LC subproblem solves). This is because MINOS terminates progress on each of its subproblems after only 40 minor iterations. In contrast, LCLOPT attempts to constrain the subproblem iterations by means of an initially loose optimality tolerance (we set $\omega_0 = \sqrt{\omega_*}$ for the numerical results). A

potential weakness of this approach vis à vis MINOS is that there is no a priori bound on the number of subproblem iterations. MINOS's aggressive (and heuristic) strategy seems effective in keeping the total minor iteration counts low. This property is particularly important during the early major iterations, when the current solution estimates are poor.

It may be possible to emulate the MINOS strategy while satisfying the requirement that the subproblem optimality tolerances ω_k converge to zero (cf. Lemma 5.2). For example, LCLOPT might specify a small subproblem iteration limit initially and only gradually increase the limit on successive major iterations. Especially during early major iterations, such a strategy may keep the accumulated number of subproblem iterations small. During later major iterations, the strategy would still ensure that the subproblem solver returns solutions within the prescribed tolerance ω_k .

On the other end of the performance spectrum lies the issue of recovering LCL's fast local convergence rate under inexact solves (cf. section 6.1). Bräuninger [7] proves that the quadratic convergence rate of Robinson's method is retained when ω_k is reduced at a rate $O(\|F(x_k, y_k, z_k)\|^2)$ (cf. Theorem 6.2). The first-order KKT conditions (3.1) for the LCL subproblem can be expressed as

$$(8.1) \quad \begin{pmatrix} \nabla^2 \mathcal{L}_k(x_k) & J_k^T \\ J_k & \end{pmatrix} \begin{pmatrix} p \\ -y \end{pmatrix} + O(\|p\|^2) = \begin{pmatrix} -g_k + J_k^T y_k \\ -c_k \end{pmatrix},$$

where $p \stackrel{\text{def}}{=} x - x_k$, and a first-order Taylor expansion was used to derive the residual term $O(\|p\|^2)$. (We have ignored bound constraints for the moment. Robinson [40, 41] shows that the correct active set is identified by the subproblems near a solution.) The nonlinear equations (8.1) are closely related to the linear equations that would be derived from applying Newton's method to (3.1) (again, ignoring bound constraints). In that case, the theory from inexact Newton methods (Dembo, Eisenstat, and Steihaug [14]) predicts that the quadratic convergence rate is recovered when the residual error is reduced at the rate $O(\|F(x_k, y_k, z_k)\|)$. The similarity between (8.1) and the Newton equations hints at the possibility of recovering the quadratic convergence rate of the LCL and sLCL methods by reducing ω_k at the rate $O(\|F(x_k, y_k, z_k)\|)$. See also Conn, Gould, and Toint [12]. We note, however, that stronger assumptions may be needed on the smoothness of the nonlinear functions. This issue deserves more study.

9. Conclusions. The stabilized LCL method developed in this paper is a generalization of the augmented Lagrangian methods discussed in section 4 and it shares the strengths of its predecessors: it is globally convergent (the BCL advantage) and it has quadratic local convergence (the LCL advantage). The ℓ_1 penalty function on the linearized constraints brings the two together (see section 3.1). Because the method operates in a reduced space (like all LCL methods), it is less sensitive than BCL methods to the choice of each penalty parameter ρ_k .

9.1. A second-derivative LC solver. We prove in section 5.7 that the sLCL method will converge to second-order stationary points if the subproblems are solved to second-order points (for example, by using a second-derivative LC solver). In practice, however, a second-derivative LC solver may be most useful as a means of reducing the overall computational work.

The sLCL method is largely independent of the method in which its subproblems are solved. An LC solver using second derivatives is likely to require fewer iterations (and hence less computational work) for the solution of each of the subproblems.

We would expect the number of required major iterations to remain constant if each subproblem solution is computed to within the prescribed tolerance ω_k . However, we would expect to *reduce* the number of required major iterations if a MINOS-like strategy is used to terminate the subproblems (see section 8.5). Over the same number of iterations, a subproblem solver using second derivatives may make more progress toward a solution than a first-derivative solver.

Any future sLCL implementation would ideally be flexible enough to allow for a variety of solvers to be used for the LC subproblems. The choice of subproblem solver could then be guided by the characteristics of the optimization problem at hand. In particular, the advent of automatic differentiation makes second derivatives increasingly available for certain problem classes, e.g., within recent versions of GAMS and AMPL, and for more general functions defined by Fortran or C code, notably ADIFOR and ADIC (Bischof, Roh, and Mauer [4] and Bischof et al. [3]). These may be used by SQP and interior methods for nonlinearly constrained (NC) problems. Certain theoretical hurdles might be avoided, however, by developing specialized second-derivative LC solvers. Such LC solvers could be extended readily to general NC problems by incorporating them into the sLCL algorithm.

9.2. Looking ahead (and behind). A Fortran 90 implementation of the sLCL algorithm, to be named Knossos, is currently under development [21]. As in Robinson [40] and the MINOS implementation [36], a key concept is *departure from linearity* (meaning the difference between the constraint functions and their linearization at the current solution estimate x_k). For problem (GNP), we define d_k and a *modified augmented Lagrangian* $\mathcal{M}_k$ as follows:

$$\begin{aligned} d_k(x, v, w) &= c(x) - \bar{c}_k(x) - v + w, \\ \mathcal{M}_k(x, v, w) &= f(x) - y_k^T d_k(x, v, w) + \frac{1}{2} \rho_k \|d_k(x, v, w)\|^2. \end{aligned}$$

We then use $\mathcal{M}_k$ in place of $\mathcal{L}_k$ in the elastic subproblem (ELC_k) of section 3:

$\begin{aligned} (\text{ELC}_k'') \quad & \underset{x, v, w}{\text{minimize}} \quad \mathcal{M}_k(x, v, w) + \sigma_k e^T (v + w) \\ & \text{subject to} \quad \bar{c}_k(x) + v - w = 0, \\ & \quad \quad \quad x, v, w \geq 0. \end{aligned}$

Note that $d_k(x, v, w)$ and $c(x)$ have the same value at any point (x, v, w) that satisfies the (elastic) linearized constraints $\bar{c}_k(x) + v - w = 0$. Hence $\mathcal{M}_k$ and $\mathcal{L}_k$ have the same value at such a point. The primal solutions for (ELC_k'') and (ELC_k) are therefore the same. For points (x, v, w) satisfying the elastic linearized constraints, let $\tilde{y}_k(x, v, w) \equiv y_k - \rho_k d_k(x, v, w)$. Then a solution $(x_k^*, v_k^*, w_k^*, \pi_k^*, z_k^*)$ of (ELC_k'') satisfies the first-order KKT conditions

$$\begin{aligned} \bar{c}_k(x) + v - w &= 0, \\ g(x) - (J(x) - J_k)^T \tilde{y}_k(x, v, w) - J_k^T \pi &= z, \\ \min(x, z) &= 0, \\ \min(v, \tilde{y}_k(x, v, w) + \sigma_k e - \pi) &= 0, \\ \min(w, -\tilde{y}_k(x, v, w) + \sigma_k e + \pi) &= 0. \end{aligned}$$

Now defining $\Delta y_k^* \equiv \pi_k^* - \tilde{y}_k(x_k^*, v_k^*, w_k^*)$, we find that $(x_k^*, v_k^*, w_k^*, \Delta y_k^*, z_k^*)$ satisfies (3.1), the first-order KKT conditions for (ELC_k). The multiplier update used in

Algorithm 2 (see step 5) is therefore

$$y_{k+1} = y_k + \Delta y_k^* - \rho_k c(x_k^*) = \tilde{y}(x_k^*, v_k^*, w_k^*) + \Delta y_k^* = \pi_k^*.$$

In other words, the new estimate y_{k+1} is the dual solution of (ELC_k'') . This is exactly the multiplier update first suggested by Robinson and used in MINOS.

The use of $\mathcal{M}_k$ in (ELC_k'') follows another important aspect of the MINOS implementation for problems in which only some of the variables enter the constraints nonlinearly. When $v = w = 0$, the Lagrangian term $y_k^T d_k$ and the penalty term $\|d_k\|^2$ are both nonlinear in the same variables, whereas $\|c(x)\|^2$ would be nonlinear in all components of x .

In retrospect, we see that when the linearized constraints $\bar{c}_k(x) = 0$ are infeasible, the MINOS strategy of relaxing those constraints in gentle stages corresponds to introducing v and w (without the help of σ_k). However, MINOS continues to define $\mathcal{M}_k$ in terms of $d_k = c(x) - \bar{c}_k(x)$ rather than $c(x) - \bar{c}_k(x) - v + w$, and therefore is no longer using the true augmented Lagrangian. This explains the increased failure rate observed with MINOS when the subproblems continue to be infeasible. The sLCL approach must regard v and w as nonlinear variables within the LC subproblems, but this is a small price to pay for improved reliability, and the actual nonlinearity of v and w fades as they become zero near a solution.

Acknowledgments. The authors are indebted to Margaret Wright, Nick Gould, and Annick Sartenaer for their exceptionally careful readings of this paper and its revisions. Many valuable suggestions helped clarify the description. In particular, Nick Gould helped revise Theorems 5.4 and 5.7, and his cautions about the ℓ_1 penalty function led to Theorem 6.3. Annick Sartenaer's remarkable lists of comments also led to a correction of Theorem 5.4, as well as refinement of the proof of Lemma 5.3 and many other detailed changes.

Many references have been given already, but we would like to emphasize the importance of certain authors' work on augmented Lagrangian methods. It was our study of Bertsekas [1], Conn, Gould, and Toint [10], and Conn et al. [9] that revealed the need to bound Δy and ultimately led to the vital subproblem (ELC_k) as well as much of the subsequent convergence analysis. We are also grateful to Philip Gill, Robert Mifflin, Walter Murray, Bruce Murtagh, Elijah (Lucien) Polak, and no doubt others for their valuable input along the way.

REFERENCES

- [1] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.
- [2] M. J. BEST, J. BRÄUNINGER, K. RITTER, AND S. M. ROBINSON, *A globally and quadratically convergent algorithm for general nonlinear programming problems*, Computing, 26 (1981), pp. 141–155.
- [3] C. BISCHOF, A. CARLE, P. HOVLAND, P. KHADEMI, AND A. MAUER, *ADIFOR 2.0 Users' Guide*, Tech. report 192, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, 1998.
- [4] C. BISCHOF, L. ROH, AND A. MAUER, *ADIC: An extensible automatic differentiation tool for ANSI-C*, Software—Practice and Experience, 27 (1997), pp. 1427–1456.
- [5] I. BONGARTZ, A. R. CONN, N. I. M. GOULD, AND PH. L. TOINT, *CUTE: Constrained and unconstrained testing environment*, ACM Trans. Math. Software, 21 (1995), pp. 123–160.
- [6] J. BRÄUNINGER, *A modification of Robinson's algorithm for general nonlinear programming problems requiring only approximate solutions of subproblems with linear equality constraints*, in Optimization Techniques, Proceedings of the 8th IFIP Conference, Würzburg, Part 2, Lecture Notes in Control and Inform. Sci., Springer-Verlag, Berlin, New York, 1977, pp. 33–41.

- [7] J. BRÄUNINGER, *A globally convergent version of Robinson's algorithm for general nonlinear programming problems without using derivatives*, J. Optim. Theory Appl., 35 (1981), pp. 195–216.
- [8] A. BROOKE, D. KENDRICK, AND A. MEERAUS, *GAMS: A User's Guide*, The Scientific Press, Redwood City, CA, 1988.
- [9] A. R. CONN, N. GOULD, A. SARTENAER, AND PH. L. TOINT, *Convergence properties of an augmented Lagrangian algorithm for optimization with a combination of general equality and linear constraints*, SIAM J. Optim., 6 (1996), pp. 674–703.
- [10] A. R. CONN, N. I. M. GOULD, AND PH. L. TOINT, *A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds*, SIAM J. Numer. Anal., 28 (1991), pp. 545–572.
- [11] A. R. CONN, N. I. M. GOULD, AND PH. L. TOINT, *LANCELOT: A Fortran Package for Large-Scale Nonlinear Optimization (Release A)*, Springer-Verlag, Berlin, 1991.
- [12] A. R. CONN, N. I. M. GOULD, AND PH. L. TOINT, *On the number of inner iterations per outer iteration of a globally convergent algorithm for optimization with general nonlinear equality constraints and simple bounds*, in Numerical Analysis 1991, Pitman Res. Notes Math. Ser. 260, D. F. Griffiths and G. A. Watson, eds., Longman Scientific and Technical, Harlow, UK, 1992, pp. 49–68.
- [13] A. R. CONN, N. I. M. GOULD, AND PH. L. TOINT, *Trust-Region Methods*, MPS/SIAM Ser. Optim. 1, SIAM, Philadelphia, 2000.
- [14] R. S. DEMBO, S. C. EISENSTAT, AND T. STEihaug, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408.
- [15] E. D. DOLAN AND J. J. MORÉ, *Benchmarking Optimization Software with COPS*, Tech. report ANL/MCS-246, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, 2000 (revised 2001).
- [16] E. D. DOLAN AND J. J. MORÉ, *Benchmarking optimization software with performance profiles*, Math. Program., 91 (2002), pp. 201–213.
- [17] R. FLETCHER, *An ℓ_1 penalty method for nonlinear constraints*, in Numerical Optimization 1984, P. T. Boggs, R. H. Byrd, and R. B. Schnabel, eds., SIAM, Philadelphia, 1985, pp. 26–40.
- [18] R. FLETCHER, *Practical Methods of Optimization*, 2nd ed., John Wiley and Sons, New York, 1987.
- [19] R. FOURER, D. M. GAY, AND B. W. KERNIGHAN, *AMPL: A Modeling Language for Mathematical Programming*, 2nd ed., Duxbury Press/Brooks/Cole, Pacific Grove, CA, 2003.
- [20] M. P. FRIEDLANDER, *A Globally Convergent Linearly Constrained Lagrangian Method for Nonlinearly Constrained Optimization*, Ph.D. thesis, Stanford University, Stanford, CA, 2002.
- [21] M. P. FRIEDLANDER AND M. A. SAUNDERS, *An LCL implementation for nonlinear optimization*, presented at the 18th International Symposium on Mathematical Programming, Copenhagen, Denmark, 2003; available online from <http://www.stanford.edu/group/SOL/talks.html>.
- [22] D. M. GAY, *Hooking Your Solver to AMPL*, Tech. report 97-4-06, Computing Sciences Research Center, Bell Laboratories, Murray Hill, NJ, 1997.
- [23] P. E. GILL, *private communication*, 2002.
- [24] P. E. GILL, W. MURRAY, AND M. A. SAUNDERS, *User's Guide for SNOPT 5.3: A Fortran Package for Large-scale Nonlinear Programming*, Numerical Analysis Report 97-5, Department of Mathematics, University of California, San Diego, La Jolla, CA, 1997.
- [25] P. E. GILL, W. MURRAY, AND M. A. SAUNDERS, *SNOPT: An SQP algorithm for large-scale constrained optimization*, SIAM J. Optim., 12 (2002), pp. 979–1006.
- [26] P. E. GILL, W. MURRAY, AND M. H. WRIGHT, *Practical Optimization*, Academic Press, London, 1981.
- [27] N. I. M. GOULD, D. ORBAN, AND PH. L. TOINT, *CUTEr and SifDec: A constrained and unconstrained testing environment, revisited*, ACM Trans. Math. Software, 29 (2003), pp. 373–394.
- [28] S.-P. HAN AND O. L. MANGASARIAN, *Exact penalty functions in nonlinear programming*, Math. Program., 17 (1979), pp. 251–269.
- [29] M. R. HESTENES, *Multiplier and gradient methods*, J. Optim. Theory Appl., 4 (1969), pp. 303–320.
- [30] W. HOCK AND K. SCHITTkowski, *Test Examples for Nonlinear Programming Codes*, Lecture Notes in Econom. and Math. Systems 187, Springer-Verlag, Berlin, New York, 1981.
- [31] O. L. MANGASARIAN, *Nonlinear Programming*, McGraw-Hill, New York, 1969.
- [32] N. MARATOS, *Exact Penalty Function Algorithms for Finite Dimensional and Optimization Problems*, Ph.D. thesis, Imperial College of Science and Technology, London, UK, 1978.
- [33] MATHWORKS, *MATLAB User's Guide*, The MathWorks, Natick, MA, 1992.

- [34] MATHWORKS, *MATLAB: External Interfaces*, The MathWorks, Natick, MA, 1995.
- [35] B. A. MURTAGH AND M. A. SAUNDERS, *Large-scale linearly constrained optimization*, Math. Program., 14 (1978), pp. 41–72.
- [36] B. A. MURTAGH AND M. A. SAUNDERS, *A projected Lagrangian algorithm and its implementation for sparse nonlinear constraints*, Math. Programming Stud., 16 (1982), pp. 84–117.
- [37] J. NOCEDAL AND S. J. WRIGHT, *Numerical Optimization*, Springer-Verlag, New York, 1999.
- [38] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solutions of Nonlinear Equations in Several Variables*, Academic Press, London, 1970.
- [39] M. J. D. POWELL, *A method for nonlinear constraints in minimization problems*, in Optimization, R. Fletcher, ed., Academic Press, London, 1969, pp. 283–298.
- [40] S. M. ROBINSON, *A quadratically-convergent algorithm for general nonlinear programming problems*, Math. Program., 3 (1972), pp. 145–156.
- [41] S. M. ROBINSON, *Perturbed Kuhn-Tucker points and rates of convergence for a class of nonlinear-programming algorithms*, Math. Program., 7 (1974), pp. 1–16.
- [42] J. B. ROSEN, *Two-phase algorithm for nonlinear constraint problems*, in Nonlinear Programming 3, O. Mangasarian, R. Meyer, and S. Robinson, eds., Academic Press, New York, 1978, pp. 97–124.
- [43] J. B. ROSEN AND J. KREUSER, *A gradient projection algorithm for non-linear constraints*, in Numerical Methods for Nonlinear Optimization, F. A. Lootsma, ed., Academic Press, London, 1972, pp. 297–300.
- [44] G. VAN DER HOEK, *Asymptotic properties of reduction methods applying linearly equality constrained reduced problems*, Math. Programming Stud., 16 (1982), pp. 162–189.
- [45] R. VANDERBEI, *Benchmarks for Nonlinear Optimization*, <http://www.princeton.edu/~rvdb/bench.html> (December 2002).

PROBABILITY GRADIENT ESTIMATION BY SET-VALUED CALCULUS AND APPLICATIONS IN NETWORK DESIGN*

GEORG CH. PFLUG[†] AND HEINZ WEISSHAUPT[‡]

Abstract. Let $\vartheta \mapsto P(\vartheta)$ be a set-valued mapping from $\mathbb{R}^d$ into the family of closed compact polyhedra in $\mathbb{R}^s$. Let ξ be a $\mathbb{R}^s$ valued random variable. Many stochastic optimization problems in computer networking, system reliability, transportation, telecommunication, finance, etc. can be formulated as a problem to minimize (or maximize) the probability $\mathbb{P}\{\xi \in P(\vartheta)\}$ under some constraints on the decision variable ϑ . For a practical solution of such a problem, one has to approximate the objective function and its derivative by Monte Carlo simulation, since a closed analytical expression is only available in rare cases. In this paper, we present a new method of approximating the gradient of $\mathbb{P}\{\xi \in P(\vartheta)\}$ w.r.t. ϑ by sampling, which is based on the concept of setwise (weak) derivative. Quite surprisingly, it turns out that it is typically easier to approximate the derivative than the objective itself.

Key words. sensitivity analysis, derivative estimation, Monte Carlo simulation, set-valued optimization

AMS subject classifications. Primary: 90C31, 65C05; Secondary: 28A15, 52B11

DOI. 10.1137/S1052623403431639

1. Introduction. In many applications in computer networking, systems reliability, transportation, telecommunication, insurance, and financial optimization, the objective is to minimize the probability of a failure (or network breakdown, ruin, financial distress) under some appropriate constraints. These problems are stochastic optimization problems with a specific type of objective: the probability of an event depending on the decision parameter. They are closely related to chance constrained stochastic optimization problems, in which probabilities of unwanted events appear as constraints; see, for instance, [12], [6].

In this article, we concentrate on the essential part of a stochastic optimization problem with a probability objective: to find the gradients of the objective w.r.t. the decision parameter. Once an accurate estimate of the gradient is found, any numerical algorithm for optimization, e.g., gradient projection, may be applied.

The setup of our problem is as follows: Assume that $\vartheta \mapsto P(\vartheta)$ is a set-valued mapping from $\mathbb{R}^d$ into the family of closed compact polyhedra in $\mathbb{R}^s$. Let ξ be a $\mathbb{R}^s$ -valued random variable with distribution μ and continuous density g . Let $p(\vartheta) = \mu(P(\vartheta)) = \mathbb{P}\{\xi \in P(\vartheta)\}$. If $P(\vartheta)$ is more complicated than a box or a disjoint union of boxes, then the only practical method to calculate $p(\vartheta)$ is by Monte Carlo simulation. Tricky ways to include $P(\vartheta)$ in a union of boxes and/or to use importance sampling are well known. Even more complicated is to calculate the gradient $\nabla p(\vartheta)$ of $p(\vartheta)$ w.r.t. ϑ .

Gradient estimation procedures based on sampling are known under the names of perturbation analysis [4], the score function method [16], and the method of weak derivatives [11]. Perturbation analysis does not work here. The push-out score function method (or transformation method) depends on finding an appropriate reparam-

*Received by the editors July 18, 2003; accepted for publication (in revised form) February 8, 2004; published electronically April 22, 2005.

<http://www.siam.org/journals/siopt/15-3/43163.html>

[†]Department of Statistics and Decision Support Systems, A-1010 Wien, Universitätsstrasse 5, University of Vienna, Vienna, Austria (georg.pflug@univie.ac.at).

[‡]Department of Statistics, University of Düsseldorf, Düsseldorf, Germany.

eterization, which transforms the parameters of the polyhedron into parameters of the probability distribution, which works only in specific cases. We will present here a weak derivative method. Let us briefly explain the idea. Suppose that $\vartheta \mapsto C(\vartheta)$ is a mapping from $\mathbb{R}^d$ into a family of subsets of $\mathbb{R}^s$. Such a function is called a set-valued function (for a complete treatment of set-valued analysis see [15]). If the images are closed (compact, convex, polyhedral, etc.), the mapping is called closed-set-valued (compact-set-valued, convex-set-valued, polyhedral-set-valued, etc.). Set-valued functions extend the usual analysis of functions. For instance, the notion of continuity of C may be expressed in terms of the Hausdorff distance: C is continuous at ϑ_0 , if $\|\vartheta - \vartheta_0\| \rightarrow 0$ implies that $\text{Haus}(C(\vartheta), C(\vartheta_0)) \rightarrow 0$. Here $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^d$ and $\text{Haus}(\cdot)$ is the Hausdorff distance defined in the appendix. There are various concepts of differentiability of set-valued functions; see, e.g., the multiaffine concept by Artstein [1] or the morphology concept by Aubin [2]. In this paper, we use the notion of weak differentiability, which is most suited for application in stochastic optimization (see section 2).

It is important to put the results in comparison to the related results by Uryasev [20], Raik [14], Kibzun and Kan [6], and others. These authors consider the differentiability of the functions

$$(1) \quad \vartheta \mapsto F(\vartheta) = \int_{\bigcap_{i=1}^k \{f_i(\vartheta, y) \leq 0\}} p(\vartheta, y) dy$$

by methods of analysis (the analytical method), while we have chosen a measure theoretic approach. We present a general result on the weak directional differentiability of intersections of weakly differentiable set-valued functions, even for cases when some of the set-valued functions coincide for some or all parameter values; however, we concentrate on linear constraints. Our analysis shows that one has to consider equivalence classes of constraints (Lemma 2) and only directional derivatives exist in general. It should be noticed that the formula presented in our Theorem 1 coincides with the formula of Theorem 3.1 in [19] if the equivalence classes are singletons.

Let us also mention here that the idea of weak set-valued derivatives is to find a signed measure on the surface of the varying set under consideration, which indicates the local changes in the shape of this set. Consequently, the derivative measure is a surface measure. There are formulas (see, e.g., [20]) which represent the surface integrals as integrals over the whole interior. However, these formulas do not lead to good estimators: Simulations show that their variance is much higher than those based on surface integrals. The explanation is that there is a curse of dimensionality in Monte Carlo estimation: sampling in lower dimensional manifolds reduces the variance of the estimators considerably.

We also tried a sampled version of the derivative formula derived by Tretjakov [18] for star-shaped sets, which is based on generating derivatives along rays emanating from some central point. Also this algorithm was considerably inferior to the one presented here. Finally, we mention that if the polyhedra are just rectangles and the distribution is normal or Gamma, some derivative formulas are contained in Prekopa [12, section 6.6.4].

This paper is organized as follows: Section 2 introduces the weak derivative method. In section 3 we present the main theorem and the algorithms. The proof of the main theorem is contained in section 4. An illustrating example is presented in section 5, and numerical results are collected in section 6. In section 7, we generalize the motivating problem and also show how the generalization may be treated. Some

auxiliary results are collected in the appendix.

2. Weak derivatives of set-valued mappings. In this section, we discuss the notion of derivatives of set-valued mappings and give some properties. The concept will be introduced for general convex bodies and not only for convex polyhedra.

Let $\Theta \subseteq \mathbb{R}^d$ be open and let $\vartheta \mapsto C(\vartheta)$ be a set-valued mapping defined on Θ such that the values $C(\vartheta)$ are convex bodies (convex compact sets with nonempty interior) in $\mathbb{R}^s$. Let μ be a probability measure on $\mathbb{R}^s$ which possesses a continuous density g with respect to Lebesgue measure λ .

We denote by ∂C the boundary of the (convex) set C and by $o_{\partial C}$ the surface area measure on ∂C (see appendix).

Given a measure μ we denote by $\mu|_A$ its restriction to A ; i.e., $\mu|_A(B) := \mu(A \cap B)$.

We now define the concept of the weak derivative of a set-valued mapping $\vartheta \mapsto C(\vartheta)$ in direction $\tau \in \mathbb{R}^d$ with respect to a measure μ on $\mathbb{R}^s$.

DEFINITION 1 (see also [11] and [21]). *We say that a convex-valued mapping $\vartheta \mapsto C(\vartheta)$ is weakly differentiable from the right at ϑ in direction τ with respect to a measure μ if there exists a (possibly signed) measure $D_\tau^\mu C(\vartheta)$ such that for all continuous functions $\psi : \mathbb{R}^s \mapsto \mathbb{R}$ with compact support we have*

$$\int \psi d(D_\tau^\mu C(\vartheta)) = \lim_{h \downarrow 0} \frac{1}{h} \int \psi [1_{C(\vartheta + \tau \cdot h)} - 1_{C(\vartheta)}] d\mu.$$

The (signed) measure $D_\tau^\mu C(\vartheta)$ is called the weak derivative.

From the definition of the weak derivative we get in the special case that $\psi = 1$ on a neighborhood of $C(\vartheta)$ that

$$(2) \quad D_\tau(\mu(C(\vartheta))) = \int_{\partial C(\vartheta)} d(D_\tau^\mu C(\vartheta)).$$

Suppose that $D_\tau^\lambda C(\vartheta) \ll o_{\partial C(\vartheta)}$. This implies for continuous $g := \frac{d\mu}{d\lambda}$ that $D_\tau^\mu C(\vartheta) \ll o_{\partial C(\vartheta)}$ and that

$$(3) \quad \frac{D_\tau^\mu C(\vartheta)}{do_{\partial C(\vartheta)}}(x) = g(x) \cdot \frac{dD_\tau^\lambda C(\vartheta)}{do_{\partial C(\vartheta)}}(x) \quad a.s. \quad do_{\partial C(\vartheta)}.$$

Sufficient condition (SC). If the derivative $D_\tau^\lambda C(\vartheta)$ exists, then a sufficient condition for $D_\tau^\lambda C(\vartheta) \ll o_{\partial C(\vartheta)}$ is (as can be easily derived from Definition 2 and Proposition 2 of the appendix), that $\vartheta \mapsto C(\vartheta)$ fulfills a Lipschitz condition with respect to the Hausdorff distance $Haus$; i.e.,

$$(SC) \quad Haus(C(\vartheta), C(\vartheta + h)) \leq c \cdot \|h\|$$

for some constant c and sufficiently small $\|h\|$, where $\|h\|$ denotes the Euclidean norm in $\mathbb{R}^d$.

In the following, we will always assume that the sufficient condition SC is fulfilled. We will treat the case of Lebesgue measure λ in $\mathbb{R}^s$, since all other cases can be brought to this case by simple density correction. For short, we call a function $\vartheta \mapsto C(\vartheta)$ which fulfills Definition 1 for the Lebesgue measure a WDDCV (weakly directionally differentiable convex-valued) function.

There is a calculus for WDDCV functions. The following lemma shows how differentiability carries over to intersections.

LEMMA 1.

- (i) Let $\vartheta \mapsto C_1(\vartheta)$ be a WDDCV function and let C_2 be a compact convex set such that $o_{\partial C_1(\vartheta)}(\partial C_2) = 0$. Then $\vartheta \mapsto C_1(\vartheta) \cap C_2$ is a WDDCV function with derivative $D_\tau^\lambda C_1(\vartheta)|_{C_2}$.
- (ii) Let $\vartheta \mapsto C_j(\vartheta)$ for $j \in J$, a finite index set, be WDDCV functions. Suppose that for all $j_1 \neq j_2$; $j_1, j_2 \in J$

$$(4) \quad \lim_{h \downarrow 0} \frac{1}{h} \lambda([C_{j_1}(\vartheta + h \cdot \tau) \Delta C_{j_1}(\vartheta)] \cap [C_{j_2}(\vartheta + h \cdot \tau) \Delta C_{j_2}(\vartheta)]) = 0,$$

where Δ denotes the symmetric difference. Then $\vartheta \mapsto \bigcap_{j \in J} C_j(\vartheta)$ is a WD-DCV function with derivative $\sum_{j \in J} [D_\tau^\lambda C_j(\vartheta)]|_{\bigcap_{i \in J} C_i}$.

Proof. (i). For all continuous functions ψ with compact support we have by assumption

$$\int \psi d(D_\tau^\lambda C_1(\vartheta)) = \lim_{h \downarrow 0} \frac{1}{h} \int \psi [1_{C_1(\vartheta + \tau \cdot h)} - 1_{C_1(\vartheta)}] d\lambda.$$

This implies that the same property holds for all functions $\tilde{\psi}$, for which the set of discontinuities has zero measure w.r.t. the limiting measure $D_\tau^\lambda C(\vartheta)$ (see Billingsley [3]). Since

$$\lim_{h \downarrow 0} \frac{1}{h} \int \psi [1_{C_1(\vartheta + \tau \cdot h) \cap C_2} - 1_{C_1(\vartheta) \cap C_2}] d\lambda = \lim_{h \downarrow 0} \frac{1}{h} \int \psi 1_{C_2} [1_{C_1(\vartheta + \tau \cdot h)} - 1_{C_1(\vartheta)}] d\lambda,$$

setting $\tilde{\psi} = \psi \cdot 1_{C_2}$ and noticing that the assumption implies that the set of discontinuities of $\tilde{\psi}$ have limiting measure zero because they are contained in ∂C_2 , this implies the first assertion.

(ii). We prove first the result for $J = \{1, 2\}$. Using (i), the assertion (ii) is shown if we establish that for all continuous ψ with compact support, bounded by 1,

$$(5) \quad \lim_{h \downarrow 0} \frac{1}{h} \int \psi(x) I(x) d\lambda(x) = 0,$$

where

$$\begin{aligned} I(x) = & [1_{C_1(\vartheta + h \cdot \tau) \cap C_2(\vartheta + h \cdot \tau)}(x) - 1_{C_1(\vartheta) \cap C_2(\vartheta)}(x)] \\ & - [1_{C_1(\vartheta + h \cdot \tau) \cap C_2(\vartheta)}(x) - 1_{C_1(\vartheta) \cap C_2(\vartheta)}(x)] \\ & - [1_{C_1(\vartheta) \cap C_2(\vartheta + h \cdot \tau)}(x) - 1_{C_1(\vartheta) \cap C_2(\vartheta)}(x)], \end{aligned}$$

i.e., the difference is equivalent to the sum of the cases when C_1 is treated as constant and C_2 is treated as constant—a form of a product rule of differentiation. The integral in (5) is bounded by

$$\frac{1}{h} \lambda([C_1(\vartheta + h \cdot \tau) \Delta C_1(\vartheta)] \cap [C_1(\vartheta + h \cdot \tau) \Delta C_1(\vartheta)]),$$

which by assumption (4) converges to zero.

For the general case one proceeds by induction. Suppose that the proof is established for $J = \{1, 2, \dots, k-1\}$. The condition needed to extend it for another index k is that

$$\lim_{h \downarrow 0} \frac{1}{h} \lambda \left(\left(\bigcap_{j \in J} C_j(\vartheta + h \cdot \tau) \Delta \bigcap_{j \in J} C_j(\vartheta) \right) \cap [C_k(\vartheta + h \cdot \tau) \Delta C_k(\vartheta)] \right) = 0.$$

This condition follows from the fact that

$$\begin{aligned} & \left[\bigcap_{j \in J} C_j(\vartheta + h \cdot \tau) \Delta \bigcap_{j \in J} C_j(\vartheta) \right] \cap [C_k(\vartheta + h \cdot \tau) \Delta C_k(\vartheta)] \\ & \subseteq \bigcup_{j \in J} [C_j(\vartheta + h \cdot \tau) \Delta C_j(\vartheta)] \cap [C_k(\vartheta + h \cdot \tau) \Delta C_k(\vartheta)] \end{aligned}$$

and the application of condition (4) for each of the sets

$$[C_j(\vartheta + h \cdot \tau) \Delta C_j(\vartheta)] \cap [C_k(\vartheta + h \cdot \tau) \Delta C_k(\vartheta)]. \quad \square$$

3. The main theorem and the algorithms. Let $\vartheta \in \mathbb{R}^d$ and let $(\mathbf{a}_i(\vartheta))_{i \in I}$ be row-vectors in $\mathbb{R}^s$, depending on ϑ . Let $(b_i(\vartheta))_{i \in I}$ be real numbers (the right-hand sides). Let $P(\vartheta)$ be the polyhedron

$$(6) \quad P(\vartheta) = \{y \in \mathbb{R}^s : y^T \mathbf{a}_i(\vartheta) \leq b_i(\vartheta); \quad i \in I\}.$$

We suppose that $\mathbf{a}_i(\vartheta)$ and $b_i(\vartheta)$ are differentiable functions in ϑ and that the convex polyhedron $P(\vartheta)$ is compact and has nonempty interior for $\vartheta \in \Theta$, an open subset of $\mathbb{R}^d$.

Let μ be a probability measure on $\mathbb{R}^s$ with continuous density g with respect to Lebesgue measure λ such that

$$\mathbb{P}\{\xi \in P(\vartheta)\} = \mu(P(\vartheta)) = \int_{P(\vartheta)} g(u) \, du.$$

Denote by $D_\tau \mu(P(\vartheta))$ the directional derivative from the right of $\mu(P(\vartheta))$ w.r.t. the direction τ , i.e.,

$$D_\tau \mu(P(\vartheta)) = \lim_{h \downarrow 0} \frac{1}{h} [\mu(P(\vartheta + h\tau)) - \mu(P(\vartheta))].$$

Introduce, in a similar manner, the directional derivatives $D_\tau \mathbf{a}_i$ and $D_\tau b_i$.

It may happen that for the fixed ϑ under consideration, two hyperplanes $\{x : \mathbf{a}_{i_1}(\vartheta)x = b_{i_1}(\vartheta)\}$ and $\{x : \mathbf{a}_{i_2}(\vartheta)x = b_{i_2}(\vartheta)\}$ coincide but are not identical for all ϑ . For this reason, we introduce the equivalence classes

$$[i] := \{j \in I \mid \mathbf{a}_j(\vartheta) = \gamma_j \mathbf{a}_i(\vartheta), b_j(\vartheta) = \gamma_j b_i(\vartheta) \text{ for some } \gamma_j > 0\}, \quad J := \{[i] \mid i \in I\},$$

$$H_i := \{x \mid \mathbf{a}_i(\vartheta)x = b_i(\vartheta)\} \quad \text{and} \quad H_{[i]} := H_i.$$

Our main theorem represents the desired directional derivative $D_\tau(\mu(P(\vartheta)))$ by an integral w.r.t. the surface area measure $o_{\partial P(\vartheta)}$ on $P(\vartheta)$.

THEOREM 1. *Under the given assumptions,*

$$(7) \quad D_\tau \mu(P(\vartheta)) = \int_{\partial P(\vartheta)} g(x) \cdot \sum_{[j] \in J} \mathbf{1}_{P(\vartheta) \cap H_{[j]}} \cdot \min_{i \in [j]} \left\{ \frac{1}{\|\mathbf{a}_i\|} [D_\tau b_i(\vartheta) - x^T D_\tau \mathbf{a}_i(\vartheta)] \right\} d o_{\partial P(\vartheta)}(x).$$

The proof of this theorem will be given in section 5.

Based on this theorem, the following algorithm implements a sampling method for getting an unbiased derivative estimate.

ALGORITHM 1.

- Construct a random variable ζ with probability distribution proportional to $g(\cdot) o_{\partial P(\vartheta)}$.
- Sample x from ζ .
- Calculate $f_i = \frac{1}{\|\mathbf{a}_i(\vartheta)\|} [D_\tau b_i(\vartheta) - x^T D_\tau \mathbf{a}_i(\vartheta)]$ for all i with $x \in H_{[i]}$ and find $f = \min f_i$.
- The derivative estimate is $f \cdot \int_{\partial P(\vartheta)} g(u) do_{\partial P(\vartheta)}(u)$.

By virtue of Theorem 1, this algorithm produces an unbiased estimate of the derivative. The algorithm needs the construction of the random variable ζ . The construction of ζ is not easy in general. The problem is that we have to sample from a distribution on the boundary of $P(\vartheta)$. We replace now this sampling on the boundary by sampling (and if necessary discarding) points on the hyperplanes $H_{[i]}$. Algorithm 3 works only if the integrals $\int_{H_{[j]}} g(x) do_{H_{[j]}}(x)$ are finite.

ALGORITHM 2.

- Construct for all $[j] \in J$ random variables $\zeta_{[j]}$ with probability distributions proportional to $g(\cdot) o_{\partial H_{[j]}}$.
- Sample $x_{[j]}$ from $\zeta_{[j]}$.
- If $x_{[j]} \in P$, then
 - Calculate $f_i = \frac{1}{\|\mathbf{a}_i(\vartheta)\|} [D_\tau b_i(\vartheta) - x_{[j]}^T D_\tau \mathbf{a}_i(\vartheta)]$ for all $i \in I$ and find for all $[j] \in J$ the value $f_{[j]} = \min_{i \in [j]} f_i$.
- Else set $f_{[j]} = 0$.
- The derivative estimate is

$$\sum_{[j] \in J} f_{[j]} \cdot \int_{H_{[j]}} g(u) do_{H_{[j]}}(u).$$

This algorithm needs the construction of random variables $\zeta_{[j]}$, i.e., sampling from a conditional distribution on hyperplanes, rather than from the distribution ζ itself. We will illustrate the method for the important case of ζ being a multivariate normal distribution, because then also the conditional distributions are multivariate normal and easy to sample.

Assuming that μ is multivariate normal, $\mu = N(m, \Sigma)$, recall that it may be transformed by a linear mapping to a standard normal distribution. For a positive definite Σ , let $\sqrt{\Sigma}$ be any matrix satisfying $\sqrt{\Sigma}^T \sqrt{\Sigma} = \Sigma$.

The algorithm specialized for multivariate normal reads as follows.

ALGORITHM 3.

- Transform the problem so that the original distribution $N(m, \Sigma)$ becomes the standard normal distribution $N(0, I)$ on $\mathbb{R}^s$:
 - Let $A^{new} := A \cdot \sqrt{\Sigma}$ and let $b^{new} := b - A \cdot m$.
 - Change the notation to $A := A^{new}$ and $b := b^{new}$.
- Let $z_{[j]} \in H_{[j]}$ be the point in $H_{[j]}$ with minimal norm, i.e.,
 $z_{[j]} = \frac{b_i \mathbf{a}_i}{\mathbf{a}_i^T \mathbf{a}_i}, i \in [j]$.
- Sample x from a standard normal distribution on $\mathbb{R}^s$.
- Calculate for all $[j] \in J$ the orthogonal projection
 $x_{[j]} = x - \frac{\mathbf{a}_i^T x - b_i}{\mathbf{a}_i^T \mathbf{a}_i} \mathbf{a}_i, i \in [j]$, of x onto $H_{[j]}$.
- If $x_{[j]} \in P(\vartheta)$, then
 - Calculate $f_i = \frac{1}{\|\mathbf{a}_i(\vartheta)\|} [D_\tau b_i(\vartheta) - x_{[j]}^T D_\tau \mathbf{a}_i(\vartheta)]$ for all $i \in I$ and find for all $[j] \in J$ the value $f_{[j]} = \min_{i \in [j]} f_i$.
- Else set $f_{[j]} = 0$.
- The derivative estimate is

$$\sum_{[j] \in J} f_{[j]} \cdot \int_{H_{[j]}} \phi(u) d_{H_{[j]}}(u) = \sum_{[j] \in J} f_{[j]} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{\|z_{[j]}\|^2}{2}}.$$

Here ϕ is the density of the standard normal distribution in $\mathbb{R}^s$.

4. The proof of Theorem 1. Let $C(\vartheta)$ be a convex body as in section 4. Later, we will specialize $C(\vartheta) = P(\vartheta)$, where P is the polyhedron defined in (6). From (2) and (3) we obtain

$$(8) \quad D_\tau \mu(C(\vartheta)) = \int_{\partial C(\vartheta)} g(x) \frac{dD_\tau^\lambda(C(\vartheta))}{d\partial C(\vartheta)}(x) d\partial C(\vartheta)(x).$$

Formula (8) entails the main formula (7) if we show that for the convex polyhedron P

$$\frac{dD_\tau^\mu(P(\vartheta))}{d\partial C(\vartheta)}(x) = \sum_{[j] \in J} \mathbf{1}_{P(\vartheta) \cap H_{[j]}} \cdot \min_{i \in [j]} \left\{ \frac{1}{\|\mathbf{a}_i\|} [D_\tau b_i(\vartheta) - x^T D_\tau \mathbf{a}_i(\vartheta)] \right\}.$$

We will first prove the main theorem in the special case that $P(\vartheta)$ is a half-space, i.e., that for all indices $i \in I$, the half-spaces coincide for some fixed ϑ , but are possibly different for all other values of ϑ . Having done this case, an application of Lemma 1 will provide the final proof.

In the following, we will alternatively use the notations $x^T y$ or $\langle x, y \rangle$ for the inner product in $\mathbb{R}^s$, depending upon whatever notation seems more appropriate.

LEMMA 2. *Let I be a finite index set and for $i \in I$ let,*

$$\mathbf{a}_i(\tilde{\vartheta}) \in \mathbb{R}^s \text{ and } b_i(\tilde{\vartheta}) \in \mathbb{R}^+$$

with $\tilde{\vartheta} \mapsto \mathbf{a}_i(\tilde{\vartheta})$, $\tilde{\vartheta} \mapsto b_i(\tilde{\vartheta})$ differentiable. Let

$$P(\tilde{\vartheta}) = \bigcap_{i \in I} \{y : \langle \mathbf{a}_i(\tilde{\vartheta}), y \rangle \leq b_i(\tilde{\vartheta})\}$$

and suppose that for some fixed $\vartheta \in \Theta$ the half-spaces $\{y : \langle \mathbf{a}_i(\vartheta), y \rangle \leq b_i(\vartheta)\}$ are all identical. Then $P(\vartheta)$ is itself a half-space $\mathcal{H}$, the function $\vartheta \mapsto P(\vartheta) \cap R\mathbb{B}$ is weakly

directional differentiable from the right at ϑ , and its directional derivative in direction τ is given by $f_\tau \cdot o$ with o the surface area measure on $\partial\mathcal{H} \cap R\mathbb{B}$ and $f_\tau : \partial\mathcal{H} \mapsto \mathbb{R}$ given by

$$f_\tau(y) = \min_{i \in I} \frac{1}{\|\mathbf{a}_i(\vartheta)\|} [D_\tau b_i(\vartheta) - y^T D_\tau \mathbf{a}_i(\vartheta)].$$

Proof. Let $\mathbf{u} = \frac{\mathbf{a}_i(\vartheta)}{\|\mathbf{a}_i(\vartheta)\|}$ denote the outward unit normal vector on $\mathcal{H}$, which—by assumption—is independent of i .

Introduce the half-spaces $\mathcal{H}_i(h)$ as

$$\mathcal{H}_i(h) = \{x : \langle x, \mathbf{a}_i(\vartheta + h \cdot \tau) \rangle \leq b_i(\vartheta + h \cdot \tau)\}.$$

By assumption, $\mathcal{H} = \mathcal{H}_i(0)$ is independent of i and can be written as

$$\mathcal{H}_i(0) = \{x : \langle x, \mathbf{a} \rangle \leq b\} = \{y + \gamma \mathbf{u} : \gamma \leq 0; \langle y, \mathbf{a} \rangle = b\},$$

where $\mathbf{a} = \mathbf{a}_i(\vartheta)$, $b = b_i(\vartheta)$. Alternatively, we may write

$$\mathcal{H}_i(h) = \{y + \gamma \mathbf{u} : \gamma \leq \xi_{i,h}(y); \langle y, \mathbf{a} \rangle = b\},$$

where $\xi_{i,h}(y)$ for $y \in \partial\mathcal{H}$ is chosen such that

$$\langle y + \xi_{i,h}(y)\mathbf{u}, \mathbf{a}_i(\vartheta + h\tau) \rangle = b_i(\vartheta + h\tau).$$

Now

$$\bigcap_{i \in I} \mathcal{H}_i(h) = \{y + \gamma \mathbf{u} : \gamma \leq \min_i \xi_{i,h}(y); \langle y, \mathbf{a} \rangle = b\}.$$

It is easy to show that in the definition of the weak derivative, the set of continuous functions with compact support may be replaced by any other dense set—for instance, the set of Lipschitz continuous functions with compact support.

Let ψ be a Lipschitz continuous function with compact support. We have to show that

$$(9) \quad \lim_{h \downarrow 0} \frac{1}{h} \int \psi(x) [1_{\bigcap_{i \in I} \mathcal{H}_i(h)}(x) - 1_{\mathcal{H}(0)}(x)] d\lambda(x) = \int \psi(y) f_\tau(y) d\sigma_H(y),$$

where H is the hyperplane $H = \partial\mathcal{H}$. A change of integration variables gives

$$\begin{aligned} & \frac{1}{h} \int \psi(x) [1_{\bigcap_{i \in I} \mathcal{H}_i(h)}(x) - 1_{\mathcal{H}(0)}(x)] d\lambda(x) \\ &= \frac{1}{h} \int_H \int_{\mathbb{R}} \psi(y + \gamma \mathbf{u}) [1_{\bigcap_{i \in I} \mathcal{H}_i(h)}(y + \gamma \mathbf{u}) - 1_{\mathcal{H}(0)}(y + \gamma \mathbf{u})] d\gamma d\sigma_H(y) \\ &= \frac{1}{h} \int_H \int_{\mathbb{R}} [\psi(y) + O(\gamma)] [1_{0 \leq \gamma \leq \min_i \xi_{i,h}(y)} - 1_{\min_i \xi_{i,h}(y) \leq \gamma \leq 0}] d\gamma d\sigma_H(y) \\ &= \int_H \int_{\mathbb{R}} \psi(y) \frac{1}{h} \min_i \xi_{i,h}(y) d\sigma_H(y) + O\left(\int_H \frac{1}{h} \left[\min_i \xi_{i,h}(y)\right]^2 d\sigma_H(y)\right) \\ &= \int_H \int_{\mathbb{R}} \psi(y) \frac{1}{h} \min_i \xi_{i,h}(y) d\sigma_H(y) + o(1) \end{aligned}$$

as $h \downarrow 0$. Here we have used the Landau symbols $O(\cdot)$ and $o(\cdot)$. An easy calculation shows that

$$\lim_{h \downarrow 0} \frac{\xi_{i,h}(y)}{h} = \frac{1}{\|\mathbf{a}_i(\vartheta)\|} [D_\tau b_i(\vartheta) - y^T D_\tau \mathbf{a}_i(\vartheta)]$$

and therefore

$$\lim_{h \downarrow 0} \min_i \frac{\xi_{i,h}(y)}{h} = \min_i \frac{1}{\|\mathbf{a}_i(\vartheta)\|} [D_\tau b_i(\vartheta) - y^T D_\tau \mathbf{a}_i(\vartheta)] = f_\tau(y).$$

Finally, by the dominated convergence theorem,

$$\lim_{h \downarrow 0} \frac{1}{h} \int \psi(x) [1_{\cap_{i \in I} \mathcal{H}_i(h)} - 1_{\mathcal{H}(0)}] d\lambda(x) = \int \psi(y) f_\tau(y) d\sigma_H(y). \quad \square$$

We remark here that a similar result holds if $b_i(\vartheta) \in \mathbb{R}^-$.

The proof of Theorem 1 in the general case. Consider a ball $R\mathbb{B}$ sufficiently big such that $P(\vartheta) \subset \text{int}(R\mathbb{B})$. Propositions 4 and 5 of the appendix ensure that $\vartheta \mapsto P(\vartheta)$ is Lipschitz in the Hausdorff sense and that the sufficient condition is fulfilled. Moreover, for half-spaces the condition (4) is fulfilled. Thus we get from Lemma 2 with

$$C_{[j]} := R\mathbb{B} \cap \bigcap_{i \in [j]} \{x : \langle \mathbf{a}_i(\vartheta), x \rangle \leq b_i(\vartheta)\}$$

and Lemma 1 (ii) applied to the index set $J = \{[j]\}$ that

$$\frac{dD_\tau^\mu(P(\vartheta))}{d\sigma_{\partial C(\vartheta)}}(x) = \sum_{[j] \in J} \mathbf{1}_{P(\vartheta) \cap H_{[j]}} \cdot \min_{i \in [j]} \left\{ \frac{1}{\|\mathbf{a}_i\|} [D_\tau b_i(\vartheta) - x D_\tau \mathbf{a}_i(\vartheta)] \right\}.$$

Inserting this in (8) we obtain (7), i.e., the main theorem has been proved.

5. An illustrating example. We consider the well-known problem of designing a communication network with random demand in an optimal way (see, e.g., [13], [9]). Assume that the topology of the network is given, the decision is to be made about the optimal capacities of the links of the network under a budget constraint. The objective is to maximize the reliability, i.e., minimize the breakdown probability. A breakdown occurs if the demand cannot be satisfied with the available capacities. The example was motivated by a project with the Austrian Telekom company.

The nodes of the network are denoted by $\mathcal{N} = \{1, \dots, N\}$. The potential links are pairs of nodes $\{(k, \ell), 1 \leq k \leq N, 1 \leq \ell \leq N\}$. Every link has a nonnegative capacity $\vartheta_{k,\ell}$. A zero capacity $\vartheta_{k,\ell} = 0$ means that the link (k, ℓ) does not exist.

The demand is given by a collection of random variables $(\xi_{m,n})$ indicating the traffic originating in node m and ending in node n , for every origin-destination (OD) relation $m \rightarrow n$. This demand is mapped onto the links through a routing system R .

We assume first that the routing is fixed. This assumption is relaxed in section 7. The proportion of traffic from m to n flowing through link (k, ℓ) is denoted by $r_{m,n,k,\ell}$. A routing must satisfy

$$\begin{aligned} \sum_{\ell} r_{m,n,m,\ell} &= 1 && \text{for all } m, n; \\ \sum_{\ell} r_{m,n,\ell,k} &= \sum_{\ell} r_{m,n,k,\ell} && \text{for } k \neq m, n; \\ \sum_{\ell} r_{m,n,\ell,n} &= 1 && \text{for all } m, n. \end{aligned}$$

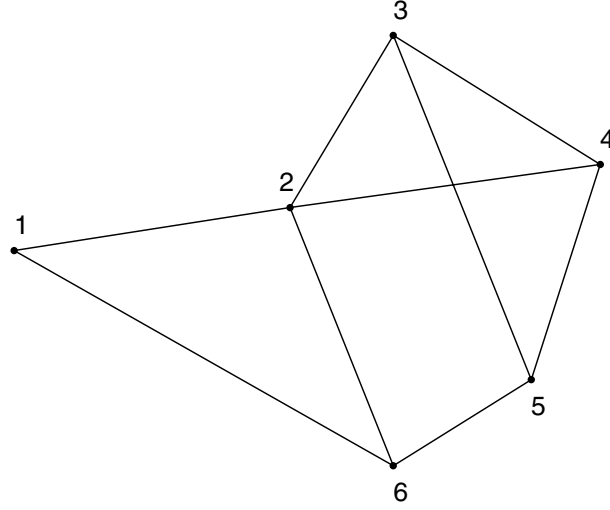


FIG. 1. A network example.

The total flow on link (k, ℓ) is $\sum_{m,n} \xi_{m,n} r_{m,n,k,\ell}$. The capacity constraint reads that

$$(10) \quad \sum_{m,n} \xi_{m,n} r_{m,n,k,\ell} \leq \vartheta_{k,\ell}.$$

The network is reliable if all capacity constraints are fulfilled.

In a compact matrix notation (after a new indexing of links and OD relationships by single indices) the constraint (10) reads

$$R\xi \leq \vartheta.$$

Denote by $P(\vartheta)$ the “reliable” polyhedron

$$P(\vartheta) = \{y \geq 0 : Ry \leq \vartheta\}.$$

The optimal network design problem is

$$(11) \quad \begin{aligned} &\text{Maximize } \mathbb{P}\{\xi \in P(\vartheta)\} \\ &h^T \vartheta \leq B. \end{aligned}$$

Here h is a cost vector and B is the available budget (the constraints may be more complicated and nonlinear as well). The main challenge in solving (11) algorithmically is to calculate the gradient $\nabla \mathbb{P}\{\xi \in P(\vartheta)\} = (\partial p(\vartheta)/\partial \vartheta_1, \dots, \partial p(\vartheta)/\partial \vartheta_d)$.

In the next section, we discuss numerical results about the precision of the proposed estimates of this gradient.

6. Numerical results. Consider the network shown in Figure 1.

The set of nodes is

$$\mathcal{N} = \{1, 2, 3, 4, 5, 6\}.$$

The set of links is

$$\begin{aligned} \mathcal{L} = \{ &l_1 = (1, 2); l_2 = (1, 6); l_3 = (2, 3); l_4 = (2, 4); l_5 = (2, 6); l_6 = (3, 4); \\ &l_7 = (3, 5); l_8 = (4, 5); l_9 = (5, 6)\}. \end{aligned}$$

The set of OD relations is

$$OD = \{(1, 2); (1, 3); (1, 4); (1, 5); (1, 6); (2, 3); (2, 4); (2, 5); (2, 6); (3, 4); (3, 5); (3, 6); (4, 5); (4, 6); (5, 6)\}.$$

For simplicity, the traffic is considered undirected in this example. With the very same method, one could also treat the links as directed and thus double their number as well as the number of OD relations.

The routing matrix is the fixed $[9 \times 15]$ matrix R given below. The ordering of the rows is according to the ordering of links in $\mathcal{L}$, and the ordering of the columns is according to the ordering of the OD relations above.

$$R = \begin{pmatrix} 1.0 & 1.0 & 1.0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.0 & 1.0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1.0 & 0 & 0 & 0 & 1.0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 1.0 & 0 & 0 & 0 & 1.0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 1.0 & 0 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 1.0 & 1.0 & 0 \\ 0 & 0 & 0 & 1.0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0.5 & 0 & 1.0 & 1.0 \end{pmatrix}.$$

We assume that the demand vector ξ follows a 15-dimensional normal distribution with mean value

$$\mathbb{E}(\xi) = (100; 200; 1500; 300; 80; 200; 1500; 300; 80; 3000; 600; 160; 4500; 1200; 240)$$

and covariance matrix

$$\Sigma = (\Sigma_1 \ \Sigma_2),$$

where

$$\Sigma_1 = \begin{pmatrix} 2.5 & 1.5 & 11.25 & 2.25 & 0.6 & 1.5 & 11.25 \\ 1.5 & 10.0 & 22.5 & 4.5 & 1.2 & 3.0 & 7.5 \\ 11.25 & 22.5 & 562.5 & 33.75 & 9.0 & 7.5 & 168.75 \\ 2.25 & 4.5 & 33.75 & 22.5 & 1.8 & 1.5 & 11.25 \\ 0.6 & 1.2 & 9.0 & 1.8 & 1.6 & 0.4 & 3.0 \\ 1.5 & 3.0 & 7.5 & 1.5 & 0.4 & 10.0 & 22.5 \\ 11.25 & 7.5 & 168.75 & 11.25 & 3.0 & 22.5 & 562.5 \\ 2.25 & 1.5 & 11.25 & 6.75 & 0.6 & 4.5 & 33.75 \\ 0.6 & 0.4 & 3.0 & 0.6 & 0.48 & 1.2 & 9.0 \\ 7.5 & 45.0 & 337.5 & 22.5 & 6.0 & 45.0 & 337.5 \\ 1.5 & 9.0 & 22.5 & 13.5 & 1.2 & 9.0 & 22.5 \\ 0.4 & 2.4 & 6.0 & 1.2 & 0.96 & 2.4 & 6.0 \\ 11.25 & 22.5 & 506.25 & 101.25 & 9.0 & 22.5 & 506.25 \\ 3.0 & 6.0 & 135.0 & 9.0 & 7.2 & 6.0 & 135.0 \\ 0.6 & 1.2 & 9.0 & 5.4 & 1.44 & 1.2 & 9.0 \end{pmatrix},$$

TABLE 1
The values of the estimated gradient vector.

	Gradient estimate	
	5000 iterations	50000 iterations
$\partial/\partial\vartheta_1$	$0.0281 * 10^{-3}$	$0.0282 * 10^{-3}$
$\partial/\partial\vartheta_2$	$0.0538 * 10^{-3}$	$0.0538 * 10^{-3}$
$\partial/\partial\vartheta_3$	$0.1394 * 10^{-3}$	$0.1388 * 10^{-3}$
$\partial/\partial\vartheta_4$	$0.0013 * 10^{-3}$	$0.0013 * 10^{-3}$
$\partial/\partial\vartheta_5$	$0.0000 * 10^{-3}$	$0.0000 * 10^{-3}$
$\partial/\partial\vartheta_6$	$0.0556 * 10^{-3}$	$0.0557 * 10^{-3}$
$\partial/\partial\vartheta_7$	$0.0784 * 10^{-3}$	$0.0786 * 10^{-3}$
$\partial/\partial\vartheta_8$	$0.0040 * 10^{-3}$	$0.0040 * 10^{-3}$
$\partial/\partial\vartheta_9$	$0.0009 * 10^{-3}$	$0.0009 * 10^{-3}$

$$\Sigma_2 = \begin{pmatrix} 2.25 & 0.6 & 7.5 & 1.5 & 0.4 & 11.25 & 3.0 & 0.6 \\ 1.5 & 0.4 & 45.0 & 9.0 & 2.4 & 22.5 & 6.0 & 1.2 \\ 11.25 & 3.0 & 337.5 & 22.5 & 6.0 & 506.25 & 135.0 & 9.0 \\ 6.75 & 0.6 & 22.5 & 13.5 & 1.2 & 101.25 & 9.0 & 5.4 \\ 0.6 & 0.48 & 6.0 & 1.2 & 0.96 & 9.0 & 7.2 & 1.44 \\ 4.5 & 1.2 & 45.0 & 9.0 & 2.4 & 22.5 & 6.0 & 1.2 \\ 33.75 & 9.0 & 337.5 & 22.5 & 6.0 & 506.25 & 135.0 & 9.0 \\ 22.5 & 1.8 & 22.5 & 13.5 & 1.2 & 101.25 & 9.0 & 5.4 \\ 1.8 & 1.6 & 6.0 & 1.2 & 0.96 & 9.0 & 7.2 & 1.44 \\ 22.5 & 6.0 & 2250.0 & 135.0 & 36.0 & 1012.5 & 270.0 & 18.0 \\ 13.5 & 1.2 & 135.0 & 90.0 & 7.2 & 202.5 & 18.0 & 10.8 \\ 1.2 & 0.96 & 36.0 & 7.2 & 6.4 & 18.0 & 14.4 & 2.88 \\ 101.25 & 9.0 & 1012.5 & 202.5 & 18.0 & 5062.5 & 405.0 & 81.0 \\ 9.0 & 7.2 & 270.0 & 18.0 & 14.4 & 405.0 & 360.0 & 21.6 \\ 5.4 & 1.44 & 18.0 & 10.8 & 2.88 & 81.0 & 21.6 & 14.4 \end{pmatrix}.$$

In this special case, the matrix A equals the routing matrix R and does not depend on ϑ . The right-hand side is equal to ϑ , i.e., $b_i(\vartheta) = \vartheta$. Further, the directions of the vectors $\mathbf{a}_i$ are all different and so all hyperplanes of the form $\{x \mid \mathbf{a}_i^T \mathbf{x} = \vartheta\}$ are different. Thus the equivalence classes $[i]$ contain just the element i . The derivative estimate in the direction e_i of the i th coordinate-axis thus becomes in the special case of our example

$$\begin{cases} \frac{1}{\|(R\sqrt{\Sigma})_{(i,:)}\|} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{\|z_i\|_2^2}{2}} & \text{if } x \in H_i \cap P_\vartheta \\ 0 & \text{otherwise.} \end{cases}$$

Here $(R\sqrt{\Sigma})_{(i,:)}$ is the i th row of the transformed routing matrix $R\sqrt{\Sigma}$.

Algorithm 3 was applied to this example to estimate $\nabla \mathbb{P}\{\xi \in P(\vartheta)\}$ at the point $\vartheta = (1890; 400; 500; 3310; 330; 3150; 715; 6150; 2070)$. The column vectors of estimates of the partial derivatives for sample sizes of 5000 and 50000 are shown in Table 1.

We compare this result with the result obtained by estimating the gradient by a more or less straightforward algorithm (Algorithm 4 below), which calculates an estimate of the directional derivative by calculating $\frac{\mathbb{P}\{\xi \in P(\vartheta + \Delta e_i) \setminus P(\vartheta)\}}{\Delta}$.

This algorithm is based on Monte Carlo simulation. Instead of sampling from the distribution of ξ it uses importance sampling from different distributions for each unit direction e_i . These distributions are chosen to improve the performance of this

TABLE 2
A comparison of the componentwise standard deviations of the estimates.

	Standard deviation	
	Algorithm 3	Algorithm 4
$\partial/\partial\vartheta_1$	$0.0627 * 10^{-4}$	$0.3158 * 10^{-3}$
$\partial/\partial\vartheta_2$	$0.0940 * 10^{-4}$	$0.2319 * 10^{-3}$
$\partial/\partial\vartheta_3$	$0.3196 * 10^{-4}$	$0.6393 * 10^{-3}$
$\partial/\partial\vartheta_4$	$0.0106 * 10^{-4}$	$0.0236 * 10^{-3}$
$\partial/\partial\vartheta_5$	$0.0000 * 10^{-4}$	$0.0000 * 10^{-3}$
$\partial/\partial\vartheta_6$	$0.0732 * 10^{-4}$	$0.8478 * 10^{-3}$
$\partial/\partial\vartheta_7$	$0.1520 * 10^{-4}$	$0.5134 * 10^{-3}$
$\partial/\partial\vartheta_8$	$0.0086 * 10^{-4}$	$0.0850 * 10^{-3}$
$\partial/\partial\vartheta_9$	$0.0038 * 10^{-4}$	$0.0104 * 10^{-3}$

simple algorithm and to allow a fair comparison of our Algorithm 3 based on weak derivatives with this simple method.

ALGORITHM 4.

- Transform the problem so that the original distribution $N(m, \Sigma)$ becomes the standard normal distribution $N(0, I)$ on $\mathbb{R}^s$:
 - Let $A^{new} := A \cdot \sqrt{\Sigma}$ and let $b^{new} := b - A \cdot m$.
 - Change the notation to $A := A^{new}$ and $b := b^{new}$.
- Let $z_{[j]} \in H_{[j]}$ be the point in $H_{[j]}$ with minimal norm, i.e., $z_{[j]} = \frac{b_j \mathbf{a}_j}{\mathbf{a}_j^T \mathbf{a}_j}, i \in [j]$.
- Sample x_i from a normal distribution on $\mathbb{R}^s$ with mean-vector z_i and covariance matrix the identity.
- If $x_i \in P(\vartheta + \Delta) \setminus P(\vartheta)$, then
 - let $f_i = 1$.
- Else set $f_i = 0$.
- The estimate of the i th directional derivative is

$$\frac{f_i}{\Delta} \cdot e^{\frac{-x_i^T \cdot x_i + (x_i - z_i)^T \cdot (x_i - z_i)}{2}}.$$

The vectors of estimated standard deviations of Algorithms 3 and 4 (for $\Delta = 0.5$) are displayed in Table 2.

Considering these standard deviations we see that about only 0.25 percent of the sample size is needed in our algorithm to obtain an equally stable estimate compared to the direct Algorithm 4. Moreover, the estimate obtained by the direct algorithm is biased, while the estimate based on Algorithm 3 is unbiased (Table 3 shows the biases).

7. Demand dependent routing. In this section we show that demand dependent routing is included in the type of problems studied already. In the demand dependent routing situation, the routing is not fixed, but is calculated through an optimal assignment problem.

Let $z_{m,n,k,\ell}$ be the traffic from origin m to destination n , which flows over the

TABLE 3
A comparison of the estimated gradient vectors.

	Gradient estimate	
	Algorithm 3 with $5 * 10^4$ iterations	Algorithm 4 with $5 * 10^5$ iterations and $\Delta = 0.5$
$\partial/\partial\vartheta_1$	$0.0281 * 10^{-3}$	$0.0270 * 10^{-3}$
$\partial/\partial\vartheta_2$	$0.0538 * 10^{-3}$	$0.0447 * 10^{-3}$
$\partial/\partial\vartheta_3$	$0.1394 * 10^{-3}$	$0.1191 * 10^{-3}$
$\partial/\partial\vartheta_4$	$0.0013 * 10^{-3}$	$0.0014 * 10^{-3}$
$\partial/\partial\vartheta_5$	$0.0000 * 10^{-3}$	$0.0000 * 10^{-3}$
$\partial/\partial\vartheta_6$	$0.0556 * 10^{-3}$	$0.0537 * 10^{-3}$
$\partial/\partial\vartheta_7$	$0.0784 * 10^{-3}$	$0.0712 * 10^{-3}$
$\partial/\partial\vartheta_8$	$0.0040 * 10^{-3}$	$0.0040 * 10^{-3}$
$\partial/\partial\vartheta_9$	$0.0009 * 10^{-3}$	$0.0009 * 10^{-3}$

link (k, ℓ) . These flows must satisfy

$$\begin{aligned}
 \sum_{\ell} z_{m,n,m,\ell} &= \xi_{m,n} && \text{for all } m, n, \\
 \sum_{\ell} z_{m,n,\ell,k} &= \sum_{\ell} z_{m,n,k,\ell} && \text{for } k \neq m, n, \\
 \sum_{\ell} z_{m,n,\ell,n} &= \xi_{n,m} && \text{for all } n, m.
 \end{aligned}
 \tag{12}$$

The capacity constraints are

$$\sum_{m,n} z_{m,n,k,\ell} \leq \vartheta_{k,\ell}.
 \tag{13}$$

We say that ξ is *admissible* if there is a $z \geq 0$ such that all equations (12) and the inequality (13) are fulfilled.

By introducing appropriate indices for links and OD relations, introducing the appropriate matrices A and M , one gets a system of the following form:

ξ is admissible, if $\exists z \geq 0$ such that

$$Az = \xi, \quad Bz = 0, \quad Mz \leq \vartheta.$$

The admissible polyhedron is

$$P(\theta) = \{\xi : \exists z \geq 0 : Az = \xi, Bz = 0, Mz \leq \vartheta\}.$$

This representation is not in the desired explicit form (6). Let's rewrite it as

$$\begin{aligned}
 P(\vartheta) &= \{\xi : \exists z \geq 0 : Az = \xi, Bz = 0, Mz \leq \vartheta\} \\
 &= \{\xi : \exists z, y \geq 0 : Az = \xi, Bz = 0, Mz + y = \vartheta\} \\
 &= \left\{ \xi : \exists x \geq 0 : Wx = \begin{pmatrix} \xi \\ 0 \\ \vartheta \end{pmatrix} \right\}.
 \end{aligned}$$

Here we have set

$$x = \begin{pmatrix} z \\ y \end{pmatrix} \quad \text{and} \quad W = \begin{pmatrix} A & 0 \\ B & 0 \\ M & I \end{pmatrix}.$$

The set $\{d : \exists x \geq 0 : Wx = d\}$ is identical with the set $\{d : d^T e_i \geq 0\}$, where $\{e_i\}$ are the extremal rays of the polyhedral cone $\{u : W^T u \geq 0\}$. Let $\{(u_i, v_i, w_i), i \in I\}$ be the extremal rays of the polyhedral cone $\{(u, v, w) : A^T u + B^T v + M^T w \geq 0, w \geq 0\}$.

We see that the polyhedron $P(\vartheta)$ has the representation

$$P(\vartheta) = \{\xi : -\xi^T u_i \leq \vartheta^T w_i, i = 1, \dots, I\},$$

which is of the required form (6).

How can one calculate the extremal rays of a polyhedral cone? Let W be a $[n \times m]$ matrix of rank d . This matrix describes a linear mapping from $\mathbb{R}^m$ onto a d -dimensional subspace of $\mathbb{R}^n$. Let k be the dimension of the kernel of W . We have that: $d = m - k$. The cone $\{u : Wu \geq 0\}$ is the set of all u 's whose image lies in the positive quadrant. We may find a basis $(b_1, \dots, b_m)$ of $\mathbb{R}^m$ such that $(b_1, \dots, b_k)$ is a basis of the kernel and $(b_{k+1}, \dots, b_m)$ is orthogonal to it. It is sufficient to intersect the linear combinations of the images of $(b_{k+1}, \dots, b_m)$ with the facets of the positive orthant. The dimension of the image is $m - k$, thus more $n - m + k$ linear conditions are needed. We solve the homogeneous equation $\sum_{i=k+1}^m \lambda_i b_i = r$, where r is zero for $n - m + k$ indices and free elsewhere. If the solution is nonnegative, an extremal ray has been found. We have to test at most $\binom{n}{n-m+k}$ equations. Notice that the extremal rays are independent of ϑ and can be stored and kept once they were found.

Appendix. In this appendix we collect some auxiliary facts, most of them without proof. Denote by $\mathbb{S}$ the unit sphere and by $\mathbb{B}$ the unit ball in $\mathbb{R}^s$.

PROPOSITION 1. *Let C be a convex body and for $h \geq 0$ let*

$$C_h^- := C \cap (\partial C + h\mathbb{B}) \quad \text{and} \quad C_h^+ := (\partial C + h\mathbb{B}) \setminus \text{int} C.$$

Then there exists a measure $o_{\partial C}$ (the surface area measure) such that in weak sense

$$\lim_{h \downarrow 0} \frac{1}{h} \lambda|_{C_h^-} = \lim_{h \downarrow 0} \frac{1}{h} \lambda|_{C_h^+} = o_{\partial C}.$$

DEFINITION 2. *The Hausdorff distance $Haus$ of two convex bodies C, D is defined by*

$$Haus(C, D) := \inf\{\gamma > 0 : D \subseteq C + \gamma\mathbb{B} \quad \text{and} \quad C \subseteq D + \gamma\mathbb{B}\}.$$

PROPOSITION 2.

- (i) *If $Haus(C, D) = \epsilon$, then $C \setminus D \subseteq \partial D + \epsilon\mathbb{B}$ and $D \setminus C \subseteq \partial C + \epsilon\mathbb{B}$.*
- (ii) *$Haus(\partial C, \partial D) \leq Haus(C, D)$.*
- (iii) *If $Haus(C, D) = \epsilon$, then $C \Delta D \subseteq \partial C + 2\epsilon\mathbb{B}$. Here Δ denotes the symmetric difference.*

Proof. (i) follows directly from the definition. For (ii) let $\gamma = Haus(C, D)$ and $x \in \partial C$. We have to find a $y \in \partial D$ such that $\|x - y\| \leq \gamma$. If $x \notin D$, then there is a $y \in \partial D$ such that $\|x - y\| \leq \gamma$ by assumption. If $x \in D$, then find a point $y \in \partial D$, such that the (convex) projection of y onto C is x . For this point, $\gamma \geq \inf\{\|y - v\| : v \in C\} = \|y - x\|$ and the assertion is shown. (iii) follows from (i) and (ii): $D \setminus C \subseteq \partial C + \epsilon\mathbb{B}$ and $C \setminus D \subseteq \partial D + \epsilon\mathbb{B} \subseteq \partial C + \epsilon\mathbb{B} + \epsilon\mathbb{B} \subseteq \partial C + 2\epsilon\mathbb{B}$. $\square$

DEFINITION 3. *For a convex body C with the property that $0 \in \text{int}(C)$, define for $x \in \mathbb{S}$ the radial function*

$$\rho_C(x) = \sup\{\gamma : \gamma x \in C\}.$$

The radial distance between two convex bodies C, D with $0 \in \text{int}(C)$, $0 \in \text{int}(D)$ is

$$\text{rad}(C, D) := \sup_{x \in \mathbb{S}} |\rho_C(x) - \rho_D(x)|.$$

PROPOSITION 3. If C, D are two convex bodies with $r\mathbb{B} \subseteq C, D \subseteq R\mathbb{B}$, then $\frac{r}{R}\text{rad}(C, D) \leq \text{Haus}(C, D) \leq \text{rad}(C, D)$. Thus rad and Haus are equivalent on the space of convex bodies C with $r\mathbb{B} \subseteq C \subseteq R\mathbb{B}$.

Proof. The inequality $\text{Haus}(C, D) \leq \text{rad}(C, D)$ is obvious. Thus we show that $\frac{r}{R}\text{rad}(C, D) \leq \text{Haus}(C, D)$. Let $p \in \partial C$ and $q \in \partial D$ be points such that $\|p - q\| = \text{rad}(C, D)$, such that p and q lie on a ray originating at 0 and suppose without loss of generality that $p \in D$ and $q \notin C$. Let $v \in \partial C$ be such that $\|v - q\| = \min_{w \in \partial C} \|w - q\|$. Then $\text{Haus}(C, D) \geq \|v - q\|$. Now let $t \in r\mathbb{S}$ the point on the ray from 0 in direction $q - v$. We have $\|t\| = r$. Notice that p, q, v, t lie in one hyperplane. Let u be the point where the line between 0 and q intersects the line between t and v . Notice that $\|u - q\| \geq \|p - q\|$ by the convexity of C . Thus

$$\frac{\text{Haus}(C, D)}{r} \geq \frac{\|v - q\|}{\|t\|} = \frac{\|u - q\|}{\|u\|} \geq \frac{\|p - q\|}{R} = \frac{\text{rad}(C, D)}{R}. \quad \square$$

Notice that the radial functions satisfy $\rho_{\cap C_i} = \min_i \rho_{C_i}$, and therefore the next proposition is an easy consequence of Proposition 3.

PROPOSITION 4. Let $(C_i)_{i=1, \dots, l}$ and $(D_i)_{i=1, \dots, l}$ be convex sets with

$$(14) \quad r\mathbb{B} \subseteq C, D \subseteq R\mathbb{B} \quad \text{for all } i = 1, \dots, l.$$

Then

$$\text{Haus}\left(\bigcap_{i=1}^l C_i, \bigcap_{i=1}^l D_i\right) \leq \frac{R}{r} \max_{i=1, \dots, l} \text{Haus}(C_i, D_i).$$

Notice that condition (14) can be weakened to the requirement that it is fulfilled for a certain translation of the sets C_i, D_i .

PROPOSITION 5. Let $\Theta \subset \mathbb{R}^d$ and let $\mathbf{a} : \Theta \rightarrow \mathbb{R}^s$ and $b : \Theta \rightarrow \mathbb{R}$ be differentiable functions with $\mathbf{a}(\vartheta) \neq 0$ for all $\vartheta \in \Theta$. Let $H(\vartheta) = \{x \in \mathbb{R}^s \mid \mathbf{a}(\vartheta)^T x \leq b(\vartheta)\}$ and suppose that $H(\vartheta) \cap R\mathbb{B} \neq \emptyset$ for all $\vartheta \in \Theta$. Then, for all $R > 0$, there exists a $c > 0$ such that for sufficiently small $h > 0$,

$$\text{Haus}(H(\vartheta) \cap R\mathbb{B}, H(\vartheta + h \cdot \tau) \cap R\mathbb{B}) \leq c \cdot h,$$

i.e., the sufficient condition SC is fulfilled.

Acknowledgment. We thank three anonymous referees for their valuable remarks and suggestions.

REFERENCES

- [1] Z. ARTSTEIN, *A calculus of set-valued maps and set-valued evolution equations*, Set-Valued Analysis, 3 (1995), pp. 213–261.
- [2] J. P. AUBIN, *Multidimensional and Morphological Analysis: Tools for Shape Regulation and Morphogenesis*, Birkhäuser, Basel, Switzerland, 1999.
- [3] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley & Sons, New York, 1968.
- [4] P. GLASSERMAN, *Gradient Estimation via Perturbation Analysis*, Kluwer Academic Publishers, Norwell, MA, 1991.

- [5] D. LIEBER, A. NEMIROVSKI, AND R. Y. RUBINSTEIN, *A fast Monte Carlo method for evaluation of reliability indices*, IEEE Trans. on Reliability Systems, 48 (1999), pp. 256–261.
- [6] A. I. KIBZUN AND YU. S. KAN, *Stochastic Programming Problems with Probability and Quantile Functions*, John Wiley & Sons, New York, 1966.
- [7] A. I. KIBZUN AND G. L. TRETYAKOV, *On differentiability of probability functions*, J. Comput. System Sci. Inter., 35 (1996), pp. 217–227.
- [8] A. I. KIBZUN AND S. URYASEV, *Differentiability of probability functions*, Stochastic Anal. Appl., 16 (1998), pp. 1101–1128.
- [9] N. KRIVULIN, *An analysis of gradient estimates in stochastic network optimization problems*, in Model-Oriented Data Analysis (Petrodvorets, 1992), Physica, Heidelberg, 1993, pp. 227–240.
- [10] K. MARTI, *Differentiation of probability functions: The transformation method*, Comput. Math. Applic., 30 (1995), pp. 361–382.
- [11] G. CH. PFLUG, *Optimization of Stochastic Models: The Interface Between Simulation and Optimization*, Kluwer Academic Publishers, Boston, MA, 1996.
- [12] A. PREKOPA, *Stochastic Programming*, Math. Applications 324, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1995.
- [13] A. PREKOPA, *Network planning using two-stage programming under uncertainty*, in Recent Results in Stochastic Programming (Proc. Meeting, Oberwolfach, 1979), Lecture Notes in Econom. and Math. Systems 179, Springer-Verlag, Berlin, 1980, pp. 215–237.
- [14] E. RAIK, *The differentiability in the parameter of the probability function and optimization via the stochastic pseudogradient method*, Izvestiya Akad. Nauk Est. SSR, Phis. Math., 24 (1975), pp. 3–6 (in Russian).
- [15] R. T. ROCKAFELLAR AND R. J. B. WETS, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
- [16] R. Y. RUBINSTEIN AND A. SHAPIRO, *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*, John Wiley & Sons, Chichester, UK, 1993.
- [17] R. Y. RUBINSTEIN, *Sensitivity analysis of discrete event systems by the “push out” method*, Ann. Oper. Res., 39 (1992), pp. 229–250.
- [18] G. TRETYAKOV, *Stochastic quasi-gradient algorithms for maximization of the probability function: A new formula for the gradient of the probability function*, in Stochastic Optimization Techniques: Numerical Methods and Technical Applications, K. Marti, ed., Springer-Verlag, Berlin, 2002, pp. 117–139.
- [19] S. URYASEV, *Derivatives of probability functions and integrals over sets given by inequalities*, J. Comput. Appl. Math., 56 (1994), pp. 197–223.
- [20] S. URYASEV, *Derivatives of probability functions and some applications*, Ann. Oper. Res., 56 (1995), pp. 287–311.
- [21] H. WEISSHAUPT, *A measure-valued approach to convex set-valued dynamics*, Set-Valued Anal., 9 (2001), pp. 337–373.

THE CONTINUOUS NEWTON–RAPHSON METHOD CAN LOOK AHEAD*

RAPHAEL HAUSER[†] AND JELENA NEDIĆ[†]

Abstract. This paper is about an intriguing property of the continuous Newton–Raphson method for the minimization of a continuous objective function f : if x is a point in the domain of attraction of a strict local minimizer x^* , then the flux line of the Newton–Raphson flow that starts in x approaches x^* from a direction that depends only on the behavior of f in arbitrarily small neighborhoods around x and x^* . In fact, if $\tilde{f}$ is a sufficiently benign perturbation of f on an open region D not containing x , then the two flux lines through x defined by the Newton–Raphson vector fields that correspond to f and $\tilde{f}$ differ from one another only within D .

Key words. continuous Newton method, Newton–Raphson method, Newton flow, qualitative analysis

AMS subject classifications. 65K10, 65H20, 58C15

DOI. 10.1137/S1052623403432633

1. Introduction. The continuous Newton method for the solution of nonlinear systems of equations arises from Newton’s method when damping with infinitesimal step sizes is applied. To numerical analysts this method is of theoretical interest because the topological behavior of the continuous dynamical system corresponding to a Newton vector field is better understood than the discrete dynamical systems arising from Newton’s method with finite damping; see, for example, the papers of Smale [9], Branin [4], Janovsky and Seige [5], Jongen, Jonker, and Twilt [6], and Neuberger [7]. Such results proved to be useful in the analysis of solutions to systems of PDEs using Sobolev gradient methods where an insufficient number of boundary conditions may not directly guarantee the existence of a unique solution; see, for example, Ball [2] and Neuberger [8]. Anstreicher [1] investigated the continuous Newton–Raphson method in the context of barrier minimization in linear programming. This is the same as the continuous Newton method applied to the zero-finding problem defined by the barrier gradient field.

Applied to the problem of minimizing an arbitrary nonlinear function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, the Newton–Raphson flow equations are defined by

$$(1.1) \quad \frac{\partial}{\partial t} \varphi(x_0, t) = -H(\varphi(x_0, t))^{-1} \nabla f(\varphi(x_0, t)), \quad \varphi(x_0, 0) = x_0,$$

where

$$\begin{aligned} \nabla f(y) &:= D_x f(x)^T|_{x=y} \in \mathbb{R}^n, \\ H(y) &:= D_x^2 f(x)|_{x=y} \in \mathbb{R}^{n \times n}. \end{aligned}$$

*Received by the editors August 4, 2003; accepted for publication (in revised form) June 15, 2004; published electronically April 22, 2005.

<http://www.siam.org/journals/siopt/15-3/43263.html>

[†]Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford, OX1 3QD, England (hauser@comlab.ox.ac.uk, jelena@comlab.ox.ac.uk). The first author was supported by EPSRC grant GR/S34472. The second author was supported by the Clarendon Fund, Oxford University Press, and ORS Award, Universities UK.

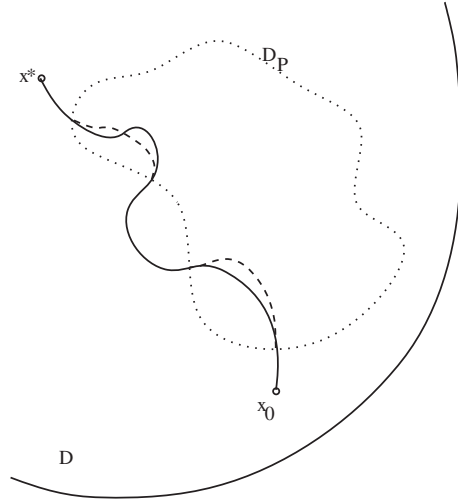


FIG. 1. Flux lines of the original function f (solid line) and of the function $\hat{f}$ which is perturbed in the dotted region $\mathcal{D}_P$ (dashed line).

The continuous Newton–Raphson method then consists of choosing a starting point x_0 and following the flux line $t \mapsto \varphi(x_0, t)$, obtained by integrating the ODE (1.1) to a stationary point $x^* = \lim_{t \rightarrow \infty} \varphi(x_0, t)$ of ∇f .

Janovsky and Seige [5] showed that, subject to certain regularity conditions, the topological behavior of the Newton flow associated with the continuous Newton method for solving systems of equations is invariant under small perturbations of the system. In this paper we consider a more restrictive setting and prove a similar but stronger property: we consider the continuous Newton–Raphson method to minimize f applied at a point x_0 in the domain of attraction of a strict local minimizer x^* . Under these conditions, if $\hat{f}$ is a benign perturbation of f in an open domain $\mathcal{D}_P$ that excludes a small neighborhood of x_0 , and if $\hat{\varphi}(x_0, t)$ denotes the flux line flowing through x_0 under the perturbed dynamical system

$$(1.2) \quad \frac{\partial}{\partial t} \hat{\varphi}(x_0, t) = -\hat{H}(\hat{\varphi}(x_0, t))^{-1} \nabla \hat{f}(\hat{\varphi}(x_0, t)), \quad \hat{\varphi}(x_0, 0) = x_0,$$

with $\nabla \hat{f} = D_x \hat{f}$ and $\hat{H} = D_x^2 \hat{f}$, then $\varphi(x_0, t)$ and $\hat{\varphi}(x_0, t)$ differ only for values of t for which $\varphi(x_0, t), \hat{\varphi}(x_0, t) \in \mathcal{D}_P$. In other words, the dynamics of φ and $\hat{\varphi}$ not only are topologically equivalent, but the flux lines actually coincide outside of the domain of perturbation; see Figure 1.

This property is intriguing, because it seems counterintuitive: $\varphi(x_0, t)$ and $\hat{\varphi}(x_0, t)$ clearly must coincide for values of t that occur before first entering $\mathcal{D}_P$, because (1.1) and (1.2) coincide on this range. However, one might expect that $\hat{\varphi}$ starts to differ from φ inside $\mathcal{D}_P$ and leads to an exit point $\hat{y} = \hat{\varphi}(x_0, \hat{\tau}) \in \partial \mathcal{D}_P$ from $\mathcal{D}_P$ which is different from the point $y = \varphi(x_0, \tau) \in \partial \mathcal{D}_P$ where φ exits. If that was the case, then the remaining parts of the flux lines

$$\begin{aligned} \varphi(x_0, t) &= \varphi(y, t - \tau), & (t > \tau), \\ \hat{\varphi}(x_0, t) &= \hat{\varphi}(\hat{y}, t - \hat{\tau}), & (t > \hat{\tau}), \end{aligned}$$

would differ from one another.

This intuitive picture is wrong, because a simple lemma shows that the normalized gradient

$$\mathbf{n}(\nabla f(y)) := \frac{\nabla f(y)}{\|\nabla f(y)\|}$$

is an invariant for points y that lie along the flux line $\varphi(x_0, t)$; see Lemma 2.1. The same of course holds true for $\mathbf{n}(\nabla \hat{f})$. This mechanism forces the exit points x and $\hat{x}$ and the exit times τ and $\hat{\tau}$ to coincide, which in turn implies that φ and $\hat{\varphi}$ will perfectly coincide after exiting $\mathcal{D}_P$, until they next enter the domain of perturbation.

If $\mathcal{D}_P$ furthermore excludes a small neighborhood of x^* , then $\varphi(x_0, t)$ and $\hat{\varphi}(x_0, t)$ coincide for all t large enough. Thus, the continuous Newton–Raphson method “knows” how it will approach x^* solely on the basis of information about f taken from neighborhoods around x_0 and x^* ; the function values of f at points that are positioned along the way have no bearing on this. For example, the limiting direction at which $\varphi(x_0, t)$ approaches x^* is given by

$$\lim_{t \rightarrow +\infty} e^t \frac{\partial}{\partial t} \varphi(x_0, t) = -H(x^*)^{-1} \nabla f(x_0);$$

see Theorem 3.1. Thus, in this specific sense the continuous Newton–Raphson method can look ahead: it knows where it is going and can only be distracted but not ultimately deviated from its predetermined path by local perturbations.

For the purposes of our analysis we will assume throughout that x^* is a point such that $\nabla f(x^*) = 0$ and the Hessian at x^* is positive definite ($H(x^*) \succ 0$) implying that x^* is a strict local minimizer of f , and that $f \in \mathcal{C}^3(\mathcal{D}_3)$ in some open neighborhood $\mathcal{D}_3$ of x^* . The domain of attraction of x^* is the set of points $x \in \mathbb{R}$ such that $\lim_{t \rightarrow \infty} \varphi(x, t) = x^*$. It is well known that the assumption $H(x^*) \succ 0$ implies that the domain of attraction of x^* contains an open neighborhood $\mathcal{D}_a$ of x^* such that $H(x) \succ 0$ for all $x \in \mathcal{D}_a$. We will use the notation $\text{int}A$, ∂A , and $\text{cl}A$ for the interior, boundary, and closure of a set A , respectively. Additionally, for convex set A we will use $\text{ext}A$ to denote the set of extreme points of A , that is to say, the set of all points $x \in A$ such that $A \setminus \{x\}$ is convex.

2. Characterization of flux lines by an invariance property. A simple but powerful property of the Newton–Raphson flow is that $\mathbf{n}(\nabla f)$ is invariant along flux lines. This claim, proved in the next Lemma, has the interesting consequences that we will formalize in the main theorems.

LEMMA 2.1. *The direction of the gradient of the function f is constant along flux lines defined by its Newton–Raphson vector field.*

Proof. Let $\varphi(x_0, t)$ denote the flux line of the Newton–Raphson vector field defined by f passing through x_0 . Then

$$\begin{aligned} \frac{\partial}{\partial t} \nabla f(\varphi(x_0, t)) &= H(\varphi(x_0, t)) \frac{\partial}{\partial t} \varphi(x_0, t) \\ &= H(\varphi(x_0, t)) (-H(\varphi(x_0, t))^{-1}) \nabla f(\varphi(x_0, t)) \\ &= -\nabla f(\varphi(x_0, t)). \end{aligned}$$

Therefore,

$$(2.1) \quad \nabla f(\varphi(x_0, t)) = e^{-t} \nabla f(\varphi(x_0, 0))$$

and $\|\nabla f(\varphi(x_0, t))\| = e^{-t}\|\nabla f(\varphi(x_0, 0))\|$, that is,

$$\mathbf{n}(\nabla f(\varphi(x_0, t))) = \mathbf{n}(\nabla f(\varphi(x_0, 0))),$$

which proves the claim. $\square$

3. Limiting behavior of flux lines. As a direct consequence of Lemma 2.1 we now obtain the following result.

THEOREM 3.1. *Let $x_0 \in \mathcal{D}$. Then the limiting direction of the flux line $\varphi(x_0, t)$ is the normalization of the vector*

$$\lim_{t \rightarrow +\infty} e^t \frac{\partial}{\partial t} \varphi(x_0, t) = -H(x^*)^{-1} \nabla f(x_0).$$

Proof. Using (2.1) we obtain

$$\lim_{t \rightarrow +\infty} e^t \frac{\partial}{\partial t} \varphi(x_0, t) = - \lim_{t \rightarrow +\infty} H(\varphi(x_0, t))^{-1} \nabla f(x_0) = -H(x^*)^{-1} \nabla f(x_0),$$

where the last equality follows from the fact that flux line $\varphi(x_0, t)$ converges to x^* as $t \rightarrow \infty$ and the assumption that the Hessian is a continuous function. $\square$

Theorem 3.1 shows that the direction at which the flux line $\varphi(x_0, t)$ approaches x^* depends only on the gradient of f at x_0 and the Hessian at x^* . In other words, knowing only the behavior of the function f in a small neighborhood around x_0 and x^* , one can determine the limiting direction of a flux line; see Figure 2. Since the limiting direction does not depend on the value of f at intermediate points along the flux line $\varphi(x_0, t)$, does this mean that if we perturb f in a region that lies in its path, the behavior of $\varphi(x_0, t)$ around x^* will be the same? If the perturbation is sufficiently benign, the answer is yes. As a matter of fact, a much stronger claim is true: the flux lines of f and its perturbation differ only in the region of perturbation! This is the subject of the next section.

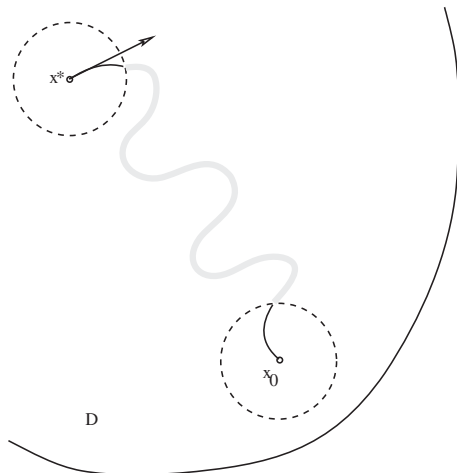


FIG. 2. The limiting direction of the flux line through x_0 depends only on information in small neighborhoods of x_0 and x^* .

4. Change of flux lines under perturbations of function f . In this section we will prove the main theorem of this paper, which we described in the introduction as counterintuitive. Thinking back to Lemma 2.1 and its use in the proof of Theorem 3.1, the result we claim here is now more plausible: the normalized gradient of the f along $\varphi(x_0, t)$ takes the constant value $\mathbf{n}(\nabla f(x_0))$. Moreover, under the regularity assumption we have made the set of points where $\mathbf{n}(\nabla f) = \mathbf{n}(\nabla f(x_0))$ is one-dimensional. Since f and its perturbation $\hat{f}$ are identical in the unperturbed region, the loci of the points on the flux lines $\varphi(x_0, t)$ and $\hat{\varphi}(x_0, t)$ have to lie in the same one-dimensional set, forcing them to coincide.

Before formalizing and proving this result, we need the following definitions and auxiliary results.

DEFINITION 4.1. Let $C \subset \mathbb{R}^n$ and $x \in \partial C$. A supporting hyperplane of C at x is $S_C(x) = \{y \in \mathbb{R}^n : \langle a; y - x \rangle = 0\}$, where $0 \neq a \in \mathbb{R}^n$ is a vector such that for all $y \in C$, $\langle a; y - x \rangle \leq 0$.

In other words, a supporting hyperplane of a set C at x is a hyperplane containing x such that the whole set C lies in one closed halfspace defined by the hyperplane. It is well known (see, e.g., [3]) that if C is convex, then a supporting hyperplane exists for every $x \in \partial C$.

DEFINITION 4.2. Let $C \subset \mathbb{R}^n$ be a convex set and $x \in C$. The normal cone of C at x is $N_C(x) = \{d \in \mathbb{R}^n : \langle d; y - x \rangle \leq 0 \ \forall y \in C\}$. The tangent cone of C at x is $T_C(x) = \{\alpha(c - x) : \alpha \in \mathbb{R}_+, c \in C\}$.

DEFINITION 4.3. Let $A \subset \mathbb{R}^n$. The polar cone of A is

$$A^- = \{y \in \mathbb{R}^n : \langle y; x \rangle \leq 0 \ \forall x \in A\}.$$

It is not difficult to prove that $N_C(x) = T_C(x)^-$ and $T_C(x) = N_C(x)^-$; see, e.g., [3].

Using the notions defined above, we can now show that for a function with a strictly convex level set L_k , that is, $\partial L_k = \text{ext}(L_k)$, there exists a unique point in ∂L_k where ∇f has a given direction.

LEMMA 4.4. Let $f \in \mathcal{C}^2$ be a function with positive definite Hessian $H \succ 0$ for every x in some neighborhood $\mathcal{D}$ of a strict local minimizer x^* , and let g be a given unit vector. Then for any fixed level $k > f(x^*)$ such that $L_k^f := \{x : f(x) \leq k\} \subseteq \mathcal{D}$ there exists a unique point $x_k \in \partial L_k$ such that $\mathbf{n}(\nabla f(x_k)) = g$.

Proof. It is easy to prove that the condition $H(x) \succ 0$ on $\mathcal{D}$ implies that all level sets $L_k^f = \{x : f(x) \leq k\}$ such that $L_k \subseteq \mathcal{D}$ are strictly convex, that is to say, $\partial L_k^f = \text{ext} L_k^f$.

Existence. Let us first find a point $x_k \in \partial L_k^f$, where g defines a supporting hyperplane. Since we want $\langle g; y - x_k \rangle = \langle g; y \rangle - \langle g; x_k \rangle \leq 0$ to hold for every $y \in L_k^f$, let us choose $x_k \in \text{argmax}\{\langle g; x \rangle : x \in L_k^f\}$. We will prove that $\mathbf{n}(\nabla f(x_k)) = g$. It is easy to see that $g \in N_{L_k^f}(x_k)$ and that $\nabla f(x_k) \neq 0$. By the implicit function theorem, ∂L_k^f is a $\mathcal{C}^1$ manifold and hence $T_{\partial L_k^f}(x_k)$ is well defined and of dimension $n - 1$. Therefore $T_{L_k^f}(x_k)$ is a linear halfspace and $N_{L_k^f}(x_k) = T_{L_k^f}(x_k)^-$ is a ray (a one-dimensional pointed cone). Since $\nabla f(x_k) \in N_{L_k^f}(x_k)$, this shows that $\mathbf{n}(\nabla f(x_k)) = g$; see Figure 3.

Uniqueness. Let us now assume that there exist two points x_k and y_k satisfying the claim. Then, both the supporting hyperplanes for x_k and y_k are defined by g . The definition of $S_{L_k^f}(x_k)$ implies that $\langle g; y_k - x_k \rangle \leq 0$, while $\langle g; x_k - y_k \rangle \leq 0$ follows

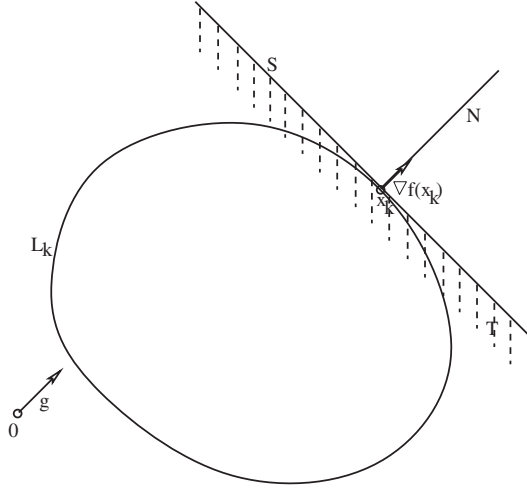


FIG. 3. Supporting hyperplane S defined by the vector g , normal cone N , and tangent cone T of level set L_k at x_k

from the definition of $S_{L_k^f}(y_k)$. Hence $\langle g; x_k - y_k \rangle = 0$ and $y_k \in S_{L_k^f}(x_k)$. But this implies that the entire line segment defined by x_k and y_k belongs to ∂L_k^f , so for every $\lambda \in (0, 1)$ we have that $\lambda x_k + (1 - \lambda)y_k \in \partial L_k^f \setminus \text{ext} L_k^f$, which contradicts the assumption that L_k^f is strictly convex. $\square$

Hereafter we will say that the function $\hat{f}$ is a *regular perturbation* of $f \in \mathcal{C}^3$ with local minimizer x^* if the following conditions are satisfied:

- (i) $\hat{f} \in \mathcal{C}^2$ and $H_{\hat{f}} \succ 0 \forall x \in \mathcal{D}$;
- (ii) $f(x) - \hat{f}(x) \neq 0 \Rightarrow x \in \mathcal{D}_P \subset \mathcal{D}$, where $\mathcal{D}_P$ is an open set.

Note that the condition (i) ensures that the Newton–Raphson vector field is well defined and no bifurcation of flux lines occurs. We will use the notation

$$\Gamma_{\varphi}(x_0) = \{x : x = \varphi(x_0, t) \text{ for some } t \geq 0\}$$

to denote the flux line $\varphi(x_0, t)$ as a set of points, and we write $A \triangle B$ for the symmetric difference between two sets A and B .

All the tools are now available to formalize and prove the main result of this paper: the Newton–Raphson–flux lines defined by a regular perturbation $\hat{f}$ of f differ from the Newton–Raphson–flux lines defined by f only in the region of perturbation $\mathcal{D}_P$.

THEOREM 4.5. *Let f be a $\mathcal{C}^2$ function with $H_f \succ 0$ for all $x \in \mathcal{D}$, and let $x^* \in \mathcal{D}$ be a strict local minimizer of f . Let $\hat{f}$ be a regular perturbation of f , and let $\varphi(x_0, t)$ and $\hat{\varphi}(x_0, t)$ be the Newton–Raphson flows defined by f and $\hat{f}$. Let x_0 be a point in $\mathcal{D} \setminus \text{cl}(\mathcal{D}_P)$ such that $L_{f(x_0)}^f \subset \mathcal{D}$ and $L_{\hat{f}(x_0)}^{\hat{f}} \subset \mathcal{D}$. Then*

$$\Gamma_{\varphi}(x_0) \triangle \Gamma_{\hat{\varphi}}(x_0) \subset \mathcal{D}_P$$

and $\varphi(x_0, t) = \hat{\varphi}(x_0, t)$ for all $t \in \mathbb{R}$ such that $\varphi(x_0, t) \notin \mathcal{D}_P$.

Proof. Let $g = \mathbf{n}(\nabla f(x_0)) = \mathbf{n}(\nabla \hat{f}(x_0))$, where the latter equality holds because $f = \hat{f}$ in a neighborhood of x_0 . We claim that

$$\begin{aligned}\Gamma_\varphi(x_0) &= \{x \in \mathcal{D} : \mathbf{n}(\nabla f(x)) = g, f(x) \leq f(x_0)\} =: \Upsilon_g(x_0), \\ \Gamma_{\hat{\varphi}}(x_0) &= \{x \in \mathcal{D} : \mathbf{n}(\nabla \hat{f}(x)) = g, \hat{f}(x) \leq \hat{f}(x_0)\} =: \hat{\Upsilon}_g(x_0).\end{aligned}$$

Since $H_f \succ 0$ and $H_{\hat{f}} \succ 0$, level sets L_k^f and $L_k^{\hat{f}}$ are strictly convex in $\mathcal{D}$ for every $k \leq f(x_0)$. Therefore $\Gamma_\varphi(x_0) \subseteq \Upsilon_g(x_0)$ by Lemma 2.1. On the other hand, assume that there exists a point $y \in \Upsilon_g(x_0) \setminus \Gamma_\varphi(x_0)$. We know that $x_0 \in \Gamma_\varphi(x_0)$ so $y \neq x_0$. Since f is continuous and strictly decreasing along $\varphi(x_0, t)$ and since $\varphi(x_0, t)$ converges to x^* , there exists a point $x_1 \in \Gamma_\varphi(x_0)$ such that $f(x_1) = f(y) < f(x_0)$ and $\mathbf{n}(\nabla f(x_1)) = \mathbf{n}(\nabla f(x_0)) = \mathbf{n}(\nabla f(y))$, where the first equality follows from the Lemma 2.1 and the second from the fact that $y \in \Upsilon_g(x_0)$. This implies that there are two distinct points in the level set $L_{f(y)}^f$ with the same gradient direction, in contradiction to Lemma 4.4. Therefore, $\Gamma_\varphi(x_0) = \Upsilon_g(x_0)$. The result for the function $\hat{f}$ can be derived in a similar fashion. Now, since $f(x) = \hat{f}(x)$ for $x \in \mathcal{D} \setminus \mathcal{D}_P$ and in particular $f(x_0) = \hat{f}(x_0)$, it is easy to see that $\Upsilon_g(x_0)$ and $\hat{\Upsilon}_g(x_0)$ are the same in the unperturbed region, i.e., $\Upsilon_g(x_0) \cap (\mathcal{D} \setminus \mathcal{D}_P) = \hat{\Upsilon}_g(x_0) \cap (\mathcal{D} \setminus \mathcal{D}_P)$. In other words, $\Gamma_\varphi(x_0) \cap (\mathcal{D} \setminus \mathcal{D}_P) = \Gamma_{\hat{\varphi}}(x_0) \cap (\mathcal{D} \setminus \mathcal{D}_P)$, showing the first claim. The last claim now follows from (2.1). $\square$

5. Numerical examples.

5.1. A two-dimensional example. Let us consider the quadratic function $f(x) = c + b^T x + \frac{1}{2} x^T B x$ for $c = 2$, $b = (1, 2)^T$, and $B = \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}$ with global minimizer $x^* = (6/7, -5/7)^T$ and perturb it in an annulus around x^* with inner and outer radii r_a and r_b , respectively. Let us choose a perturbation term given by a polynomial of degree 8 with 4-tuple zeros r_a and r_b : $\hat{f}(x) = f(x) + p(\|x - x^*\|)$, where

$$(5.1) \quad p(s) = \begin{cases} \varepsilon(s - r_a)^4(s - r_b)^4 & \text{for } r_a < s < r_b, \\ 0 & \text{otherwise.} \end{cases}$$

The figures in this section show the results for $r_a = 1$, $r_b = 3$, and $\varepsilon = 1$ obtained in numerical computations performed in MATLAB using the ODE integrator `ode45`. Figure 4 shows a few level sets of the perturbed function $\hat{f}$, the annulus where the original function was perturbed, and the flux lines φ and $\hat{\varphi}$, respectively, corresponding to f (solid lines) and $\hat{f}$ (dashed curves). It can be clearly seen that the flux lines differ only in the region of perturbation, as proved in Theorem 4.5. In fact, this example shows that the convexity assumptions of Theorem 4.5 are sufficient but not necessary for its claim to hold.

Note that the time scale of the flux lines for f and $\hat{f}$ is in general different in the perturbed region. Therefore, to measure the difference between points on $\Gamma_\varphi(x_0)$ and $\Gamma_{\hat{\varphi}}(x_0)$, we have reparametrized the flux lines in terms of the distance to x^* . Accordingly, Figure 5 shows the norm of the difference of flux lines shown in Figure 4 as functions of $\|x - x^*\|$. The flux lines differ only in the region $1 < \|x - x^*\| < 3$, confirming the claim of Theorem 4.5. Note that $p(s)$ has exactly one local extremum s^* for $r_a < s < r_b$. Therefore, for $\bar{x}$ such that $\|\bar{x} - x^*\| = s^*$ we have $\nabla \hat{f}(\bar{x}) = \nabla f(\bar{x}) + p'(s^*) \frac{d}{dx} \|x - x^*\| |_{x=\bar{x}} = \nabla f(\bar{x})$; so recalling the proof of Theorem 4.5 we see that the perturbed and unperturbed flux lines intersect at such points, which explains the dip observed in Figure 5.

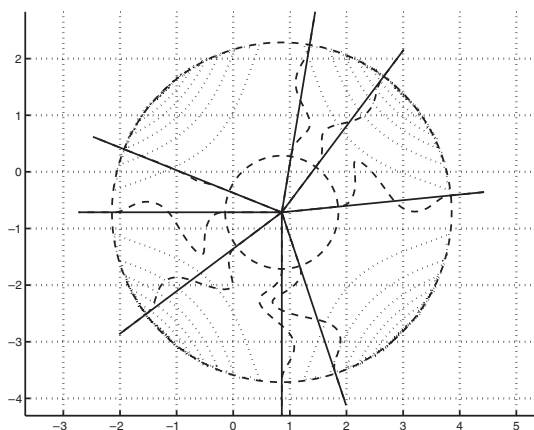


FIG. 4. *Perturbed region, level sets of the perturbed function, and flux lines of the original (solid line) and the perturbed (dashed curves) function.*

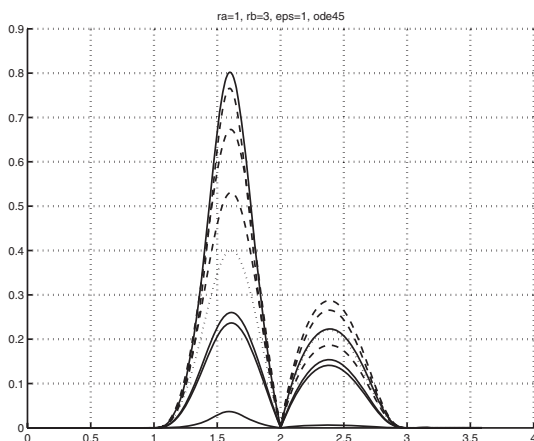


FIG. 5. *Norms of the difference of flux lines for f and $\hat{f}$ as functions of $\|x - x^*\|$.*

To emphasize the fact that this property is closely related to the Newton–Raphson vector field and does not necessarily hold for all vector fields, the same tests were performed on a negative gradient field, and the results are shown in Figures 6 and 7. It is clear that flux lines for the perturbed function exit the region of perturbation at the points different from exit points of the flux lines for the original function and therefore converge to the solution x^* following different flows.

5.2. A three-dimensional example. The following three-dimensional example was constructed in a similar fashion. Let us again consider the quadratic function $f(x) = c + b^T x + \frac{1}{2} x^T B x$ and its perturbation $\hat{f} = f(x) + p(\|x - x^*\|)$ which differs from f on a sphere of inner radius r_a and outer radius r_b around the exact minimizer x^* . The perturbation term is given by the polynomial $p(s)$ defined in (5.1).

The particular values $r_a = 1$, $r_b = 3$, $\varepsilon = 1$, $c = 2$, $b = (2, -5, 3)^T$, and

$$B = \begin{pmatrix} 5 & 2 & 1 \\ 2 & 4 & 1 \\ 1 & 1 & 4 \end{pmatrix}$$

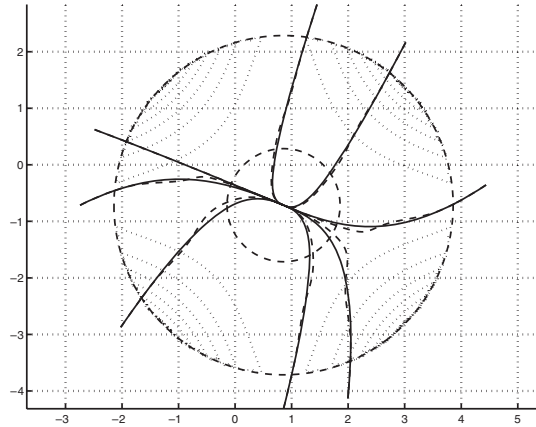


FIG. 6. Flux lines of the negative gradient field for the original (solid line) and the perturbed (dashed curves) function.

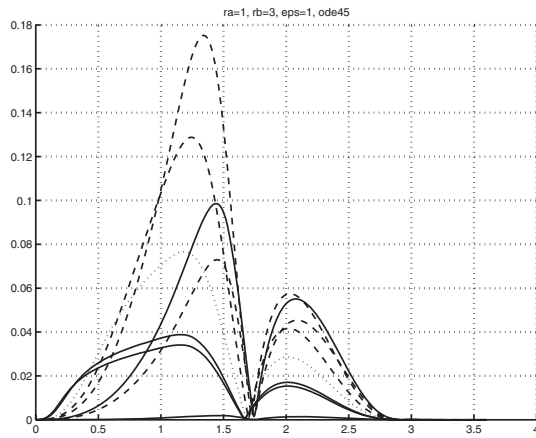


FIG. 7. Norms of the difference of the negative gradient field flux lines for f and $\hat{f}$ as functions of $\|x - x^*\|$.

were used in the numerical experiments. In this case the exact minimizer is $x^* = (1, -2, 1)^T$.

Figures 8 and 9 show that the flux lines for f and $\hat{f}$ again differ only in the region of perturbation.

6. Discussion. The main theorems of this paper highlight an interesting property of the global dynamics of the continuous Newton–Raphson method. For the purposes of a succinct analysis we assumed that f and its perturbation $\hat{f}$ are both strictly convex functions. This assumption is merely a sufficient condition that could be relaxed in many ways. In fact, using the topological results of Janovsky and Seige [5], it is possible in principle to extend the results presented here to the case where perturbations are small and where neither $H_f(x)$ nor $H_{\hat{f}}(x)$ is singular anywhere along $\varphi(x_0, t)$ and $\hat{\varphi}(x_0, t)$, guaranteeing that the flux lines cannot bifurcate. Our numerical experiments confirm this where we have chosen nonconvex examples which tend to make the claim of Theorem 4.5 more clearly visible than convex examples.

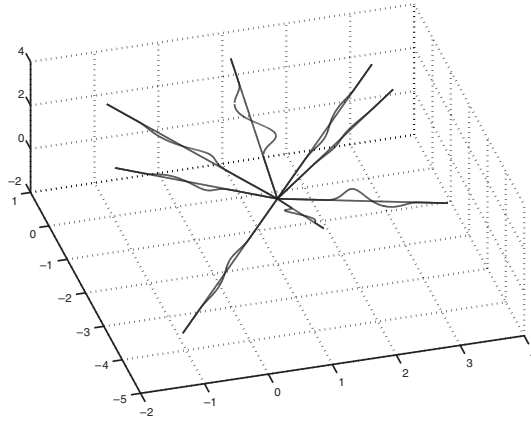
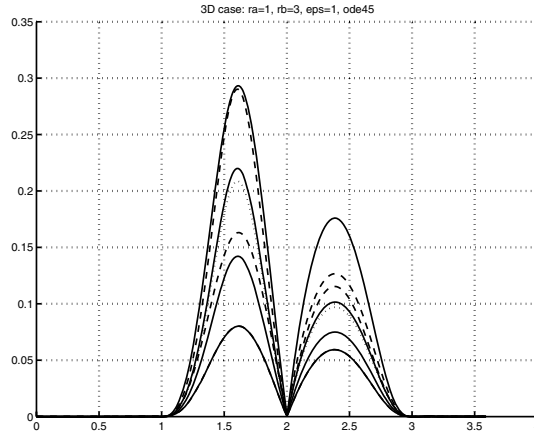


FIG. 8. Flux lines of original and perturbed function in the three-dimensional case.

FIG. 9. Norms of a difference of flux lines for f and $\hat{f}$ as functions of $\|x - x^*\|$ in the three-dimensional case.

As pointed out by one of the referees, Theorem 4.5 can be proved without resorting to the tools of convex analysis.

LEMMA 6.1. *Let $f \in C^2$ be a function with positive definite Hessian $H \succ 0$ for every x in some neighborhood $\mathcal{D}$ of a strict local minimizer x^* , and $y, z \in \mathcal{D}$. If $\nabla f(y) = \nabla f(z)$, then $y = z$.*

Proof. First, note that the condition $H(x) \succ 0$ on $\mathcal{D}$ implies that the function f is strictly convex on $\mathcal{D}$. Let $p := \nabla f(y) = \nabla f(z)$ and let us define a new function $g(x) = f(x) - \langle p; x \rangle$. Since g is obtained by adding a linear term to a strictly convex function f , g is also strictly convex. By straightforward calculation one obtains that $\nabla g(y) = \nabla g(z) = 0$, which, together with strict convexity of g , implies that both y and z are global minimizers of g and therefore $g(y) = g(z)$. If $y \neq z$, this implies that for every point $w = \lambda y + (1 - \lambda)z$, $\lambda \in (0, 1)$ one gets $g(y) = g(w) = g(z)$, which is a contradiction with g being strictly convex. Therefore $y = z$ must hold. $\square$

This approach yields a slightly more general statement of Theorem 4.5 obtained by excluding the conditions $L_{f(x_0)}^f \subset \mathcal{D}$ and $L_{\hat{f}(x_0)}^{\hat{f}} \subset \mathcal{D}$.

THEOREM 6.2. Let f be a C^2 function with $H_f \succ 0$ for all $x \in \mathcal{D}$, and let $x^* \in \mathcal{D}$ be a strict local minimizer of f . Let $\hat{f}$ be a regular perturbation of f , and let $\varphi(x_0, t)$ and $\hat{\varphi}(x_0, t)$ be the Newton-Raphson-flows defined by f and $\hat{f}$. Let x_0 be a point in $\mathcal{D} \setminus cl(\mathcal{D}_P)$. Then

$$\Gamma_\varphi(x_0) \Delta \Gamma_{\hat{\varphi}}(x_0) \subset \mathcal{D}_P$$

and $\varphi(x_0, t) = \hat{\varphi}(x_0, t)$ for all $t \in \mathbb{R}$ such that $\varphi(x_0, t) \notin \mathcal{D}_P$ or $\hat{\varphi}(x_0, t) \notin \mathcal{D}_P$.

Proof. Suppose $t > 0$ is such that $\varphi(x_0, t) \notin \mathcal{D}_P$. Then

$$\nabla \hat{f}(\varphi(x_0, t)) = \nabla f(\varphi(x_0, t)) \stackrel{(2.1)}{=} e^{-t} \nabla f(x_0) = e^{-t} \nabla \hat{f}(x_0) \stackrel{(2.1)}{=} \nabla \hat{f}(\hat{\varphi}(x_0, t)),$$

where the first and the third equality hold because $f = \hat{f}$ outside the region of perturbation $\mathcal{D}_P$. Since $\hat{f}$ is strictly convex and $\nabla \hat{f}(\varphi(x_0, t)) = \nabla \hat{f}(\hat{\varphi}(x_0, t))$, Lemma 6.1 implies that $\varphi(x_0, t) = \hat{\varphi}(x_0, t)$. In a similar fashion, one can show that if $\hat{\varphi}(x_0, t) \notin \mathcal{D}_P$, then $\nabla f(\hat{\varphi}(x_0, t)) = \nabla f(\varphi(x_0, t))$, and therefore strict convexity of f and Lemma 6.1 imply that again $\varphi(x_0, t) = \hat{\varphi}(x_0, t)$. This shows that the flux lines can differ only inside the region of perturbation, which concludes the proof. $\square$

The authors would like to thank the anonymous referee for this interesting observation.

REFERENCES

- [1] K. ANSTREICHER, *Linear programming and the Newton-barrier flow*, Math. Programming, 41 (1988), pp. 367–373.
- [2] J. M. BALL, *Continuity properties and global attractors of generalized semiflows and the Navier-Stokes equations*, Nonlinear Sci., 7 (1997), pp. 475–502.
- [3] J. M. BORWEIN AND A. S. LEWIS, *Convex Analysis and Nonlinear Optimization: Theory and Examples*, Springer, New York, 2000.
- [4] F. H. BRANIN, *A widely convergent method for finding multiple solutions of simultaneous nonlinear equations*, IBM J. Res. Develop., 16 (1972), pp. 504–522.
- [5] V. JANOVSKÝ AND V. SEIGE, *Qualitative analysis of Newton's flow*, SIAM J. Numer. Anal., 33, (1996), pp. 2068–2097.
- [6] H. TH. JONGEN, P. JONKER, AND F. TWILT, *The continuous desingularized Newton method for meromorphic functions*, Acta Appl. Math., 13 (1988), pp. 81–121.
- [7] J. W. NEUBERGER, *Continuous Newton's method for polynomials*, Math. Intelligencer, 21 (1999), pp. 18–23.
- [8] J. W. NEUBERGER, *Sobolev Gradients and Differential Equations*, Lecture Notes in Math. 1670, Springer-Verlag, New York, 1997.
- [9] S. SMALE, *A convergence process in price adjustment and global Newton methods*, J. Math. Econ., 3 (1976), pp. 107–120.

EXTENSIONS OF THE KUHN–TUCKER CONSTRAINT QUALIFICATION TO GENERALIZED SEMI-INFINITE PROGRAMMING*

FRANCISCO GUERRA VÁZQUEZ[†] AND JAN-J. RÜCKMANN[‡]

Abstract. This paper deals with the class of generalized semi-infinite programming problems (GSIPs) in which the index set of the inequality constraints depends on the decision vector and all emerging functions are assumed to be continuously differentiable. We introduce two extensions of the Kuhn–Tucker constraint qualification (which is based on the existence of a tangential continuously differentiable arc) to the class of GSIPs, prove a corresponding Karush–Kuhn–Tucker theorem, and discuss its assumptions. Finally, we present several examples which illustrate for the class of GSIPs some interrelations between the considered extensions of the Mangasarian–Fromovitz constraint qualification, the Abadie constraint qualification, and the Kuhn–Tucker constraint qualification.

Key words. generalized semi-infinite programming, extended Kuhn–Tucker constraint qualification, extended Mangasarian–Fromovitz constraint qualification, extended Abadie constraint qualification, Karush–Kuhn–Tucker theorem, local minimizer

AMS subject classifications. 90C34, 90C26, 90C46, 65K10

DOI. 10.1137/S1052623403431500

1. Introduction. This paper deals with the class of generalized semi-infinite programming problems (GSIPs), which recently became a very active research area in mathematical programming; e.g., see the survey paper [7], the recent monographs [5, 17, 24], as well as the compilations [6, 18], which include various tutorial papers on theory, numerics, and applications in generalized semi-infinite programming.

As a starting point of this paper we define a GSIP as follows:

$$(1.1) \quad \text{Minimize } f(x) \text{ subject to } x \in M$$

with the feasible set

$$M = \{x \in \mathbb{R}^n \mid G(x, y) \geq 0, y \in Y(x)\}$$

and the index set

$$Y(x) = \{y \in \mathbb{R}^r \mid u_k(x, y) = 0, k \in K, v_l(x, y) \geq 0, l \in L\},$$

where K and L are finite sets with $|K| < r$ ($|\cdot|$ denotes the cardinality) and all emerging functions are assumed to be real-valued and continuously differentiable. If the index set does not depend on x (i.e., $Y(x) = Y$ for all $x \in \mathbb{R}^n$), then problem (1.1) reduces to a standard semi-infinite programming problem.

It is well known in mathematical programming that one of the main elements for obtaining necessary first order optimality conditions of Karush–Kuhn–Tucker

*Received by the editors July 14, 2003; accepted for publication (in revised form) July 16, 2004; published electronically April 22, 2005. This work was partially supported by CONACyT (Mexico) grant 44003.

<http://www.siam.org/journals/siopt/15-3/43150.html>

[†]Universidad de las Américas, Escuela de Ciencias, Sta. Catarina Mártir, Cholula, Puebla 72820, México (fguerra@mail.udlap.mx).

[‡]Universidad de las Américas, Departamento de Física y Matemáticas, Sta. Catarina Mártir, Cholula, Puebla 72820, México (rueckman@mail.udlap.mx).

type is an appropriate constraint qualification, e.g., an Abadie constraint qualification (ACQ), Mangasarian–Fromovitz constraint qualification (MFCQ), or Kuhn–Tucker constraint qualification (KTCQ). Although most of these constraint qualifications have been originally used in the framework of finite programming (see, e.g., [1, 4, 13, 15] as well as [16] for an exhaustive review of constraint qualifications in finite-dimensional spaces), some of them have been generalized to a broader class of problems; e.g., a generalization of MFCQ to standard semi-infinite programming is used in [7, 8, 12] and an extension of KTCQ to infinite programming is proposed in [26]. For our class of GSIPs, first order necessary optimality conditions of the Fritz–John type and a corresponding Karush–Kuhn–Tucker theorem have been stated in [11]; in particular, an appropriate extension of MFCQ was introduced there: the extended Mangasarian–Fromovitz constraint qualification (EMFCQ) for GSIP (cf. section 2 of this paper). Since then, several results on necessary and sufficient optimality conditions and on constraint qualifications for GSIP have been proved (cf., e.g., [19, 20, 21, 24, 25]). In particular, we mention the recent papers [22] and [23]: in [22], a generalization of ACQ to GSIP is presented (the extended Abadie constraint qualification (EACQ); cf. section 2) and in [23] a unified framework for several constraint qualifications in semi-infinite programming is discussed by using a nonsmooth approach.

Returning to our paper, its goal is twofold. First, we define two appropriate extensions, EKTCQ1 and EKTCQ2, of KTCQ to GSIP and present corresponding first order optimality conditions of Karush–Kuhn–Tucker type. Then, we discuss interrelations between the following constraint qualifications for GSIP: EMFCQ, EACQ, EKTCQ1, and EKTCQ2.

The paper is organized as follows. Section 2 contains definitions and basic results, as well as lemmas, which are used later. In particular, the above-mentioned extensions of KTCQ to GSIP are introduced. Then, in section 3, a Karush–Kuhn–Tucker theorem is proved and its assumptions are discussed. Furthermore, several examples are presented which illustrate some interrelations between the considered constraint qualifications. Section 4 summarizes the results and presents some conclusions.

2. Auxiliary results, EKTCQ1 and EKTCQ2. Throughout the paper let $\mathbb{R}^n$ be the n -dimensional Euclidean space and $x_i, i = 1, \dots, n$, the components of $x \in \mathbb{R}^n$. Denote by $C^1(\mathbb{R}^n, \mathbb{R})$ the set of continuously differentiable real-valued functions defined on $\mathbb{R}^n$. If $g \in C^1(\mathbb{R}^n, \mathbb{R})$, then $Dg(\bar{x})$ ($D_{x^1}g(\bar{x})$) represents the derivative (partial derivative with respect to the subvector x^1 of x) of g at $\bar{x} \in \mathbb{R}^n$.

The GSIP. Let the GSIP be defined as in (1.1) with the feasible set M and the index set $Y(x)$, which represents the inequality constraints of M . Assume that

- the set $Y(x)$ is compact for all $x \in \mathbb{R}^n$ and
- the set-valued mapping $Y : x \in \mathbb{R}^n \mapsto Y(x) \subset \mathbb{R}^r$ is upper semicontinuous at each $x \in \mathbb{R}^n$ in the sense of Berge [3].

These two conditions on $Y(x)$ are equivalent to a uniform compactness assumption for $Y(x)$ at each $x \in \mathbb{R}^n$ (cf., e.g., [9]).

For $\bar{x} \in M$ we define the index set of active constraints at $\bar{x}$ as

$$Y_0(\bar{x}) = \{y \in Y(\bar{x}) \mid G(\bar{x}, y) = 0\}.$$

For the sake of simplicity we omitted equality constraints in the description of M . All the results of this paper can easily be generalized to the more general case, where the description of M contains additionally finitely many equality constraints.

The lower level problem. Let $\bar{x} \in M$ and $\bar{y} \in Y_0(\bar{x})$. Then $\bar{y}$ is a (global) minimizer of the (parametric) *lower level problem*

$$(2.1) \quad \underset{y}{\text{minimize}} G(\bar{x}, \cdot) \text{ subject to } y \in Y(\bar{x})$$

($\bar{x} \in \mathbb{R}^n$ is the parameter). Define

$$L_0(\bar{x}, \bar{y}) = \{l \in L \mid v_l(\bar{x}, \bar{y}) = 0\}$$

as the index set of active constraints at $\bar{y}$ with respect to (2.1). By the Fritz–John optimality condition [10], there exist—not necessarily unique—multipliers $\bar{\alpha}, \bar{\beta} = (\bar{\beta}_k, k \in K)$ and $\bar{\gamma} = (\bar{\gamma}_l, l \in L_0(\bar{x}, \bar{y}))$ satisfying

$$(2.2) \quad \begin{aligned} \bar{\alpha} &\geq 0, \quad \bar{\gamma}_l \geq 0, \quad l \in L_0(\bar{x}, \bar{y}), \\ \bar{\alpha} + \sum_{k \in K} |\bar{\beta}_k| + \sum_{l \in L_0(\bar{x}, \bar{y})} \bar{\gamma}_l &= 1, \\ D_y \mathfrak{L}^{(\bar{x}, \bar{y})}(\bar{x}, \bar{y}, \bar{\alpha}, \bar{\beta}, \bar{\gamma}) &= 0, \end{aligned}$$

where

$$\mathfrak{L}^{(\bar{x}, \bar{y})}(x, y, \alpha, \beta, \gamma) = \alpha G(x, y) - \sum_{k \in K} \beta_k u_k(x, y) - \sum_{l \in L_0(\bar{x}, \bar{y})} \gamma_l v_l(x, y)$$

is the corresponding Lagrangian. Hence, as we have seen in [11], the set

$$F(\bar{x}, \bar{y}) = \{(\bar{\alpha}, \bar{\beta}, \bar{\gamma}) \in \mathbb{R} \times \mathbb{R}^{|K|} \times \mathbb{R}^{|L_0(\bar{x}, \bar{y})|} \mid (\bar{\alpha}, \bar{\beta}, \bar{\gamma}) \text{ satisfies (2.2)}\}$$

is nonempty and compact for $\bar{x} \in M$ and $\bar{y} \in Y_0(\bar{x})$. Furthermore, we define for $\bar{x} \in M$

$$V(\bar{x}) = \bigcup_{\bar{y} \in Y_0(\bar{x})} \{D_x \mathfrak{L}^{(\bar{x}, \bar{y})}(\bar{x}, \bar{y}, \bar{\alpha}, \bar{\beta}, \bar{\gamma}) \mid (\bar{\alpha}, \bar{\beta}, \bar{\gamma}) \in F(\bar{x}, \bar{y})\}.$$

Cones. Let $S \subset \mathbb{R}^n$ and $s \in S$. Then $\text{cl}(S)$, $\text{co}(S)$, and S^0 denote the closure of S , the convex cone of S (consisting of all finite nonnegative combinations of elements of S with $\text{co}(\emptyset) = \{0\}$), and the dual cone $\{x \in \mathbb{R}^n \mid x^\top s \geq 0, s \in S\}$ of S , respectively.

The *inner tangent cone* $\Gamma(s, S)$ of S at $s \in S$ is defined as follows: $\bar{d} \in \Gamma(s, S)$ if and only if there exist a neighborhood D of $\bar{d}$ and some $\tau > 0$ such that

$$s + td \in S \quad \text{for all } t \in (0, \tau) \text{ and all } d \in D.$$

The *contingent cone* $\Gamma^*(s, S)$ of S at $s \in S$ is defined as follows: $\bar{d} \in \Gamma^*(s, S)$ if and only if there exist sequences $\{t^\nu\}_{\nu \in \mathbb{N}}$ and $\{d^\nu\}_{\nu \in \mathbb{N}}$ such that

$$t^\nu \downarrow 0, \quad d^\nu \rightarrow \bar{d}, \quad \text{and } s + t^\nu d^\nu \in S \quad \text{for all } \nu \in \mathbb{N}.$$

For $\bar{x} \in M$ we define

$$\begin{aligned} \mathcal{L}_0(\bar{x}, M) &= \{\xi \in \mathbb{R}^n \mid w^\top \xi > 0, w \in V(\bar{x})\}, \\ \mathcal{L}'(\bar{x}, M) &= \{\xi \in \mathbb{R}^n \mid w^\top \xi \geq 0, w \in V(\bar{x})\}, \\ \tilde{\mathcal{L}}(\bar{x}, M) &= \left\{ \xi \in \mathbb{R}^n \mid \inf_{y \in Y_0(\bar{x})} \max_{(\alpha, \beta, \gamma) \in F(\bar{x}, y)} D_x \mathfrak{L}^{(\bar{x}, y)}(\bar{x}, y, \alpha, \beta, \gamma)^\top \xi \geq 0 \right\}, \end{aligned}$$

and the cone of attainable directions (cf., e.g., [2])

$$A(\bar{x}, M) = \left\{ \xi \in \mathbb{R}^n \setminus \{0\} \mid \begin{array}{l} \text{There exist some } \tau > 0 \text{ and a continuously} \\ \text{differentiable arc } C : [0, \tau) \rightarrow \mathbb{R}^n \text{ such that} \\ C(0) = \bar{x}, \quad C'(0) = \xi, \text{ and } C(t) \in M, \quad t \in [0, \tau) \end{array} \right\}$$

(where $C'(t) = (DC_1(t), \dots, DC_n(t))$).

Constraint qualifications.

EACQ. The *extended Abadie constraint qualification (EACQ)* is said to hold at $\bar{x} \in M$ if $\Gamma^*(\bar{x}, M) = \tilde{\mathcal{L}}(\bar{x}, M)$. This extension of the well-known ACQ (cf. [1]) to GSIP was introduced in [22].

EMFCQ. The *extended Mangasarian–Fromovitz constraint qualification (EMFCQ)* is said to hold at $\bar{x} \in M$ if $\mathcal{L}_0(\bar{x}, M) \neq \emptyset$. This extension of the well-known MFCQ (cf. [15]) to GSIP was introduced in [11].

Now, we introduce extensions of the well-known KTCQ (cf. [13]).

EKTCQ1. The *extended Kuhn–Tucker constraint qualification 1 (EKTCQ1)* is said to hold at $\bar{x} \in M$ if $\mathcal{L}'(\bar{x}, M) \subset \text{cl}(A(\bar{x}, M))$.

EKTCQ2. The *extended Kuhn–Tucker constraint qualification 2 (EKTCQ2)* is said to hold at $\bar{x} \in M$ if $\tilde{\mathcal{L}}(\bar{x}, M) \subset \text{cl}(A(\bar{x}, M))$.

An objective of this paper is to ask for sufficiently weak extensions of KTCQ to GSIP and to analyze the extent to which a corresponding Karush–Kuhn–Tucker theorem depends on these extensions. Obviously, EKTCQ1 and EKTCQ2 are generalizations of KTCQ (we have $\mathcal{L}'(\bar{x}, M) = \tilde{\mathcal{L}}(\bar{x}, M)$ for finite programming problems). Since $\mathcal{L}'(\bar{x}, M) \subset \tilde{\mathcal{L}}(\bar{x}, M)$, EKTCQ1 is a weaker condition than EKTCQ2. Section 3 contains a Karush–Kuhn–Tucker theorem under EKTCQ1.

Remark 2.1. Note that the set $\mathcal{L}'(\bar{x}, M)$ can be much smaller than $\tilde{\mathcal{L}}(\bar{x}, M)$ and, in particular, need not contain the contingent cone $\Gamma^*(\bar{x}, M)$ (see the subsequent example, Example 2.2). Furthermore, since the set $\tilde{\mathcal{L}}(\bar{x}, M)$ represents the analogue of the linearization cone, which appears in KTCQ (an exhaustive discussion on this subject can be found in [24, section 4.2]), the concepts of the conditions EKTCQ2 and KTCQ have more in common than those of EKTCQ1 and KTCQ.

Example 2.2. In this example the set $\mathcal{L}'(\bar{x}, M)$ does not contain the contingent cone $\Gamma^*(\bar{x}, M)$. Let

$$M = \{x \in \mathbb{R}^2 \mid yx_2 \geq 0, y \in Y(x)\}$$

and

$$Y(x) = \{y \in \mathbb{R} \mid x_1 - y^2 \geq 0\}.$$

We obtain

$$Y(x) = \begin{cases} [-\sqrt{x_1}, \sqrt{x_1}] & \text{if } x_1 \geq 0, \\ \emptyset & \text{if } x_1 < 0, \end{cases}$$

and

$$M = \{x \in \mathbb{R}^2 \mid x_1 \leq 0\} \cup \{x \in \mathbb{R}^2 \mid x_1 > 0, x_2 = 0\}.$$

At $\bar{x} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ we get $Y_0(\bar{x}) = Y(\bar{x}) = \{0\}$, $V(\bar{x}) = \{ \begin{pmatrix} -\gamma \\ 0 \end{pmatrix} \mid \gamma \in [0, 1] \}$, as well as

$$\begin{aligned} \mathcal{L}'(\bar{x}, M) &= \{\xi \in \mathbb{R}^2 \mid \xi_1 \leq 0\}, \\ \tilde{\mathcal{L}}(\bar{x}, M) &= \mathbb{R}^2, \\ \Gamma^*(\bar{x}, M) &= M, \\ A(\bar{x}, M) &= M \setminus \{0\}. \end{aligned}$$

Hence, we have that $\Gamma^*(\bar{x}, M) \not\subset \mathcal{L}'(\bar{x}, M)$, $\mathcal{L}'(\bar{x}, M) \subset \text{cl}(A(\bar{x}, M))$ (EKTCQ1 holds at $\bar{x}$) and $\tilde{\mathcal{L}}(\bar{x}, M) \not\subset \text{cl}(A(\bar{x}, M))$ (EKTCQ2 does not hold at $\bar{x}$).

Lemmas. We present some auxiliary results which are used in the following section.

LEMMA 2.3 (see [11, Lemma 3.3]). *Let $\bar{x} \in M$. Then, the set $V(\bar{x})$ is compact.*

LEMMA 2.4. *The following assertions hold:*

- (i) $\Gamma^*(\bar{x}, M)$ is closed.
- (ii) $\mathcal{L}_0(\bar{x}, M) \subset A(\bar{x}, M)$.
- (iii) $A(\bar{x}, M) \subset \Gamma^*(\bar{x}, M) \subset \tilde{\mathcal{L}}(\bar{x}, M)$.

Proof. (i) is proved in [14].

(ii) By [22, Corollary 2.14], we have $\mathcal{L}_0(\bar{x}, M) \subset \Gamma(\bar{x}, M)$. Now, let $\xi \in \Gamma(\bar{x}, M) \setminus \{0\}$. Then, there exists some $\tau > 0$ such that $\bar{x} + t\xi \in M$, $t \in [0, \tau)$, and, therefore, $\xi \in A(\bar{x}, M)$.

(iii) First, we prove $A(\bar{x}, M) \subset \Gamma^*(\bar{x}, M)$ (this result is already mentioned without proof in [16] for finite programming problems). Let $\xi \in A(\bar{x}, M)$, and let $C : [0, \tau) \mapsto \mathbb{R}^n$ be a continuously differentiable arc with $C(0) = \bar{x}$, $C'(0) = \xi$, and $C(t) \in M$, $t \in [0, \tau)$. Then, a (componentwise) Taylor expansion provides $C(t) = \bar{x} + t\xi + R(t)$, $t \in [0, \tau_1)$, for some $\tau_1 \in (0, \tau)$, where $R(t) \in \mathbb{R}^n$ and $\lim_{t \downarrow 0} \frac{R(t)}{t} = 0$. Then, for $t^\nu \in (0, \tau_1)$, $t^\nu \downarrow 0$, and $\xi^\nu = \xi + \frac{R(t^\nu)}{t^\nu}$ we obtain $C(t^\nu) = \bar{x} + t^\nu \xi^\nu$, and, therefore, $\xi \in \Gamma^*(\bar{x}, M)$. The inclusion $\Gamma^*(\bar{x}, M) \subset \tilde{\mathcal{L}}(\bar{x}, M)$ is proved in [24, section 4.2.3]. $\square$

By Lemma 2.4 (i),(iii) we obtain the following.

COROLLARY 2.5. *Let $\bar{x} \in M$ and assume that EKTCQ2 holds at $\bar{x}$. Then EACQ holds at $\bar{x}$ as well.*

LEMMA 2.6 (see [2, section 2.5]). *Let $S_1, S_2 \subset \mathbb{R}^n$. Then we have the following:*

- (i) *If $S_1 \subset S_2$, then $S_2^0 \subset S_1^0$.*
- (ii) $S_1^0 = \text{co}(S_1)^0$.
- (iii) *If S_1 is a closed convex cone, then $S_1 = S_1^{00}$.*

LEMMA 2.7 (see [8, Lemma 3.1.12]). *Let $S \subset \mathbb{R}^n$ be a compact set.*

- (i) *If there exists a solution $\xi \in \mathbb{R}^n$ of the system*

$$\xi^\top s > 0, \quad s \in S,$$

then $\text{co}(S)$ is closed.

- (ii) *If S is a finite set, then $\text{co}(S)$ is closed.*

3. Results and examples.

First order optimality condition of GSIP (FOC–GSIP), a corresponding Karush–Kuhn–Tucker theorem, and the interrelation between EMFCQ and EKTCQ1. We say that the FOC–GSIP is fulfilled at $\bar{x} \in M$ if

$$(3.1) \quad Df(\bar{x}) \in \text{co}(V(\bar{x})).$$

If $\bar{x} \in M$ and $Y_0(\bar{x}) \neq \emptyset$, then (3.1) is equivalent to the existence of finitely many

$$y^j \in Y_0(\bar{x}), \quad (\alpha^j, \beta^j, \gamma^j) \in F(\bar{x}, y^j), \quad \mu_j \geq 0, \quad j = 1, \dots, p,$$

satisfying

$$Df(\bar{x}) - \sum_{j=1}^p \mu_j D_x \mathfrak{L}^{(\bar{x}, y^j)}(\bar{x}, y^j, \alpha^j, \beta^j, \gamma^j) = 0.$$

It is shown in [11, Corollary 4.1] that if $\bar{x}$ is a local minimizer of GSIP, and if EMFCQ holds at $\bar{x}$, then FOC–GSIP is fulfilled at $\bar{x} \in M$. Before we state a

corresponding Karush–Kuhn–Tucker theorem under EKTCQ1, we make the following observation. By Lemma 2.6, if $\text{co}(V(\bar{x}))$ is closed, then (3.1) is equivalent to

$$(3.2) \quad V(\bar{x})^0 (= \mathcal{L}'(\bar{x}, M)) \subset \{Df(\bar{x})\}^0.$$

THEOREM 3.1. *Let $\bar{x}$ be a local minimizer of GSIP. Assume that EKTCQ1 (or EKTCQ2) holds at $\bar{x}$ and that $\text{co}(V(\bar{x}))$ is closed. Then, FOC–GSIP is fulfilled at $\bar{x} \in M$.*

Proof. Since $\bar{x}$ is a local minimizer of GSIP, we have

$$\Gamma^*(\bar{x}, M) \subset \{Df(\bar{x})\}^0$$

(cf. [2, Theorem 5.1.2]). We obtain (3.2) from the following chain of inclusions:

$$\begin{array}{ccccccc} \mathcal{L}'(\bar{x}, M) & \subset & (\tilde{\mathcal{L}}(\bar{x}, M) \subset & \text{cl}(A(\bar{x}, M)) & \subset & \Gamma^*(\bar{x}, M) & \subset \{Df(\bar{x})\}^0. \\ \uparrow & & \uparrow & & \uparrow & & \uparrow \\ (\text{EKTCQ1}) & & (\text{EKTCQ2}) & & \text{Lemma} & & \bar{x} \text{ is a local} \\ & & & & 2.4 \text{ (i),(iii)} & & \text{minimizer.} \quad \square \end{array}$$

The next proposition and Example 3.3 show that EKTCQ1 is a weaker condition than EMFCQ.

PROPOSITION 3.2. *Assume that EMFCQ holds at $\bar{x} \in M$. Then, EKTCQ1 holds at $\bar{x}$ as well.*

Proof. By EMFCQ, we obtain that $\mathcal{L}_0(\bar{x}, M) \neq \emptyset$, $\text{cl}(\mathcal{L}_0(\bar{x}, M)) \neq \emptyset$, and, by Lemma 2.4(ii), $\text{cl}(\mathcal{L}_0(\bar{x}, M)) \subset \text{cl}(A(\bar{x}, M))$. Furthermore, $\mathcal{L}_0(\bar{x}, M) \neq \emptyset$ also implies that $\text{cl}(\mathcal{L}_0(\bar{x}, M)) = \mathcal{L}'(\bar{x}, M)$. Therefore we get $\mathcal{L}'(\bar{x}, M) \subset \text{cl}(A(\bar{x}, M))$. $\square$

The following example illustrates Theorem 3.1.

Example 3.3. Consider GSIP with $n = 2$, $r = 1$, $K = \emptyset$, $L = \{1, 2\}$, and

$$\begin{aligned} f(x) &= -x_1, \\ G(x, y) &= y^5 - 3x_2^2, \\ v_1(x, y) &= -y, \\ v_2(x, y) &= y^5 + 4x_1^2 + x_2^2 - 1. \end{aligned}$$

A short calculation yields

$$Y(x) = \begin{cases} [(-4x_1^2 - x_2^2 + 1)^{1/5}, 0] & \text{if } -4x_1^2 - x_2^2 + 1 \leq 0, \\ \emptyset & \text{if } -4x_1^2 - x_2^2 + 1 > 0, \end{cases}$$

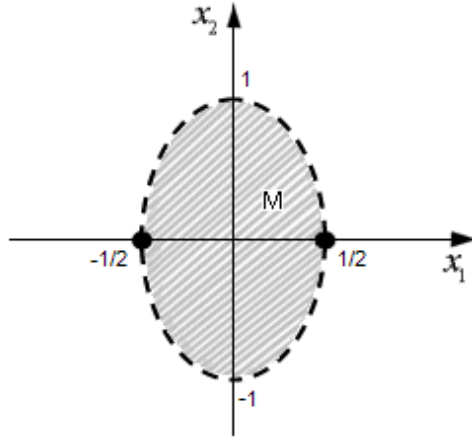
and the feasible set (cf. Figure 3.1)

$$M = \{x \in \mathbb{R}^2 \mid 4x_1^2 + x_2^2 < 1\} \cup \left\{ \begin{pmatrix} 1/2 \\ 0 \end{pmatrix}, \begin{pmatrix} -1/2 \\ 0 \end{pmatrix} \right\}.$$

Then, the point $\bar{x} = \begin{pmatrix} 1/2 \\ 0 \end{pmatrix}$ is a (global) minimizer with $Y_0(\bar{x}) = \{\bar{y}\} = \{0\}$, $\mathcal{L}(\bar{x}, \bar{y})(x, y, \alpha, \gamma_1, \gamma_2) = \alpha(y^5 - 3x_2^2) - \gamma_1(-y) - \gamma_2(y^5 + 4x_1^2 + x_2^2 - 1)$, $F(\bar{x}, \bar{y}) = \{(\alpha, 0, 1 - \alpha) \mid \alpha \in [0, 1]\}$, $V(\bar{x}) = \{ \begin{pmatrix} -4(1-\alpha) \\ 0 \end{pmatrix} \mid \alpha \in [0, 1] \}$ and we get the following:

- Since $0 \in V(\bar{x})$ (for $\alpha = 1$), EMFCQ does not hold at $\bar{x}$.
- $\text{co}(V(\bar{x})) = \{ \begin{pmatrix} \delta \\ 0 \end{pmatrix} \mid \delta \leq 0 \}$ and, therefore, $\text{co}(V(\bar{x}))$ is closed.
- It is easily seen that $\mathcal{L}'(\bar{x}, M) = \{ \xi \in \mathbb{R}^2 \mid \xi_1 \leq 0 \}$, $\{ \xi \in \mathbb{R}^2 \mid \xi_1 < 0 \} \subset A(\bar{x}, M)$, and, hence,

$$\{ \xi \in \mathbb{R}^2 \mid \xi_1 \leq 0 \} \subset \text{cl}(A(\bar{x}, M)).$$

FIG. 3.1. The feasible set M .

The latter fact means that EKTCQ1 holds at $\bar{x}$. Therefore, the following assumptions of Theorem 3.1 are fulfilled:

- $\bar{x}$ is a local minimizer.
- EKTCQ1 holds at $\bar{x}$.
- $\text{co}(V(\bar{x}))$ is closed.

FOC-GSIP is fulfilled at $\bar{x}$ and we obtain

$$(3.3) \quad Df(\bar{x}) - \mu D_x \mathfrak{L}^{(\bar{x}, \bar{y})}(\bar{x}, \bar{y}, \alpha, \gamma_1, \gamma_2) = \begin{pmatrix} -1 \\ 0 \end{pmatrix} - \mu \begin{pmatrix} -4(1-\alpha) \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

with $\mu = \frac{1}{4(1-\alpha)}$ for $\alpha \in [0, 1)$. The present example illustrates the following properties:

- The violation of EMFCQ does not force the violation of EKTCQ1; i.e., by Proposition 3.2, EKTCQ1 is a weaker condition than EMFCQ.
- The cone $\text{co}(V(\bar{x}))$ is closed although EMFCQ does not hold at $\bar{x}$.
- For the contingent cone we obtain $\Gamma^*(\bar{x}, M) = \{\xi \in \mathbb{R}^2 \mid \xi_1 \leq 0\}$, which is different from

$$\tilde{\mathcal{L}}(\bar{x}, M) = \left\{ \xi \in \mathbb{R}^2 \mid \max_{\alpha \in [0, 1]} (-4(1-\alpha), 0) \xi \geq 0 \right\} = \mathbb{R}^2.$$

Consequently, EACQ and, by Corollary 2.5, EKTCQ2, do not hold at $\bar{x}$; in general, EKTCQ1 implies neither EACQ nor EKTCQ2.

• Note that in (3.3) we cannot find a corresponding μ if we choose $\alpha = 1$; i.e., in general, there does not exist a corresponding linear combination satisfying FOC-GSIP for each possible choice of elements of $F(\bar{x}, \bar{y})$ (cf. [19] for a further discussion).

• A main difference between GSIP and a standard (semi-infinite or finite) programming problem is that the feasible set of GSIP need not be closed! This property may cause serious difficulties in the design of solution algorithms (cf. [11, 24] for more details on the topological structure of the feasible set).

The assumption of closedness of $\text{co}(V(\bar{x}))$. As already mentioned above, the optimality condition FOC-GSIP is fulfilled at a local minimizer $\bar{x}$ of GSIP if EMFCQ holds at $\bar{x}$. However, if we assume only that one of the conditions EKTCQ1 or

EKTCQ2 holds at $\bar{x}$, then we have to assume additionally that $\text{co}(V(\bar{x}))$ is closed in order to obtain FOC–GSIP (cf. Theorem 3.1). The following example shows that we cannot omit the assumption of closedness of $\text{co}(V(\bar{x}))$.

Example 3.4. Let $n = 2$, $r = 2$, $K = \emptyset$, $L = \{1, 2, 3, 4, 5\}$, and consider a corresponding GSIP with $f(x) = -x_1$, $G(x, y) = -y_2 + x_2 y_1^4$, $v_1(x, y) = -y_2 + x_1 y_1^2$, $v_2(x, y) = y_1 + 1$, $v_3(x, y) = -y_1 + 1$, $v_4(x, y) = y_2 + 1$, $v_5(x, y) = -y_2 + 1$.

Then we have the feasible set (cf. Figure 3.2)

$$M = \{x \in \mathbb{R}^2 \mid -1 \leq x_1 \leq 0, x_2 \geq x_1\} \cup \{x \in \mathbb{R}^2 \mid x_1 < -1, x_2 \geq -x_1^2\}$$

and $\bar{x} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ is a local minimizer with

$$Y_0(\bar{x}) = \{(y_1, 0) \mid y_1 \in [-1, 1]\}, \quad F(\bar{x}, \bar{y}) = \left\{ \left(\frac{1}{2}, \frac{1}{2} \right)^\top \right\} \quad \text{for } \bar{y} \in Y_0(\bar{x}),$$

and

$$V(\bar{x}) = \left\{ \frac{1}{2} \begin{pmatrix} -y_1^2 \\ y_1^4 \end{pmatrix}, y_1 \in [-1, 1] \right\}.$$

Obviously, EMFCQ does not hold at $\bar{x}$ (with $y_1 = 0$ we have $0 \in V(\bar{x})$). With respect to Theorem 3.1 we observe the following situation:

- $\bar{x}$ is a local minimizer.
- EKTCQ1 and EKTCQ2 hold at $\bar{x}$ (it is $\mathcal{L}'(\bar{x}, M) = \tilde{\mathcal{L}}(\bar{x}, M) = M$).
- $\text{co}(V(\bar{x}))$ is not closed.
- FOC–GSIP is not fulfilled at $\bar{x}$; since any nonnegative linear combination of elements of $V(\bar{x})$ results in either $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ or $\begin{pmatrix} a \\ b \end{pmatrix}$ with $a < 0$ and $b > 0$, the gradient $Df(\bar{x}) = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$ cannot be obtained.

The latter example illustrates that EKTCQ1 and EKTCQ2 do not imply the closedness of $\text{co}(V(\bar{x}))$. However, the following proposition shows that EMFCQ always implies the closedness of $\text{co}(V(\bar{x}))$.

PROPOSITION 3.5. *Let EMFCQ be satisfied at $\bar{x} \in M$. Then, $\text{co}(V(\bar{x}))$ is closed.*

Proof. By Lemma 2.3, $V(\bar{x})$ is a compact set, and EMFCQ implies $\mathcal{L}_0(\bar{x}, M) \neq \emptyset$. Then, the result is an immediate consequence of Lemma 2.7(i). $\square$

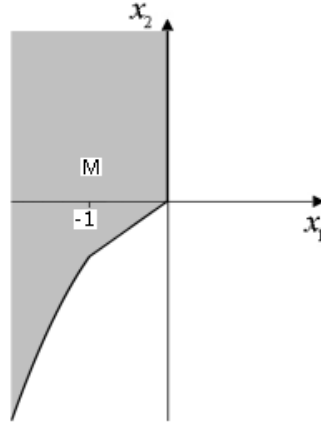
Obviously, the price one has to pay for using the weaker constraint qualification EKTCQ1 (instead of EMFCQ) is the additional assumption of the closedness of $\text{co}(V(\bar{x}))$. The following proposition shows that in the important particular case of finitely many active constraints (i.e., $Y_0(\bar{x})$ is a finite set) this additional assumption is always fulfilled.

PROPOSITION 3.6. *Let $\bar{x} \in M$ and let $Y_0(\bar{x})$ be a finite set. Then $\text{co}(V(\bar{x}))$ is closed.*

Proof. For the sake of simplicity the proof is given for the case $Y_0(\bar{x}) = \{\bar{y}\}$ (it can be generalized straightforwardly to the general case $Y_0(\bar{x}) = \{y^1, \dots, y^p\}$). After substituting the equalities $u_k(x, y) = 0$, $k \in K$, by $2|K|$ inequalities $u_k(x, y) \geq 0$, $-u_k(x, y) \geq 0$, $k \in K$, suppose without loss of generality that $K = \emptyset$.

Now, letting $\xi^\nu \in \text{co}(V(\bar{x}))$, $\xi^\nu \rightarrow \bar{\xi}$, we will show that $\bar{\xi} \in \text{co}(V(\bar{x}))$. If $\bar{\xi} = 0$, then we are done; assume that $\bar{\xi} \neq 0$. The definition of $V(\bar{x})$ yields

$$\begin{pmatrix} \xi^\nu \\ 0 \end{pmatrix} \in \text{co}(W), \quad \nu \in \mathbb{N},$$

FIG. 3.2. The feasible set M .

where

$$W = \left\{ \begin{pmatrix} D_x G(\bar{x}, \bar{y}) \\ D_y G(\bar{x}, \bar{y}) \end{pmatrix}, - \begin{pmatrix} D_x v_l(\bar{x}, \bar{y}) \\ D_y v_l(\bar{x}, \bar{y}) \end{pmatrix}, l \in L_0(\bar{x}, \bar{y}) \right\}.$$

By Lemma 2.7(ii), $\text{co}(W)$ is closed and, therefore,

$$\begin{pmatrix} \bar{\xi} \\ 0 \end{pmatrix} \in \text{co}(W) \quad \text{with } 0 \in \mathbb{R}^r,$$

$$\begin{pmatrix} \bar{\xi} \\ 0 \end{pmatrix} = \alpha \begin{pmatrix} D_x G(\bar{x}, \bar{y}) \\ D_y G(\bar{x}, \bar{y}) \end{pmatrix} - \sum_{l \in L_0(\bar{x}, \bar{y})} \gamma_l \begin{pmatrix} D_x v_l(\bar{x}, \bar{y}) \\ D_y v_l(\bar{x}, \bar{y}) \end{pmatrix}$$

for some $\alpha \geq 0$, $\gamma_l \geq 0$, $l \in L_0(\bar{x}, \bar{y})$.

Since $\bar{\xi} \neq 0$, the latter equality implies $\bar{\xi} \in \text{co}(V(\bar{x}))$. $\square$

The final example illustrates more interrelations between EMFCQ, EACQ, EKTCQ1, and EKTCQ2.

Example 3.7. Let $n = 2$, $r = 1$, $K = \emptyset$, and $L = \{1, 2\}$. Consider the feasible set

$$M = \{x \in \mathbb{R}^2 \mid y(x_2 - g(x_1)) + (y - 1)x_2 \geq 0, y \in Y(x)\},$$

where

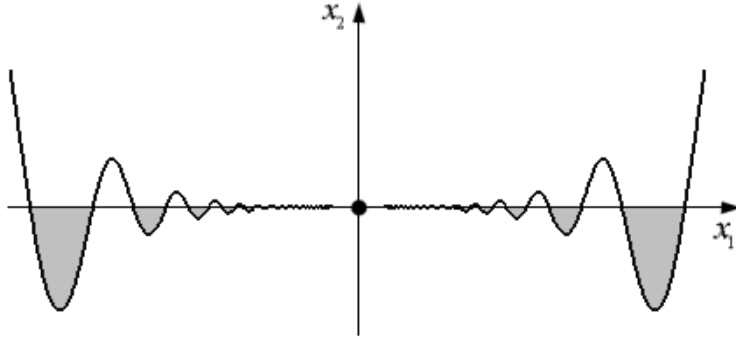
$$g(x_1) = \begin{cases} x_1^4 \cos \frac{1}{x_1} & \text{if } x_1 \neq 0, \\ 0 & \text{if } x_1 = 0, \end{cases}$$

and

$$Y(x) = \{y \in \mathbb{R} \mid 1 - y \geq 0, y \geq 0\}.$$

Obviously, $g \in C^1(\mathbb{R}, \mathbb{R})$, $Y(x) = Y$ does not depend on $x \in \mathbb{R}^2$ and we have the feasible set (cf. Figure 3.3)

$$M = \left\{ x \in \mathbb{R}^2 \mid x_1 \neq 0, x_2 \leq 0, x_2 \geq x_1^4 \cos \frac{1}{x_1} \right\} \cup \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right\}.$$

FIG. 3.3. The feasible set M .

For $\bar{x} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ we obtain

$$\begin{aligned} Y_0(\bar{x}) &= [0, 1], \\ V(\bar{x}) &= \left\{ \begin{pmatrix} 0 \\ 2y - 1 \end{pmatrix} \mid y \in [0, 1] \right\}, \\ \mathcal{L}'(\bar{x}, M) &= \tilde{\mathcal{L}}(\bar{x}, M) = \Gamma^*(\bar{x}, M) = \{\xi \in \mathbb{R}^2 \mid \xi_2 = 0\}, \end{aligned}$$

where the latter fact implies that EACQ holds at $\bar{x}$. Since $0 \in V(\bar{x})$ (with $y = \frac{1}{2}$), EMFCQ does not hold at $\bar{x}$. However, M consists of infinitely many connected components (locally in a neighborhood of $\bar{x}$) and we obtain $A(\bar{x}, M) = \emptyset$.

Hence, $\mathcal{L}'(\bar{x}, M)$ is not a subset of $\text{cl}(A(\bar{x}, M))$ and, therefore, EKTCQ1 and EKTCQ2 do not hold at $\bar{x}$. The current example shows that, in general,

- EACQ implies neither EMFCQ, EKTCQ2, nor EKTCQ1.

This property is not surprising since it generalizes an analogous property from finite programming (which is illustrated by a similar example in [1]).

4. Conclusions and summary.

1. In this paper we introduced two extensions, EKTCQ1 and EKTCQ2, of KTCQ to GSIP and proved a corresponding Karush–Kuhn–Tucker theorem. For finite programming problems the following chain of implications holds for the constraint qualifications MFCQ, KTCQ, and ACQ at a point $\bar{x} \in M$ (cf., e.g., [16]):

$$(\text{MFCQ}) \longrightarrow (\text{KTCQ}) \longrightarrow (\text{ACQ}).$$

For the class GSIP there does not exist a straightforward generalization of this chain for the extended constraint qualifications EMFCQ, EKTCQ1, EKTCQ2, and EACQ, as shown in the following:

- According to [22], EMFCQ does not imply EACQ in general. Then, by Corollary 2.5, EMFCQ does not imply EKTCQ2 in general as well:

$$(4.1) \quad \left. \begin{array}{l} (\text{EMFCQ}) \not\rightarrow (\text{EACQ}) \\ (\text{EMFCQ}) \not\rightarrow (\text{EKTCQ2}) \end{array} \right\}.$$

- In the present paper, we proved the following:

$$(4.2) \quad \left. \begin{array}{l} (\text{EKTCQ2}) \longrightarrow (\text{EACQ}) \\ \downarrow \\ (\text{EKTCQ1}) \longleftarrow (\text{EMFCQ}) \end{array} \right\}.$$

The examples in section 3 and (4.1) show that no implications (other than those shown in (4.2)) between two constraint qualifications from EMFCQ, EKTCQ1, EKTCQ2, and EACQ generally hold.

2. We summarize some of the properties used above in the following chain of inclusions. Let $\bar{x} \in M$:

$$\begin{aligned} \mathcal{L}'(\bar{x}, M) & \cap \\ \mathcal{L}_0(\bar{x}, M) \subset A(\bar{x}, M) \subset \text{cl}(A(\bar{x}, M)) \subset \Gamma^*(\bar{x}, M) \subset \tilde{\mathcal{L}}(\bar{x}, M). \\ \left(\begin{array}{l} \text{If } \bar{x} \text{ is a local minimizer} \end{array} \longrightarrow \bigcap \{Df(\bar{x})\}^0 \right). \end{aligned}$$

Then, the following results can easily be verified:

- The implications in (4.2).
- If $\bar{x}$ is a local minimizer of GSIP and $\text{co}(V(\bar{x}))$ is closed, then

$$\begin{array}{lcl} \text{(EMFCQ)} \text{ (that implies } \text{cl}(\mathcal{L}_0(\bar{x}, M)) = \mathcal{L}'(\bar{x}, M) & \longrightarrow & \\ & \text{(EACQ)} & \longrightarrow \\ & \text{(EKTCQ1)} & \longrightarrow \\ & \text{(EKTCQ2)} & \longrightarrow \end{array} \left. \vphantom{\begin{array}{l} \text{(EMFCQ)} \\ \text{(EACQ)} \\ \text{(EKTCQ1)} \\ \text{(EKTCQ2)} \end{array}} \right\} \begin{array}{l} \text{(FOC-GSIP)} \\ \text{(see (3.2)).} \end{array}$$

• EMFCQ implies that $\mathcal{L}'(\bar{x}, M) \subset \Gamma^*(\bar{x}, M)$. The latter inclusion is another constraint qualification that for finite programming problems also coincides with ACQ.

Acknowledgments. The authors would like to thank the referees for their careful reading of the manuscript and constructive comments. The improved version of the paper has benefited from their valuable suggestions and detailed remarks.

REFERENCES

- [1] J. M. ABADIE, *On the Kuhn-Tucker theorem*, in Nonlinear Programming, J. Abadie, ed., Wiley, New York, 1967, pp. 21–36.
- [2] M. S. BAZARAA, H. D. SHERALI, AND C. M. SHETTY, *Nonlinear Programming. Theory and Algorithms*, Wiley, New York, 1993.
- [3] C. BERGE, *Topological Spaces*, Oliver and Boyd, London, 1963.
- [4] A. V. Fiacco AND G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, Wiley, New York, 1968.
- [5] M. A. GOBERNA AND M. A. LÓPEZ, *Linear Semi-Infinite Optimization*, Wiley, New York, 1998.
- [6] M. A. GOBERNA AND M. A. LÓPEZ, EDS., *Semi-infinite Programming—Recent Advances*, Kluwer, Boston, 2001.
- [7] R. HETTICH AND K. O. KORTANEK, *Semi-infinite programming: Theory, methods, and applications*, SIAM Rev., 35 (1993), pp. 380–429.
- [8] R. HETTICH AND P. ZENCKE, *Numerische Methoden der Approximation und semi-infiniten Optimierung*, Teubner Studienbücher, Stuttgart, 1982.
- [9] W. W. HOGAN, *Point-to-set mappings in mathematical programming*, SIAM Rev., 15 (1973), pp. 591–603.
- [10] F. JOHN, *Extremum problems with inequalities as subsidiary conditions*, in Studies and Essays Presented to R. Courant on His 60th Birthday, K. O. Friedrichs, O. E. Neugebauer, and J. J. Stoker, eds., Interscience Publishers, New York, 1948, pp. 187–204.
- [11] H. TH. JONGEN, J.-J. RÜCKMANN, AND O. STEIN, *Generalized semi-infinite optimization: A first order optimality condition and examples*, Math. Program., 83 (1998), pp. 145–158.
- [12] H. TH. JONGEN, F. TWILT, AND G.-W. WEBER, *Semi-infinite optimization: Structure and stability of the feasible set*, J. Optim. Theory Appl., 72 (1992), pp. 529–552.
- [13] H. W. KUHN AND A. W. TUCKER, *Nonlinear programming*, in Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, J. Neyman, ed., University of California Press, Berkeley, 1951, pp. 481–492.
- [14] P.-J. LAURENT, *Approximation et Optimisation*, Hermann, Paris, 1972.

- [15] O. L. MANGASARIAN AND S. FROMOVITZ, *The Fritz John necessary optimality conditions in the presence of equality and inequality constraints*, J. Math. Anal. Appl., 17 (1967), pp. 37–47.
- [16] D. W. PETERSON, *A review of constraint qualifications in finite-dimensional spaces*, SIAM Rev., 15 (1973), pp. 639–654.
- [17] E. POLAK, *Optimization. Algorithms and Consistent Approximations*, Springer, New York, 1997.
- [18] R. REEMTSSEN AND J.-J. RÜCKMANN, EDS., *Semi-Infinite Programming*, Kluwer, Boston, 1998.
- [19] J.-J. RÜCKMANN AND A. SHAPIRO, *First-order optimality conditions in generalized semi-infinite programming*, J. Optim. Theory Appl., 101 (1999), pp. 677–691.
- [20] J.-J. RÜCKMANN AND O. STEIN, *On convex lower level problems in generalized semi-infinite optimization*, in *Semi-infinite Programming—Recent Advances*, M. A. Goberna and M. A. López, eds., Kluwer, Boston, 2001, pp. 121–134.
- [21] A. SHAPIRO, *First and second order optimality conditions and perturbation analysis of semi-infinite programming problems*, in *Semi-infinite Programming*, R. Reemsten and J.-J. Rückmann, eds., Kluwer, Boston, 1998, pp. 103–134.
- [22] O. STEIN, *First order optimality conditions for degenerate index sets in generalized semi-infinite programming*, Math. Oper. Res., 26 (2001), pp. 565–582.
- [23] O. STEIN, *On constraint qualifications in nonsmooth optimization*, J. Optim. Theory Appl., 121 (2004), pp. 647–671.
- [24] O. STEIN, *Bi-level Strategies in Semi-infinite Programming*, Kluwer, Boston, 2003.
- [25] O. STEIN AND G. STILL, *On optimality conditions for generalized semi-infinite programming problems*, J. Optim. Theory Appl., 104 (2000), pp. 443–458.
- [26] R. A. TAPIA AND M. W. TROSSET, *An extension of the Karush-Kuhn-Tucker necessity conditions to infinite programming*, SIAM Rev., 36 (1994), pp. 1–17.

PROBABILISTICALLY CONSTRAINED LINEAR PROGRAMS AND RISK-ADJUSTED CONTROLLER DESIGN*

CONSTANTINO M. LAGOA[†], XIANG LI[†], AND MARIO SZNAIER[†]

Abstract. The focal point of this paper is the probabilistically constrained linear program (PCLP) and how it can be applied to control system design under risk constraints. The PCLP is the counterpart of the classical linear program, where it is assumed that there is random uncertainty in the constraints and, therefore, the deterministic constraints are replaced by probabilistic ones. It is shown that for a wide class of probability density functions, called log-concave symmetric densities, the PCLP is a convex program. An equivalent formulation of the PCLP is also presented which provides insight into numerical implementation. This concept is applied to control system design. It is shown how the results in this paper can be applied to the design of controllers for discrete-time systems to obtain a closed loop system with a well-defined risk of violating the so-called property of *superstability*. Furthermore, we address the problem of risk-adjusted pole placement.

Key words. probabilistic constraints, linear program, convexity, risk-adjusted controller

AMS subject classifications. 90C15, 90C25, 90C05, 93C99

DOI. 10.1137/S1052623403430099

1. Introduction. Recently there has been a growing interest in the development of control systems design procedures which are able to handle risk constraints. The main motivation for considering this problem is that “classical” robust controllers tend to be complex; i.e., often “classical” controller design algorithms produce high order controllers. The hope is that, if one is willing to tolerate a small, well-defined risk of violation of performance specifications, one would obtain a significantly less complex controller. Also, there are several problems which naturally lead to a formulation involving risk constraints. An example is emergency system operation. Here one would like to allow for a small, well-defined risk of violation of stability in exchange for a better chance of handling the emergency situation. For example, in the control of an aircraft, one may want to have more power available to avoid a collision, although this might lead to a small risk of engine failure. For situations like this, a different kind of controller should be designed. It should be a controller that maximizes performance at the expense of a very small, well-defined risk of instability. In other words, we relax the conditions on stability in exchange for an improvement in performance.

The problem of designing risk-adjusted controllers has been considered in several papers, e.g., see [1], [2], [3], and [4]. However, there is a fundamental difference between the work mentioned above and the results presented in this paper. In contrast to the work presented in the papers mentioned above, where the search for the controller parameters is done using randomized algorithms, this paper integrates a new line of research which aims at developing fast deterministic algorithms for risk-adjusted controller design. The work presented in [5] and [6] indicates that there are several risk-adjusted controller design problems which are convex and, hence, numer-

*Received by the editors June 19, 2003; accepted for publication (in revised form) July 19, 2004; published electronically April 22, 2005. Funding for this research was provided by the National Science Foundation under grants ECS-9984260 and ECS-0115946 and the Air Force Office of Scientific Research under grant AFSOR-F49620-00-1-0020.

<http://www.siam.org/journals/siopt/15-3/43009.html>

[†]Electrical Engineering Department, The Pennsylvania State University, University Park, PA 16802 (lagoa@engr.psu.edu, xxl124@psu.edu, msznaier@frodo.ee.psu.edu).

ically solvable. In this paper, we extend the class of risk-adjusted design problems which are known to be convex. The main paradigm underlying the results presented is the concept of the probabilistically constrained linear program.

1.1. Probabilistically constrained linear program. The main result of this paper concerns the convexity of the so-called probabilistically constrained linear program (PCLP). We show that, for a large class of probability distributions, the probabilistic version of the classical linear program is convex. Furthermore, we show how it can be applied in a controller design context.

Indeed, consider the “classical” linear program described by

$$\min c^T x$$

subject to

$$x^T a^i \leq b_i; \quad i = 1, 2, \dots, k,$$

where $c, x, a^i \in \mathbf{R}^\ell$, and $b_i \in \mathbf{R}$, $i = 1, 2, \dots, k$.

In the PCLP framework, the constraint vectors a^i and b above are treated as random and the deterministic constraints are replaced by probabilistic constraints. There are a number of versions of the PCLP problem and the one that is used throughout this paper is described below.

DEFINITION 1.1. *Given acceptable risk levels $0 \leq \varepsilon_i \leq 1$, $i = 1, 2, \dots, k$, the PCLP is defined as the following optimization problem:*

$$\min c^T x$$

subject to

$$\text{Prob}\{x^T a^i \leq b_i\} \geq 1 - \varepsilon_i; \quad i = 1, 2, \dots, k,$$

where $c, x \in \mathbf{R}^\ell$, and a^i, b are random vectors of appropriate dimensions.

1.2. Convexity of the feasible set. A fundamental question about the PCLP is the following: Is the PCLP a convex program? In other words, is the feasible set

$$\mathcal{X}_\varepsilon \doteq \{x \in \mathbf{R}^\ell : \text{Prob}\{x^T a^i \leq b_i\} \geq 1 - \varepsilon_i \text{ for } i = 1, 2, \dots, k\}$$

convex? It turns out that, without additional conditions on the distribution of the pair (a^i, b_i) , one can easily generate examples in which the answer to this question is “no.”

In this paper, we prove that the PCLP is a convex program when the probability density function of the random parameters is log-concave and symmetric; see section 2 for a precise definition of this class of distributions.

1.3. Why consider individual probability constraints? In the setup above, one considers only individual probability constraints; i.e., one looks at the probability of satisfaction of each of the constraints independently. In this paper, only this case is considered since joint probability constraints result, in general, in a nonconvex feasible set. As an example, consider the feasible set

$$\mathcal{X}_{0.45} \doteq \{x \in \mathbf{R}^2 : \text{Prob}\{x^T [\Delta a_1 \ \Delta a_2] \leq 0.15 \text{ and } x^T [0 \ \Delta a_2] \leq 0.15\} \geq 1 - 0.45\},$$

where $\Delta a = [\Delta a_1 \ \Delta a_2]^T$ is uniformly distributed over the set

$$\{\Delta a \in \mathbf{R}^2 : |\Delta a_1| \leq 1, |\Delta a_2| \leq 1\}.$$

One can see that both $x^1 = [1 \ 0]^T$ and $x^2 = [0 \ 1]^T$ belong to the set $\mathcal{X}_{0.45}$. Now consider

$$x^3 = \frac{1}{2}x^1 + \frac{1}{2}x^2 = [0.5 \ 0.5]^T.$$

It can be proven that

$$\text{Prob}\{(x^3)^T[\Delta a_1 \ \Delta a_2]^T \leq 0.15 \text{ and } (x^3)^T[0 \ \Delta a_2] \leq 0.15\} = 0.525,$$

and hence x^3 does not belong to $\mathcal{X}_{0.45}$. In other words, the set $\mathcal{X}_{0.45}$ is not convex.

1.4. Previous results. Convexity results are available for specific kinds of distributions. For example, in [15] it is proven that $\mathcal{X}_\varepsilon$ is convex when $0 \leq \varepsilon_i \leq 1/2$ and b_i and the components of the a^i are independent and normally distributed. This result was later extended to the case when a^i and b_i have stable distributions, e.g., see [16]. The convexity of the PCLP has also been established for the case when a^i are deterministic and b_i are random with a log-concave distribution; see [17]. Using recent results from geometry, in this paper we extend these earlier results and prove that, for a large class of joint distributions for a^i and b_i (which includes many of the commonly used distributions), the PCLP is a convex program.

Several results are also available for PCLPs with joint probability constraints. It has been shown that there are two instances in which the jointly constrained PCLP is convex: (i) when the quantities a^i are deterministic and the vector b has a log-concave distribution (not necessarily symmetric) or a so-called α -concave distribution; and (ii) when a^i and the vector b have a joint Gaussian distribution with some restrictions on the covariance matrix. Also, there are several results involving quasi-concave constraint functions instead of the linear functions considered in this paper. For an overview of such results see [17].

1.5. The sequel. Section 2 is dedicated to the definition of the class of admissible distributions for the uncertain parameters: log-concave symmetric distributions. The main result of this paper is presented in section 3, which states that the PCLP is a convex program. In section 4, we provide some insight into a numerical implementation of the PCLP. Section 5 is dedicated to the application of our results to control system design. Finally, in section 6, some concluding remarks are presented and several directions for further research are outlined.

2. Preliminaries: Log-concavity. In order to communicate the main result, we need to elaborate on which probability density functions are admissible for the vector of uncertain parameters. To this end, we require a definition of log-concave functions, e.g., see [17]. In this section, we also introduce a concept that is central to the results presented in this paper: floating body of a probability measure.

2.1. Log-concave functions and probability densities. A function $f : \mathbf{R}^\ell \rightarrow [0, \infty)$ is said to be *log-concave* if the following condition holds: Given any $x^0, x^1 \in \mathbf{R}^\ell$, and $\lambda \in [0, 1]$, it follows that

$$f((1 - \lambda)x^0 + \lambda x^1) \geq [f(x^0)]^{1-\lambda} [f(x^1)]^\lambda.$$

In what follows, we let $\mathcal{L}$ denote the class of log-concave symmetric probability density functions. Without loss of generality, one can assume that the center of symmetry is the origin; i.e., if $f \in \mathcal{L}$, then for any $x \in \mathbf{R}^\ell$, we have

$$f(x) = f(-x).$$

Throughout this paper we assume that the probability density function f of the vector of uncertain parameters is log-concave and symmetric, i.e., $f \in \mathcal{L}$.

2.2. Remarks. It is important to note that the class $\mathcal{L}$ of log-concave symmetric density functions is quite rich. For example, most “common” probability density functions used in the control field (such as uniform distribution over compact convex symmetric sets or Gaussian) are readily shown to be log-concave and symmetric. In other words, the main result to follow applies to typical density functions which have appeared in the probabilistic robustness literature to date.

2.3. Floating body. Central to the results presented in this paper is the concept of a *floating body* of a probability measure. This concept was first introduced by Dupin [7] in the context of uniform distributions. Given a risk level $0 < \varepsilon < 1/2$, the floating body K_ε of a probability distribution is a convex symmetric set for which each supporting hyperplane “cuts off” a set of probability ε .

More precisely, let $0 < \varepsilon < 1/2$ be given and consider a set K_ε . For any direction $u \in \mathbf{R}^\ell$, $\|u\|_2 = 1$, let $H(u)$ be the supporting hyperplane of K_ε normal to u . Also, let $\mathcal{H}^+(u)$ be the half-space defined by $H(u)$ which does not contain the origin. Then, K_ε is the *floating body associated with ε* of the given probability measure if, given any direction $\|u\|_2 = 1$,

$$\text{Prob}(\mathcal{H}^+(u)) = \varepsilon.$$

Not every probability measure has a floating body. However, the results in [9] indicate that every probability distribution with a probability density function belonging to the class $\mathcal{L}$ does have a compact convex symmetric floating body K_ε for any $0 < \varepsilon < 1/2$.

2.4. Why consider only symmetric probability density functions? It turns out that symmetry of the probability density function plays an important role in the results to follow. In general, nonsymmetric probability density functions result in a distribution that does not have a floating body. In other words, they result in a PCLP which is not a convex program. As an example, consider uncertainty $\Delta a = [\Delta a_1 \ \Delta a_2]^T$ uniformly distributed over the set

$$\{\Delta a \in \mathbf{R}^2 : \Delta a_1^2 + \Delta a_2^2 \leq 1, \Delta a_1 \geq 0, \Delta a_2 \geq 0\},$$

which has a nonsymmetric log-concave probability density function. Now, consider a risk level $\varepsilon = 0.45$. One can easily see that for $x^1 = [1 \ 0]^T$ and $x^2 = [0 \ 1]^T$, one has

$$\text{Prob}\{(x^1)^T \Delta a \leq 0.45\} = \text{Prob}\{(x^2)^T \Delta a \leq 0.45\} = 0.552 \geq 1 - \varepsilon.$$

Moreover, if one takes

$$x^3 = \frac{1}{2}x^1 + \frac{1}{2}x^2 = [0.5 \ 0.5]^T,$$

we have

$$\text{Prob}\{(x^3)^T \Delta a \leq 0.45\} = 0.5157.$$

Now, let

$$\mathcal{A}_i \doteq \{\Delta a \in \mathbf{R}^2 : (x^i)^T \Delta a \leq 0.45\},$$

and, proceeding by contradiction, assume that the floating body $K_{0.45}$ exists for this probability distribution. From the results above, we have $K_{0.45} \subseteq \mathcal{A}_1$ and $K_{0.45} \subseteq \mathcal{A}_2$. Hence,

$$K_{0.45} \subseteq \mathcal{A}_1 \cap \mathcal{A}_2.$$

However, $K_{0.45} \not\subseteq \mathcal{A}_3$. This leads to a contradiction since $\mathcal{A}_1 \cap \mathcal{A}_2 \subseteq \mathcal{A}_3$. Hence, the floating body $K_{0.45}$ does not exist for the distribution above.

2.5. Additional notation. Let $\|\cdot\|$ be a norm in $\mathbf{R}^\ell$. We define the dual norm as

$$\|x\|_* \doteq \max\{x^T y : \|y\| \leq 1\}.$$

Now, recalling that the probability density functions considered in this paper are log-concave and symmetric, given a probability density function $f \in \mathcal{L}$, let $\|\cdot\|_\varepsilon$ be the Minkowski functional associated with its floating body $K_\varepsilon \neq \emptyset$, i.e.,

$$\|x\|_\varepsilon \doteq \inf \left\{ \rho \in \mathbf{R}^+ : \frac{1}{\rho} x \in K_\varepsilon \right\}.$$

Note that the quantity above is a norm since the set K_ε is compact, convex, and symmetric. Furthermore, let $\|\cdot\|_{\varepsilon,*}$ denote its dual norm as defined above.

3. Convexity of the PCLP. In this section, we present the main result of the paper. Theorem 3.1 to follow indicates that, if the distribution of the uncertain parameters is log-concave and symmetric and if the risk levels satisfy $0 \leq \varepsilon_i \leq 1/2$, the PCLP is a convex program. Although the result below involves only the convexity of a PCLP with one constraint, the extension to the case with an arbitrary number of constraints is immediate. This extension is a consequence of the fact that an intersection of convex sets is still convex. More precisely, for multiple constraints defining $\mathcal{X}_\varepsilon$ as an intersection of

$$\mathcal{X}_{\varepsilon,i} \doteq \{x \in \mathbf{R}^\ell : \text{Prob}\{x^T a^i \leq b_i\} \geq 1 - \varepsilon_i\},$$

the convexity result is immediate because each $\mathcal{X}_{\varepsilon,i}$ is readily shown to be convex.

Throughout this paper, we write

$$a^i = a_0^i + \Delta a^i; \quad i = 1, 2, \dots, k,$$

and

$$b = b_0 + \Delta b,$$

where a_0 and b_0 are the nominal values of the random parameters a_i and b_i , respectively. We assume that the pair $(\Delta a^i, \Delta b_i)$ has a log-concave symmetric probability density function. For simplicity of exposition, it is assumed that the vector b is deterministic, i.e., $b = b_0$. The more general result can be easily obtained by noting that

$$\begin{aligned} \mathcal{X}_{\varepsilon,i} &= \{x \in \mathbf{R}^\ell : \text{Prob}\{x^T a^i - \Delta b_i \leq b_{0,i}\} \geq 1 - \varepsilon_i\} \\ &= \{\tilde{x} \in \mathbf{R}^{\ell+1} : \text{Prob}\{\tilde{x}^T \tilde{a}^i \leq b_{0,i}\} \geq 1 - \varepsilon_i\} \cap \{\tilde{x} \in \mathbf{R}^{\ell+1} : \tilde{x}_{\ell+1} = -1\}, \end{aligned}$$

where $\tilde{a}^i = [a^{iT} \Delta b_i]^T$.

THEOREM 3.1. *Let $a_0 \in \mathbf{R}^\ell$, $b \in \mathbf{R}$, and the risk level $0 \leq \varepsilon \leq 1/2$ be given. Also, let the random vector Δa have a log-concave symmetric distribution. Then, the set*

$$\mathcal{X}_\varepsilon \doteq \{x \in \mathbf{R}^\ell : \text{Prob}\{x^T(a_0 + \Delta a) \leq b\} \geq 1 - \varepsilon\}$$

is convex.

Proof. For the given $0 \leq \varepsilon \leq 1/2$, we first note that proving the convexity of the set

$$\mathcal{X}_\varepsilon \doteq \{x \in \mathbf{R}^\ell : \text{Prob}\{x^T(a_0 + \Delta a) \leq b\} \geq 1 - \varepsilon\}$$

is equivalent to proving the quasi concavity of the function

$$\varphi(x) \doteq \text{Prob}\{x^T(a_0 + \Delta a) \leq b\}$$

on the set

$$\mathcal{D} = \{x \in \mathbf{R}^\ell : \varphi(x) \geq 1/2\}.$$

In other words, given $x^0, x^1 \in \mathcal{D}$ one must show that

$$\varphi((1 - \lambda)x^0 + \lambda x^1) \geq \min\{\varphi(x^0), \varphi(x^1)\}$$

for all $0 \leq \lambda \leq 1$. Now, define

$$Q_{good}(x) \doteq \{\Delta a \in \mathbf{R}^\ell : x^T(a_0 + \Delta a) \leq b\}$$

leading to

$$\varphi(x) = \text{Prob}\{Q_{good}(x)\}$$

and extend the definition of the convex floating body to include risk levels $\varepsilon = 1/2$ and $\varepsilon = 0$ as

$$K_{1/2} = \{0\}; \quad K_0 = \{\Delta a \in \mathbf{R}^\ell : f(\Delta a) \neq 0\},$$

where $f(\cdot)$ denotes the probability density function. Note that both are convex symmetric sets and that, for any $0 \leq \varepsilon \leq 1/2$,

$$\varphi(x) \geq 1 - \varepsilon \Leftrightarrow Q_{good}(x) \supseteq K_\varepsilon.$$

Now, take $x^0, x^1 \in \mathcal{D}$ and let $0 \leq \varepsilon \leq 1/2$ be defined as

$$\varepsilon \doteq 1 - \min\{\varphi(x^0), \varphi(x^1)\}.$$

Now, since

$$\min\{\varphi(x^0), \varphi(x^1)\} = 1 - \varepsilon,$$

the definition of convex floating body implies the inclusions

$$Q_{good}(x^0) \supseteq K_\varepsilon; \quad Q_{good}(x^1) \supseteq K_\varepsilon.$$

Hence, for any $\lambda \in [0, 1]$,

$$Q_{good}((1 - \lambda)x^0 + \lambda x^1) \supseteq Q_{good}(x^0) \cap Q_{good}(x^1) \supseteq K_\varepsilon.$$

Therefore,

$$\varphi((1 - \lambda)x^0 + \lambda x^1) \geq 1 - \varepsilon = \min\{\varphi(x^0), \varphi(x^1)\}$$

or, in other words, $\varphi(x)$ is a quasi-concave function for all $x \in \mathcal{D}$. This completes the proof. $\square$

4. On the numerical implementation of the PCLP. The result in the previous section indicates that the PCLP is a convex program. However, it does not provide any indication of how to numerically solve the resulting optimization problem. In this section, we use the concept of *floating body* introduced in section 2.3 to provide some insight into how one can numerically solve the PCLP.

4.1. A closed-form formulation of the PCLP. Since the set

$$\{\Delta a \in \mathbf{R}^\ell : x^T(a_0 + \Delta a) \leq b\}$$

is a half-space, the definition of the floating body presented in section 2.3 indicates that requiring

$$\text{Prob}\{x^T(a_0 + \Delta a) \leq b\} \geq 1 - \varepsilon$$

for $0 < \varepsilon < 1/2$ is equivalent to requiring

$$x^T(a_0 + \Delta a) \leq b$$

for all $\Delta a \in K_\varepsilon$, where K_ε is the floating body of the probability distribution of Δa as defined in section 2.3. Now, given the definition of dual norm, this is equivalent to

$$\|x\|_{\varepsilon,*} \leq b - x^T a_0.$$

Therefore, the probabilistic constraints of the PCLP can be replaced by convex ones of the form above. Hence, if the quantity $\|x\|_{\varepsilon,*}$ can be easily determined, this leads to an immediate numerical implementation for solving the PCLP.

4.2. Elliptical log-concave distributions. It turns out that there are cases where $\|x\|_{\varepsilon,*}$ is easily determined. An example is the case where the probability distribution of the uncertain parameters is an *elliptical log-concave* distribution. An elliptical log-concave distribution is a distribution whose probability density function is of the form

$$f(y) = g(y^T W y),$$

where $g : \mathbf{R}_0^+ \rightarrow \mathbf{R}_0^+$ is a log-concave nonincreasing function and W is a positive definite matrix. Examples of such distributions are multivariable normal distributions and uniform distributions over balls. Note that the function above is indeed log-concave. This can be proven by using the fact that the composition of a log-concave nonincreasing function with a convex function is itself log-concave. For such probability distributions one can determine the “shape” of the floating body. This result is presented below.

LEMMA 4.1. *Consider an elliptical log-concave distribution. Then, for any $0 < \varepsilon < 1/2$, there exists a radius $r(\varepsilon)$ leading to*

$$K_\varepsilon = \{\Delta a \in \mathbf{R}^\ell : \Delta a^T W \Delta a \leq r^2(\varepsilon)\}.$$

Sketch of proof. As mentioned in the definition of an elliptical log-concave distribution, its probability density function is of the form

$$f(\Delta a) = g((\Delta a)^T W \Delta a),$$

where $g(\cdot)$ is log-concave. Now let

$$\Delta\tilde{a} = \frac{1}{\det(W^{1/2})} W^{1/2} \Delta a,$$

where the symmetric matrix $W^{1/2}$ is such that $W = W^{1/2} W^{1/2}$. Since the Jacobian of the linear transformation above can be easily shown to be 1, it follows that the probability density function of $\Delta\tilde{a}$ is given by

$$f_{\Delta\tilde{a}}(\Delta\tilde{a}) = g \left[\det^2(W^{1/2}) (\Delta\tilde{a})^T \Delta\tilde{a} \right].$$

The contours of $f_{\Delta\tilde{a}}$ are hyperspheres and, hence, this probability distribution is “direction independent”; i.e., given any constant b and any two vectors $\|u\| = 1$ and $\|v\| = 1$, we have

$$\text{Prob}\{u^T \Delta\tilde{a} \leq b\} = \text{Prob}\{v^T \Delta\tilde{a} \leq b\}.$$

Hence, given any $0 < \varepsilon < 1/2$, the floating body of this distribution is of the form

$$\tilde{K}_\varepsilon = \{\Delta\tilde{a} \in \mathbf{R}^\ell : \Delta\tilde{a}^T \Delta\tilde{a} \leq \tilde{r}^2(\varepsilon)\}$$

for some $\tilde{r}(\varepsilon)$. Since there is a linear relationship between Δa and $\Delta\tilde{a}$, one can see that the floating body of the probability distribution of Δa is given by

$$K_\varepsilon = \{\Delta a \in \mathbf{R}^\ell : \Delta a^T W \Delta a \leq \tilde{r}^2(\varepsilon) \det^2(W^{1/2})\}.$$

This completes the proof. $\square$

The actual “radius” of the ellipsoid K_ε can be determined analytically for some probability distributions. If one cannot determine this radius analytically, a one line search optimization problem can be set up to numerically approximate this value by performing several Monte Carlo simulations. Therefore, for such probability distributions, one can easily see that the PCLP reduces to a convex quadratic optimization problem, i.e., minimization of a linear function subject to quadratic constraints. More precisely, consider an elliptical log-concave probability density function of the form described above. Then, for any $0 < \varepsilon < 1/2$, the floating body K_ε is of the form

$$K_\varepsilon = \{\Delta a \in \mathbf{R}^\ell : \Delta a^T W \Delta a \leq r^2(\varepsilon)\},$$

where $r(\varepsilon) > 0$ is such that the half-space $H = \{\Delta a \in \mathbf{R}^\ell : \Delta a_1 \leq \zeta\}$ with

$$\text{Prob}(H) = 1 - \varepsilon$$

results in a hyperplane $\{\Delta a \in \mathbf{R}^\ell : \Delta a_1 = \zeta\}$ tangent to the ellipsoid

$$\{\Delta a \in \mathbf{R}^\ell : \Delta a^T W \Delta a \leq r^2(\varepsilon)\}.$$

Now let

$$\Delta\gamma = \frac{1}{r(\varepsilon)} W^{1/2} \Delta a,$$

where $W^{1/2}$ is a positive definite matrix satisfying $W^{1/2} W^{1/2} = W$. Note that $\Delta a \in K_\varepsilon$ if and only if $\|\Delta\gamma\|_2 \leq 1$. Then, requiring

$$x^T(a_0 + \Delta a) \leq b$$

for all $\Delta a \in K_\varepsilon$ is equivalent to requiring

$$x^T(a_0 + r(\varepsilon)W^{-1/2}\Delta\gamma) \leq b$$

for all $\|\Delta\gamma\|_2 \leq 1$. Hence, since the dual of the 2-norm is the 2-norm itself, the expression above is equivalent to

$$\|r(\varepsilon)W^{-1/2}x\|_2 \leq b - x^T a_0,$$

which is a convex quadratic constraint on x .

5. Application to control systems design. In this section, we exemplify how the concept of a PCLP can be applied to controller design. We start by applying the PCLP to the design of superstable systems. A second example shows how the theory in this paper can be applied to robust pole assignment.

5.1. Superstability. In contrast to the concept of ℓ^∞ stability, where only input/output behavior is considered, superstability allows for computing the worst-case value of the ℓ^∞ norm of the output due to ℓ^∞ bounded disturbances and initial conditions. In addition, it provides an upper bound on the ℓ^∞ induced norm of the system (this bound is exact in the case of finite impulse response (FIR) systems). The reader is referred to [10] and [11] for proofs and additional properties.

Consider a discrete-time linear time invariant system

$$y(z) = G(z)w(z), \quad G(z) = \frac{b(z)}{1 + a(z)},$$

where the w are exogenous disturbances, y is the output, z is the delay operator $zx[k] = x[k-1]$, and where the polynomial $a(z)$ does not have a constant term, i.e.,

$$a(z) = a_1z + a_2z^2 + \cdots + a_nz^n.$$

Define

$$\|a\|_1 = \sum_{i=1}^n |a_i|.$$

A system is said to be superstable if

$$\|a\|_1 < 1.$$

Moreover, it can be shown (see [10]) that in this case the ℓ^∞ induced norm of the system is bounded by

$$\|G(z)\|_{\ell^\infty \rightarrow \ell^\infty} \leq \frac{\|b\|_1}{1 - \|a\|_1}.$$

This property was exploited in [10] to synthesize low order ℓ^1 controllers as follows. Consider a feedback controller of the form

$$G_c(z) = \frac{b_c(z)}{1 + a_c(z)}.$$

Synthesizing a controller of the form described above such that the ℓ^1 norm of the closed loop system is less than or equal to a given μ reduces to finding controller coefficients satisfying

$$\mu \|d_{cl}\|_1 + \|n_{cl}\|_1 \leq \mu,$$

where

$$d_{cl}(z) = a(z) + a_c(z) + a(z)a_c(z) + b(z)b_c(z); \quad n_{cl}(z) = b(z)b_c(z).$$

Hence, in this case, controller design is equivalent to finding a feasible point of a set of linear inequalities.

Now, assume that the plant is subject to parametric uncertainty of the form

$$G(z) = \frac{b(z)}{1+a(z)} = \frac{b_{0,0} + \Delta b_0 + (b_{0,1} + \Delta b_1)z + \cdots + (b_{0,m} + \Delta b_m)z^m}{1 + (a_{0,1} + \Delta a_1)z + \cdots + (a_{0,n} + \Delta a_n)z^n},$$

where $b_{0,i}$ and $a_{0,i}$ are the nominal values of the coefficients of the numerator and denominator, respectively, and Δb_i and Δa_i represent the uncertainty.¹ In this case robust performance is achieved if the inequality

$$\mu \|d_{cl}\|_1 + \|n_{cl}\|_1 \leq \mu$$

holds for all admissible values of the uncertainty, which is a linear program. More precisely, given $\mu > 0$, the problem above can be recast as

$$A(\Delta a_0, \dots, \Delta a_n, \Delta b_0, \dots, \Delta b_m) \begin{bmatrix} a_{c,0} \\ \vdots \\ a_{c,n_c} \\ b_{c,0} \\ \vdots \\ b_{c,m_c} \end{bmatrix} \leq \mu \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}^T,$$

where $a_{c,0}, \dots, a_{c,n_c}$ and $b_{c,0}, \dots, b_{c,m_c}$ denote the coefficients of the denominator and the numerator of the controller, respectively. However, there is a major difference between the nominal and robust performance case: while it can be shown that the former always admits a solution if the controller order is chosen to be at least equal to the order of the plant, the latter may not have a solution even for high order controllers. On the other hand, as we illustrate next with a simple example, it might be possible to find low order risk-adjusted controllers, even for very small risk values of violating the constraints; i.e., given a distribution for the uncertain parameters and a small, well-defined risk ε , one may find a low order controller such that the probability of violating each of the resulting linear constraints is less than ε . In other words, in order to reduce controller complexity, one solves a PCLP for a given risk level ε .

¹For notational simplicity, here we assume that the controller is not subject to uncertainty; however, the proposed procedure can be easily modified to take controller uncertainty into account.

5.2. Numerical example. We now consider the example in [10]. The discrete-time system presented has nominal transfer function

$$P(z) = \frac{n(z)}{d(z)} = \frac{z - 2.5z^2 + 1.501z^3}{1 - 2.7z + 23.5z^2 - 4.6z^3}$$

and we assume that all coefficients are subject to uncertainty. Moreover, we assume that the uncertainty vector is uniformly distributed on a ball with radius 0.05. In searching for the controller, we assume that it has the form

$$G_c(z) = \frac{b_c(z)}{1 + a_c(z)} = \frac{b_{c,0} + b_{c,1}z + \cdots + b_{c,m_c}z^{m_c}}{1 + a_{c,1}z + \cdots + a_{c,n_c}z^{n_c}}.$$

It can be shown that there is no controller of order less than or equal to $n_c = 6$ that renders the closed loop system robustly superstable. Next, we allowed for a risk of $\varepsilon = 1.25 \times 10^{-4}$. Since, in this case, the distribution of the uncertainty is an elliptical log-concave distribution, we can use the results in section 4 to determine a risk-adjusted controller. As mentioned in section 4.2, the floating body for this distribution is a ball and, hence, one can set up a quadratic program to solve the problem at hand. Matlab was used to solve the resulting quadratic program, and the third order risk-adjusted controller

$$C(z) = \frac{4.5819 - 17.7802z - 1.0245z^2 + 0.8795z^3}{1 - 1.8819z + 0.6538z^2 + 0.287z^3}$$

was obtained. Next, a Monte Carlo simulation was performed to assess the risk of violating superstability (recall that ε is the risk of violating each inequality, not the actual risk of violating at least one of the inequalities). The number of samples used was 10^7 and the estimated probability of violating superstability obtained is 0.78%, showing that one can obtain a low order controller even for small risk levels.

5.3. Robust pole assignment. We now describe how one can apply the results in this paper to the problem of robust pole assignment. We start with a continuous uncertain open loop plant described by the transfer function

$$G(s) = \frac{n(s)}{d(s)} = \frac{(b_{0,0} + \Delta b_0) + (b_{0,1} + \Delta b_1)s + \cdots + (b_{0,m} + \Delta b_m)s^m}{(a_{0,0} + \Delta a_0) + (a_{0,1} + \Delta a_1)s + \cdots + (a_{0,n} + \Delta a_n)s^n},$$

where $b_{0,i}$ and $a_{0,i}$ are the nominal values of the coefficients of the numerator and denominator, respectively, and Δb_i and Δa_i represent the uncertainty. Now, since uncertainty is present, one cannot determine a controller that will assign the closed loop poles to a specific location. As in [13], one instead aims at designing a controller such that the closed loop characteristic polynomial belongs to a set defined by the performance specifications. More precisely, each of the coefficients of the closed loop characteristic polynomial should belong to a given interval; i.e., given a controller of the form

$$G_c(s) = \frac{n_c(s)}{d_c(s)} = \frac{b_{c,0} + b_{c,1}s + \cdots + b_{c,m_c}s^{m_c}}{a_{c,0} + a_{c,1}s + \cdots + a_{c,n_c}s^{n_c}},$$

one aims at determining the coefficients of the closed loop controller such that the closed loop characteristic polynomial

$$n_c(s)n(s) + d_c(s)d(s)$$

belongs to the family of polynomials

$$s^{n_{cl}} + [\delta_{n_{cl}-1}^-, \delta_{n_{cl}-1}^+] s^{n_{cl}-1} + \cdots + [\delta_1^-, \delta_1^+] s + [\delta_0^-, \delta_0^+]$$

for all admissible uncertainty values, where $n_{cl} = n_c + n$ is the degree of the closed loop characteristic polynomial. As can be easily seen, the search for the coefficients of the controller reduces to finding a feasible solution to a set of linear inequalities of the form

$$\begin{bmatrix} \delta_0^- \\ \vdots \\ \delta_{n_{cl}-1}^- \end{bmatrix} \leq A(\Delta a_0, \dots, \Delta a_n, \Delta b_0, \dots, \Delta b_m) \begin{bmatrix} a_{c,0} \\ \vdots \\ a_{c,n_c} \\ b_{c,0} \\ \vdots \\ b_{c,m_c} \end{bmatrix} \leq \begin{bmatrix} \delta_0^+ \\ \vdots \\ \delta_{n_{cl}-1}^+ \end{bmatrix},$$

where $A(\Delta a_0, \dots, \Delta a_n, \Delta b_0, \dots, \Delta b_m)$ is an affine function of its arguments. To satisfy the performance specifications, the inequalities above should be satisfied for all admissible values of $\Delta a_0, \dots, \Delta a_n$ and $\Delta b_0, \dots, \Delta b_m$. If the uncertainty support set is convex and symmetric, the problem above turns out to be convex. This is easily seen if one rewrites the inequalities above as follows: Let $q = [\Delta a_0, \dots, \Delta a_n, \Delta b_0, \dots, \Delta b_m]^T$ and $x = [a_{c,0}, \dots, a_{c,n_c}, b_{c,0}, \dots, b_{c,m_c}]^T$. Then, for the uncertainty vector q , whose support is a convex symmetric set Q , the inequalities above can be written in the form

$$g_i(x)^T q \leq h_i(x) + k_i; \quad i = 1, 2, \dots, n_c + m_c + 2,$$

for all $\|q\| \leq 1$, where $\|\cdot\|$ is the norm whose unit ball is the set Q and $g_i(x)$ and $f_i(x)$ are affine functions of x . Hence, we can rewrite the inequalities above as

$$\|g_i(x)\|_* \leq h_i(x) + k_i; \quad i = 1, 2, \dots, n_c + m_c + 2,$$

which is easily seen to be a convex constraint in x .

However, even though in many cases the problem above is convex, it can lead to complex controllers. To avoid this, we take a risk-adjusted point of view of the problem above; i.e., rather than requiring that each inequality be satisfied for all admissible values of the uncertain parameters, we require that the risk of violating each of the inequalities be less than or equal to a prescribed risk level ε . In other words, we solve a PCLP version of the problem above.

5.4. Numerical example. The example presented here is a small modification of one of the examples in [13]. Consider an uncertain plant with transfer function

$$G(s) = \frac{(0.75 + \Delta b_1)s + 1.25 + \Delta b_0}{s^2 + (0.75 + \Delta a_1)s + \Delta a_0},$$

where the uncertain parameter vector is uniformly distributed over an ellipsoid. More precisely

$$\begin{bmatrix} \Delta a_0 \\ \Delta a_1 \\ \Delta b_0 \\ \Delta b_1 \end{bmatrix} = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \Delta \alpha,$$

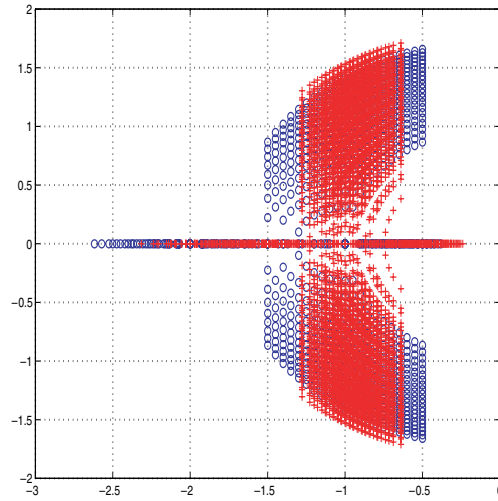


FIG. 5.1. *Desired pole location “o” and actual pole location “+”.*

where $\Delta\alpha \in \mathbf{R}^4$ is uniformly distributed over the ball of radius 0.25. We now aim at designing a controller such that the closed loop polynomial belongs to the family

$$\Delta_T(s) = s^2 + [1, 3]s + [1, 3].$$

Therefore, the controller transfer function is the constant

$$G_c(s) = b_0.$$

It can be proven that no constant controller can meet the specification for all admissible values of the uncertain parameters. Next, we allowed for a risk of $\varepsilon = 0.02$. As in the previous numerical example, the distribution of the uncertainty is elliptical and log-concave and, hence, one can use the results in section 4 to determine a risk-adjusted controller. Again, the floating body for this distribution is a ball, and one can set up a quadratic program to solve the problem at hand. Matlab was used to solve the resulting quadratic program and the risk-adjusted constant controller

$$G_s(s) = 1.555$$

was obtained. The pole cluster distributions of the desired system and the actual closed loop system are shown in Figure 5.1. A Monte Carlo simulation was performed to estimate the actual risk of violating the specifications. The estimated value of the risk obtained was approximately 3.6%, illustrating another instance where, even for low risk values, one can obtain risk-adjusted controllers in cases where a robust controller does not exist. Furthermore, in this case, we were able to obtain robust stability as an added benefit; see Figure 5.1.

6. Conclusions and further research. In this paper, we show that the PCLP is a convex optimization problem for any log-concave symmetric distribution. Also, a closed-form formulation of the PCLP was provided which can be easily implemented in the case of elliptical distributions, such as normal or uniform over a ball. This result was applied to systems design, showing that classical control theory can lead to complex controllers. Namely, we have presented examples where, even for very low

levels of risk, one can obtain controllers that are substantially less complex than their robust counterparts.

The results in this paper suggest several directions for further research. First, we believe that effort should be put into the development of numerical tools for solving the PCLP when the distribution is other than elliptical. Also, it seems that the “ratio” between the complexity of a robust controller and the complexity of the risk-adjusted controller increases with the dimension of the uncertainty vector. Therefore, it would be of interest to quantify how this difference in complexity depends on the uncertainty dimension. Finally, we note that results to date deal only with risk constraints. It would be of great interest to develop design procedures that would take into account both risk and robust constraints. For example, one might want to design a control system that would result in a robustly stable system having a risk of violating the performance specification that is less than or equal to a given risk level.

REFERENCES

- [1] L. R. RAY AND R. F. STENGEL, *A Monte Carlo approach to the analysis of control system robustness*, Automatica, 29 (1993), pp. 229–236.
- [2] R. F. STENGEL AND L. R. RAY, *Stochastic robustness of linear time-invariant control systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 82–87.
- [3] A. YOON, P. KHARGONEKAR, AND K. HEBBALE, *Design of computer experiments for open-loop control and robustness analysis of clutch-to-clutch shifts in automatic transmissions*, in Proceedings of the American Control Conference (Albuquerque, NM), Vol. 5, IEEE, Piscataway, NJ, 1997, pp. 3359–3364.
- [4] X. CHEN AND K. ZHOU, *Constrained optimal synthesis and robustness analysis by randomized algorithms*, in Proceedings of the American Control Conference (Philadelphia, PA), Vol. 3, IEEE, Piscataway, NJ, 1998, pp. 1429–1433.
- [5] B. R. BARMISH AND C. M. LAGOA, *On convexity of the probabilistic design problem for quadratic stabilizability*, in Proceedings of the American Control Conference (San Diego, CA), Vol. 1, IEEE, Piscataway, NJ, 1999, pp. 430–434.
- [6] C. M. LAGOA, *A convex parameterization of risk-adjusted stabilizing controllers*, in Proceedings of the 38th Annual IEEE Conference on Decision and Control, Phoenix, AZ, Vol. 1, 1999, pp. 534–539.
- [7] C. DUPIN, *Application de Géométrie et de Mécanique à la Marine*, aux Ponts et Chaussées, Paris, 1822.
- [8] S. MEYER AND S. REISNER, *A geometric property of the boundary of symmetric convex bodies and convexity of flotation surfaces*, Geom. Dedicata, 37 (1991), pp. 327–337.
- [9] M. MEYER AND S. REISNER, *Characterizations of affinity-rotation-invariant log-concave measures by section-centroid location*, in Geometric Aspects of Functional Analysis (1989–90), Lecture Notes in Math. 1469, Springer, Berlin, 1991, pp. 145–152.
- [10] F. BLANCHINI AND M. SZNAIER, *A convex optimization approach to fixed-order controller design for disturbance rejection in SISO systems*, in Proceedings of the 36th Annual IEEE Conference on Decision and Control, San Diego, CA, 1997, pp. 1540–1545.
- [11] B. T. POLYAK AND M. E. HALPERN, *The use of a new optimization criterion for discrete-time feedback controller design*, in Proceedings of the 38th Annual IEEE Conference on Decision and Control, Phoenix, AZ, 1999, pp. 1540–1545.
- [12] L. H. KEEL AND S. P. BHATTACHARYYA, *Robust, fragile or optimal*, IEEE Trans. Automat. Control, 42 (1997), pp. 1098–1105.
- [13] L. H. KEEL AND S. P. BHATTACHARYYA, *A linear programming approach to controller design*, in Proceedings of the 36th Annual IEEE Conference of Decision and Control, San Diego, CA, 1997, pp. 2139–2148.
- [14] B. R. BARMISH AND C. M. LAGOA, *The uniform distribution: A rigorous justification for its use in robustness analysis*, Math. Control Signals Systems, 10 (1997), pp. 203–222.
- [15] C. VAN DE PANNE AND W. POPP, *Minimum-cost cattle feed under probabilistic protein constraints*, Management Sci., 9 (1963), pp. 405–430.
- [16] S. VAJDA, *Probabilistic Programming*, Academic Press, New York, 1972.
- [17] A. PRÉKOPA, *Stochastic Programming*, Kluwer Academic Publishers, Boston, 1995.

PROXIMAL METHODS IN VECTOR OPTIMIZATION*

HENRI BONNEL[†], ALFREDO NOEL IUSEM[‡], AND BENAR FUX SVAITER[‡]

Abstract. We consider the vector optimization problem of finding weakly efficient points for maps from a Hilbert space X to a Banach space Y , with respect to the partial order induced by a closed, convex, and pointed cone $C \subset Y$ with a nonempty interior. We develop for this problem an extension of the proximal point method for scalar-valued convex optimization. In this extension, the subproblems consist of finding weakly efficient points for suitable regularizations of the original map. We present both an exact and an inexact version, in which the subproblems are solved only approximately, within a constant relative tolerance. In both cases, we prove weak convergence of the generated sequence to a weakly efficient point, assuming convexity of the map with respect to C and C -completeness of the initial section. In cases where this last assumption fails, we still establish that the generating sequence is, in a suitable sense, a minimizing one. We also exhibit a particular instance of the algorithm for which, under a mild hypothesis on C , the weak limit of the generated sequence is an efficient, rather than a weakly efficient, point.

Key words. vector optimization, proximal point, inexact algorithm

AMS subject classifications. 49J52, 49M99, 65K05, 65K10, 90C25, 90C29, 90C48

DOI. 10.1137/S1052623403429093

1. Introduction. We discuss methods for vector-valued optimization in the following setting. We consider maps $F : X \rightarrow Y$, where X is a real Hilbert space and Y is a real Banach space containing a closed, convex, and pointed cone C with nonempty interior, which defines a partial order $\preceq_C$ in Y , given by $y \preceq_C y'$ if and only if $y' - y$ belongs to C , with its associate relation $\prec_C$, given by $y \prec_C y'$ if and only if $y' - y$ belongs to the interior $\text{int}(C)$ of C . Actually, we admit the possibility that F takes the value $+\infty$; this is made precise in section 2. Our goal is to analyze methods for finding a weakly efficient minimizer of F with respect to $\preceq_C$, meaning a point $a \in X$ such that there exists no $x \in X$ satisfying $F(x) \prec_C F(a)$.

This paper is part of a wider research program consisting of the extension to vector-valued optimization of several iterative methods for scalar-valued methods. In these extensions, we attempt to define the iterates in the vector-valued case by considering the order $\preceq_C$ in Y , mimicking, whenever possible, the role of the usual order in $\mathbb{R}$ in the corresponding algorithm for scalar-valued optimization.

Several methods already have been extended in this fashion in a finite dimensional setting. The case of the steepest descent method for multiobjective optimization (i.e., when C is the nonnegative orthant of $\mathbb{R}^n$) was dealt with in [7]; the same method for general finite dimensional vector optimization (i.e., for partial orders given by rather general cones in $\mathbb{R}^n$) is analyzed in [10]. An extension of the projected gradient method to the case of convexly constrained vector optimization, for the order given by a rather general cone in $\mathbb{R}^n$, can be found in [9].

*Received by the editors June 3, 2003; accepted for publication (in revised form) August 12, 2004; published electronically June 3, 2005. Part of this work was done while the first author visited the Instituto de Matemática Pura e Aplicada, Rio de Janeiro and was supported by the “Agreement Brasil/France in Mathematics” between CNRS and CNPq.

<http://www.siam.org/journals/siopt/15-4/42909.html>

[†]Université de la Nouvelle-Calédonie, ERIM, B.P. 4477, 98847, Nouméa, France (bonnel@univ-nc.nc). Former name: Serban Bolintineanu.

[‡]Instituto de Matemática Pura e Aplicada, Estrada Dona Casterina, 110, CEP 22460-320, Rio de Janeiro, Brazil (iusp@impa.br, benar@impa.br).

In this paper, we perform a similar extension for the case of the proximal point method for scalar-valued convex optimization. Thus, we give next a brief description of this method. Given a Hilbert space X and a point-to-set operator $T : X \rightarrow \mathcal{P}(X)$, the proximal point method, in its so-called exact version, is an iterative procedure for finding a zero of T , i.e., a point $x^* \in X$ such that $0 \in T(x^*)$. The method generates a sequence $\{x_n\} \subset X$, starting from an arbitrary $x_0 \in X$, through the following iteration: given a bounded exogenous sequence of positive real numbers $\{\alpha_n\}$ (called regularization parameters) and the current iterate x_n , x_{n+1} is the unique vector in X such that

$$(1) \quad 0 \in T_n(x_{n+1}),$$

where $T_n : X \rightarrow \mathcal{P}(X)$ is defined as $T_n(x) = T(x) + \alpha_n(x - x_n)$.

The origins of the proximal method can be traced back to the 1960s (see, e.g., [16], [13], and [14]); it attained the formulation given above in Rockafellar's works in the 1970s (see [19] and [20]).

It has been proved in [20] that whenever T is maximal monotone and has zeroes, the sequence $\{x_n\}$ generated by the proximal point method converges to a zero of T in the weak topology of X .

Now take a convex function $f : X \rightarrow \mathbb{R} \cup \{+\infty\}$ and a closed and convex subset $K \subset X$. The solutions of the convex optimization problem $\min f(x)$ subject to $x \in K$ are precisely the zeroes of the maximal monotone operator $T = \partial(f + I_K)$, where ∂ denotes the subdifferential of a convex function and I_K is the indicator function of K , defined as $I_K(x) = 0$ if x belongs to K and $I_K(x) = +\infty$ otherwise. Thus, the proximal point method can be used to solve the above-mentioned optimization problem; in this case, it is easy to check that the iteration (1) takes the form

$$(2) \quad x_{n+1} \in \operatorname{argmin}_{x \in K} \left\{ f(x) + \frac{\alpha_n}{2} \|x - x_n\|^2 \right\}.$$

Next we discuss inexact versions of the method, in which case x_{n+1} need not be the exact solution of the optimization subproblem in (2), but just an approximate solution of it. Clearly the analysis of an inexact version is essential if one wants the convergence results to hold for actual implementations of the method. Inexact versions were proposed as early as 1976 (see [20]), in which the n th subproblem was allowed to be solved within a prescribed tolerance τ_n , and it was demanded that $\sum_{n=1}^{\infty} \tau_n < \infty$. Similar error criteria, requesting always summability of the tolerances, appeared in several papers later on; all these criteria are quite demanding and difficult to check in actual implementations. More recently, less stringent error criteria, based on relative rather than absolute errors, were introduced by Solodov and Svaiter [21]. In one of these approaches, the following iteration is considered for the problem $\min f(x)$ subject to $x \in X$, with a convex $f : X \rightarrow \mathbb{R} \cup \{+\infty\}$: given x_n , take as x_{n+1} any vector $x \in X$ such that there exists $\varepsilon_n \in \mathbb{R}_+$ satisfying

$$(3) \quad 0 \in \partial_{\varepsilon_n} f(x) + \alpha_n(x - x_n),$$

$$(4) \quad \varepsilon_n \leq \sigma \frac{\alpha_n}{2} \|x - x_n\|^2,$$

where $\sigma \in [0, 1)$ is a relative error constant.

We recall that, for $\varepsilon \geq 0$,

$$(5) \quad \partial_{\varepsilon} f(x) = \{u \in X : f(y) - f(x) - \langle u, x - y \rangle \geq -\varepsilon \ \forall y \in X\}.$$

The conditions in (3)–(4) simply mean that $\alpha_n(x_n - x_{n+1})$ must belong to $\partial_{\varepsilon_n} f(x_{n+1})$, with the definition of ∂_{ε_n} given above, and with ε_n satisfying the inequality in (4) for $x = x_{n+1}$.

The algorithm given in (3)–(4) reduces to the exact version for the optimization case as in (2), when $\sigma = 0$ (or, equivalently, $\varepsilon_n = 0$ for all n).

This algorithm was developed in [21] and [5] for the more general problem of finding zeroes of operators. The exact version of this method is the one presented in (1), where, instead of $\partial_{\varepsilon_n} f_n$ with $f_n : x \mapsto f(x) + \frac{\alpha_n}{2} \|x - x_n\|^2$, an adequate enlargement of the operator T_n , related to ε_n , is used. This type of enlargement was introduced in [4].

The results in [21] and [5] establish that the sequence generated by (3)–(4) converges to a minimizer of f in the weak topology of X under the same assumptions required in the exact case, namely, convexity of f and existence of solutions of the optimization problem.

The above discussion refers, of course, to the proximal method for scalar-valued convex optimization. The essence of this paper consists of the extension of both the exact proximal method (2) and its inexact counterpart (3)–(4) to the vector-valued optimization problem introduced at the beginning of this section. Basically, in the exact case the n th subproblem will consist of finding weakly efficient minimizers of $F_n : X \rightarrow Y$, with

$$(6) \quad F_n(x) = F(x) + \alpha_n \|x - x_n\|^2 e_n,$$

restricted to the set $\Omega_n \subset X$ defined as $\Omega_n = \{x \in X : F(x) \preceq_C F(x_n)\}$, where e_n is an exogenously selected vector belonging to $\text{int}(C)$ and such that $\|e_n\| = 1$.

For our inexact version, we consider the topological dual space Y^* of Y , the positive polar cone $C^+ \subset Y^*$, given by $C^+ = \{z \in Y^* : \langle y, z \rangle \geq 0 \text{ for all } y \in C\}$ (where $\langle \cdot, \cdot \rangle : Y \times Y^* \rightarrow \mathbb{R}$ is the duality pairing), and the indicator function I_{Ω_n} , of the set Ω_n , defined as above. We take an exogenous sequence $\{z_n\} \subset C^+$, with $\|z_n\| = 1$ for all $n \in \mathbb{N}$, and define, at iteration n , the function $f_n : X \rightarrow \mathbb{R} \cup \{+\infty\}$ as

$$(7) \quad f_n(x) = \langle F(x), z_n \rangle + I_{\Omega_n}(x).$$

Then we take as x_{n+1} any vector $x \in X$ such that there exists $\varepsilon_n \in \mathbb{R}_+$ satisfying

$$(8) \quad 0 \in \partial_{\varepsilon_n} f_n(x) + \alpha_n \langle e_n, z_n \rangle (x - x_n),$$

$$(9) \quad \varepsilon_n \leq \sigma \frac{\alpha_n}{2} \langle e_n, z_n \rangle \|x_n - x\|^2,$$

where $\sigma \in [0, 1)$ is again a measure of the relative error.

We will establish that any sequence generated by either our exact or inexact version converges, in the weak topology of X , to a weakly efficient minimizer of F under the following two assumptions:

- (i) F is C -convex, i.e., $F(\lambda x + (1 - \lambda)x') \preceq_C \lambda F(x) + (1 - \lambda)F(x')$ for all $x, x' \in X$ and all $\lambda \in [0, 1]$.
- (ii) The set $(F(x_0) - C) \cap F(X)$ is C -complete; i.e., for every sequence $\{a_n\} \subset X$, with $a_0 = x_0$, such that $F(a_{n+1}) \preceq_C F(a_n)$ for all $n \in \mathbb{N}$, there exists $a \in X$ such that $F(a) \preceq_C F(a_n)$ for all $n \in \mathbb{N}$.

In the absence of assumption (ii), we establish weaker convergence results, namely, that the generated sequence is a minimizing one for our problem, meaning that $\{F(x_n)\}$ approaches the set of infimal values of F , in a sense made precise in Propositions 3.3 and 3.4 of section 3.

We mention that the exact form of the algorithm is indeed a particular case of the inexact one, corresponding to choosing $\sigma = 0$, or, equivalently $\varepsilon_n = 0$ for all n , in the sense that any vector x_{n+1} satisfying (7)–(9) with $\sigma = 0$ is a weakly efficient minimizer of F_n as defined in (4). Thus, a separate analysis of the exact version might seem superfluous. We have included it because the analysis of the inexact version is somewhat obscured by the introduction of the sequences ε_n and z_n ; the exact version is “cleaner” and allows for an easier visualization of the essential arguments in the convergence proofs. Also, both versions are presented somewhat differently: the subproblems of the exact one are (vector-valued) optimization problems, whereas in each subproblem of the inexact version we look for zeroes of approximate subdifferentials of scalar-valued convex functions. This rather inessential difference is elaborated upon in Remark 5 of section 5.

We emphasize that our analysis holds without requiring reflexivity of the Banach space Y . This detail is not negligible, because we do need the cone C to have a nonempty interior. The prototypical infinite dimensional Banach spaces are the $\mathcal{L}^p$ spaces ($1 \leq p \leq \infty$), and the most relevant cones in them are the so-called positive cones, consisting of all p -integrable functions, which are nonnegative almost everywhere. As is well known, these cones have empty interior, except for the case of $\mathcal{L}^\infty$, which happens to be nonreflexive. Thus, our analysis covers at least one meaningful example, where the order is induced by a cone in an infinite dimensional space.

Next, we discuss the relation between our results and some previously published proximal procedures for vector optimization.

A so-called multiobjective proximal bundle method can be found in [15]. It deals with the finite dimensional and multiobjective case, i.e., with $X = \mathbb{R}^p$, $Y = \mathbb{R}^m$, and $C = \mathbb{R}_+^m$. If $F(x) = (f_1(x), \dots, f_m(x))$, an *improvement function* $H : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$ is introduced, defined as $H(x, y) = \max_{1 \leq i \leq m} \{f_i(x) - f_i(y)\}$. It is the case that the weakly efficient points for F are precisely the points $x^* \in \mathbb{R}^p$ such that $x^* = \operatorname{argmin}_{x \in \mathbb{R}^p} H(x, x^*)$, assuming C -convexity of F and a Slater condition. The method consists of generating a sequence $\{x^n\} \subset \mathbb{R}^p$ such that x^{n+1} is of the form $x^n + \omega_n d^n$, where $\omega_n \in \mathbb{R}_{++}$ is a step-size, and the direction $d^n \in \mathbb{R}^p$ should be, in principle, the solution of the problem $\min H(x^n + d, x^n)$ subject to $d \in \mathbb{R}^p$. In the actual implementation, the objective function of this subproblem is replaced by a piecewise linear approximation $\hat{H}^n : \mathbb{R}^p \rightarrow \mathbb{R}$, and a proximal-like regularization term is added so that the direction d^n is indeed the solution of $\min_{d \in \mathbb{R}^p} \{\hat{H}^n(x^n + d) + \alpha_n \|d\|^2\}$, where $\alpha_n \in \mathbb{R}_{++}$ is a regularization parameter. Apart from the specification of X , Y , and C , and the addition of the line-search, this method differs from ours in the sense that it is a scalar proximal method, applied to the piecewise linearization of the improvement function, which is indeed specific to the multiobjective case. In our case, the regularization term is a vector-valued function added to the original vector-valued function F .

Another vectorial proximal method is discussed in section 4.2 of [8]. It is a generalization of algorithms for specific instances of a vector optimization problem (VOP): a particular control approximation problem in [1] and certain location problems in [17]. In the presentation given in [8], it deals with a problem more general than VOP, namely, the *vector equilibrium problem* (VEP), which we describe next. The data are

$X = \mathbb{R}^m$, Y , and C as in this paper, and additionally a closed and convex subset M of X and a function $Q : M \times M \rightarrow Y$. The problem consists of finding $\bar{x} \in M$ such that $Q(\bar{x}, x) \notin -\text{int}(C)$ for all $x \in M$. It can be checked that VOP becomes a particular case of VEP by taking $M = \text{dom}(F)$ and $Q(u, v) = F(v) - F(u)$. VEP can be scalarized in the following way: fix $u_0 \in \text{int}(C)$, define $\psi : Y \rightarrow \mathbb{R}$ as $\psi(z) = \inf\{t \in \mathbb{R} \mid z \in tu_0 - C\}$, and $f : X \times X \rightarrow \mathbb{R}$ as $f(u, v) = \psi(Q(u, v))$. It can be seen that the solutions of the scalarized equilibrium problem for f , namely, the points $\bar{x} \in M$ such that $f(\bar{x}, x) \geq 0$ for all $x \in M$, are solutions of VEP. The authors then propose a scalar proximal method for this equilibrium problem. We refrain from making explicit the iterative formula of the method because, in the case of VOP, it ends up being just the standard scalar proximal point method, as given in (2), applied to the scalar function $\langle F(x), z_n \rangle$, ($z_n \in C^+$, $\|z_n\| = 1$), except for the fact that a more general regularization term is used: the quadratic function $\|x - x_n\|^2$ of (2) is replaced by the *Bregman distance* $D_g(x, x^n)$ between x and x^n , induced by a *Bregman function* $g : X \rightarrow \mathbb{R} \cup \{\infty\}$ (see, e.g., [6] for the definition of Bregman functions and distances). The convergence analysis in [8] is also restricted to the finite dimensional case. The fact that the method in [8] is essentially a scalar proximal method establishes a basic difference from ours. Also the inexact version presented in our paper is more suitable for numerical implementation.

The paper is organized as follows. Section 2 states formally the problem and introduces some required preliminary material. The exact version of the algorithm is analyzed in section 3. Section 4 discusses a particular instance of the algorithm for which the weak limit of the generated sequence is an efficient point, rather than a weakly efficient one, and section 5 develops the inexact version. Section 6 contains several final remarks.

2. Problem statement and notation. Recall that X is a real Hilbert space, Y is a real Banach space, and $\langle \cdot, \cdot \rangle$ denotes the scalar product in X as well as the duality scalar product between Y^* (the topological dual of Y) and Y . Any norm will be denoted by $\|\cdot\|$. F is an extended-valued map from X to $Y \cup \{+\infty_C\}$. The extended space $\bar{Y} = Y \cup \{-\infty_C, +\infty_C\}$ was introduced in [2], where a neighborhood of $+\infty_C$ is defined as a set $N \subset \bar{Y}$ containing $a + C \cup \{+\infty_C\}$ for some $a \in Y$, and its opposite $-N$ is a neighborhood of $-\infty_C$. The binary relations $\preceq_C$ and $\prec_C$ (defined in the previous section) are extended to $\bar{Y}$ by

$$\forall y \in Y, \quad -\infty_C \prec_C y \prec_C +\infty_C, \quad -\infty_C \preceq_C y \preceq_C +\infty_C.$$

Note that the embedding $Y \subset \bar{Y}$ is continuous and dense. The map F is assumed to be proper, i.e., not identically equal to $+\infty_C$.

We denote by $\text{dom}(F) = \{x \in X \mid F(x) \neq +\infty_C\}$ the effective domain of F . We extend by continuity every $z \in C^+ \setminus \{0\}$ to $\bar{Y}$, by putting $\langle \pm\infty_C, z \rangle = \pm\infty$ (see [2] and [3] for more details). For a set $T \subset \bar{Y}$, we denote its topological closure in the topological space $\bar{Y}$ by $\bar{T}$.

To a given set $T \subset \bar{Y}$, we associate the following sets:

- the *infimal set* $C\text{-INF}(T) = \{y \in \bar{T} \mid \nexists z \in T \setminus \{y\} : z \preceq_C y\}$,
- the *weakly infimal set* $C\text{-INF}_w(T) = \{y \in \bar{T} \mid \nexists z \in T : z \prec_C y\}$,
- the *properly infimal set* $C\text{-INF}_p(T) = \{y \in \bar{T} \mid \exists K \subset Y \text{ pointed closed convex cone: } C \setminus \{0\} \subset \text{int } K, y \in K\text{-INF}(T)\}$.

For a vector optimization problem of the form

$$C\text{-MINIMIZE } G(x) \text{ subject to } x \in S,$$

where $G : S \rightarrow Y \cup \{+\infty_C\}$ and $S \subset X$, the point $a \in X$ is called

- *efficient (or Pareto)* if $a \in S$ and $G(a) \in C\text{-INF}(G(S))$,
- *weakly efficient* if $a \in S$ and $G(a) \in C\text{-INF}_w(G(S))$,
- *properly efficient* if $a \in S$ and $G(a) \in C\text{-INF}_p(G(S))$.

Thus, the sets of efficient (resp., weakly efficient or properly efficient) solutions, which we will denote as $C\text{-ARGMIN}\{G(x) \mid x \in S\}$ (resp., $C\text{-ARGMIN}_w\{G(x) \mid x \in S\}$ or $C\text{-ARGMIN}_p\{G(x) \mid x \in S\}$), verify the relations

$$C\text{-ARGMIN}\{G(x) \mid x \in S\} = S \cap G^{-1}(C\text{-INF}(G(S))),$$

$$C\text{-ARGMIN}_w\{G(x) \mid x \in S\} = S \cap G^{-1}(C\text{-INF}_w(G(S))),$$

$$C\text{-ARGMIN}_p\{G(x) \mid x \in S\} = S \cap G^{-1}(C\text{-INF}_p(G(S))).$$

It is easy to verify that

$$\begin{aligned} C\text{-ARGMIN}_p\{G(x) \mid x \in S\} &\subset C\text{-ARGMIN}\{G(x) \mid x \in S\} \\ &\subset C\text{-ARGMIN}_w\{G(x) \mid x \in S\}. \end{aligned}$$

For $y \in Y$, $T \subset Y \cup \{+\infty_C\}$, $T \neq \{+\infty_C\}$, we denote $d(y, T) = \inf\{\|y - z\| \mid z \in T \cap Y\}$.

We deal with the following VOP:

$$C\text{-MINIMIZE}\{F(x) \mid x \in X\}.$$

Notice that any *constrained* vector optimization problem (CVOP)

$$C\text{-MINIMIZE } F_0(x) \text{ subject to } x \in S,$$

where $S \subset X$ is the feasible set and F_0 is a map from S to Y , is equivalent to the unconstrained extended-valued VOP with

$$F(x) = \begin{cases} F_0(x) & \text{if } x \in S, \\ +\infty_C & \text{if } x \in X \setminus S. \end{cases}$$

The CVOP and VOP are equivalent in the sense that they have the same weakly efficient solutions and the same weakly infimal set.

A map $G : S \subset X \rightarrow Y \cup \{+\infty_C\}$, where S is a convex set, is called *C-convex* if

$$\forall x, x' \in S, \lambda \in [0, 1], \quad G((1 - \lambda)x + \lambda x') \preceq_C (1 - \lambda)G(x) + \lambda G(x')$$

(with the conventions $\infty_C + \infty_C = \infty_C$, $\lambda \cdot (+\infty_C) = +\infty_C$ for each positive number λ).

A map $G : X \rightarrow Y \cup \{+\infty_C\}$ will be called *positively lower semicontinuous* if, for every $z \in C^+$, the extended-valued scalar function $x \mapsto \langle G(x), z \rangle$ is lower semicontinuous.

Throughout this paper we consider a *C-convex* VOP, which means that the map F is *C-convex*.

We recall the following scalarization result (see, e.g., [12]) known for finite-valued maps, but which is immediately generalized to extended-valued ones. Denote $C_s^+ = \{z \in Y^* : \langle y, z \rangle > 0 \text{ for all } y \in C \setminus \{0\}\}$.

THEOREM 2.1. *If $S \subset X$ is a convex set and $G : S \rightarrow Y \cup \{+\infty_C\}$ is a C -convex proper map, then*

$$C\text{-ARGMIN}_w\{G(x) \mid x \in S\} = \bigcup_{z \in C^+ \setminus \{0\}} \operatorname{argmin}\{\langle G(x), z \rangle \mid x \in S\}$$

and

$$C\text{-ARGMIN}_p\{G(x) \mid x \in S\} = \bigcup_{z \in C_s^+} \operatorname{argmin}\{\langle G(x), z \rangle \mid x \in S\}.$$

Remark 1. Note that, in Theorem 2.1, the set $\operatorname{argmin}\{\langle G(x), z \rangle \mid x \in S\}$ may be empty for some $z \in C^+ \setminus \{0\}$, as we can see from the following example:

$$S = X = Y = \mathbb{R}^2, \quad C = \mathbb{R}_+^2, \quad G(x_1, x_2) = (x_1^2 - x_2, x_2).$$

For every $z = (z_1, z_2) \in C$ with $z_1 \neq z_2$ the set $\operatorname{argmin}\{\langle G(x), z \rangle \mid x \in S\}$ is empty! Thus the set of useful scalarizing vectors is the half line $z_1 = z_2 > 0$, a set of measure zero in $\mathbb{R}^2$.

3. An exact proximal algorithm. The exact version of our proximal algorithm (to be called **ALG1**) requires two exogenous sequences: a bounded sequence of positive real numbers $\{\alpha_n\}$, and a sequence $\{e_n\} \subset \operatorname{int}(C)$ such that $\|e_n\| = 1$ for all n . The method generates a sequence $\{x_n\} \subset X$ in the following way:

Initialization: Choose $x_0 \in \operatorname{dom}(F)$.

Stopping rule: Given x_n , if $x_n \in C\text{-ARGMIN}_w\{F(x) \mid x \in X\}$, then $x_{n+p} = x_n$ for all $p \geq 1$.

Iterative step: Given x_n , if $x_n \notin C\text{-ARGMIN}_w\{F(x) \mid x \in X\}$, then take as the next iterate any $x_{n+1} \in X$ such that

$$(10) \quad x_{n+1} \in C\text{-ARGMIN}_w \left\{ F(x) + \frac{\alpha_n}{2} \|x - x_n\|^2 e_n \mid x \in \Omega_n \right\},$$

with $\Omega_n := \{x \in X \mid F(x) \preceq_C F(x_n)\}$.

We now make the following assumption on the map F and the initial iterate x_0 :

(A) The set $(F(x_0) - C) \cap F(X)$ is C -complete, meaning that for all sequences $\{a_n\} \subset X$, with $a_0 = x_0$, such that $F(a_{n+1}) \preceq_C F(a_n)$ for all $n \in \mathbb{N}$, there exists $a \in X$ such that $F(a) \preceq_C F(a_n)$ for all $n \in \mathbb{N}$.

Next, we state and prove our main convergence result.

THEOREM 3.1. *Let $F : X \rightarrow Y \cup \{+\infty_C\}$ be a proper C -convex map, positively lower semicontinuous. Under assumption (A) above, any sequence $\{x_n\}$ generated by algorithm **ALG1** is convergent, with respect to the weak topology of X , to a weakly efficient solution of the VOP.*

Proof. We divide the proof into five steps.

Step 1. Existence of iterates. x_0 is chosen in the initialization step. Assuming that the algorithm reached iteration n , we show next that an appropriate x_{n+1} exists. In view of the stopping rule, this is certainly the case if x_n is weakly efficient for VOP. Otherwise, take any $z \in C^+ \setminus \{0\}$. Since e_n belongs to $\operatorname{int}(C)$, it follows from the definition of C^+ that $\langle e_n, z \rangle > 0$. Define $\varphi_n : X \rightarrow \mathbb{R} \cup \{+\infty\}$ as

$$(11) \quad \varphi_n(x) = \langle F(x), z \rangle + I_{\Omega_n}(x) + \frac{\alpha_n}{2} \langle e_n, z \rangle \|x - x_n\|^2.$$

C -convexity of F implies convexity of $\langle F(\cdot), z \rangle$ and convexity of Ω_n . Lower semicontinuity of F implies closedness of Ω_n , so that $\langle F(\cdot), z \rangle + I_{\Omega_n}$ is convex and lower semicontinuous. Since $\langle e_n, z \rangle > 0$, φ_n is strongly convex, and the existence of minimizers of φ_n results from the standard argument for existence of iterates of the scalar-valued proximal method (see [20]): the subdifferential of φ_n is maximal monotone and strongly monotone, and hence onto, by virtue of Minty's theorem. Thus, the subdifferential has some zero, which is a minimizer of φ_n . In view of Theorem 2.1, such a minimizer satisfies (10) and can be taken as x_{n+1} .

Step 2. Fejér convergence to the set of lower bounds of the initial section. If the stopping rule applies at some iteration, then the sequence remains constant thereafter, and thus it is strongly convergent to the stopping iterate, which is a weakly efficient solution of VOP. We assume from now on that the stopping rule never applies. Define $E \subset X$ as

$$E = \{x \in X : F(x) \preceq_C F(x_n) \ \forall n \in \mathbb{N}\}.$$

By assumption (A), E is nonempty. Now, since x_{n+1} solves the vector optimization problem in (10), by Theorem 2.1 there exists $z_n \in C^+ \setminus \{0\}$ such that x_{n+1} solves the problem

$$(12) \quad \min \eta_n(x)$$

$$(13) \quad \text{subject to } x \in \Omega_n,$$

where $\eta_n : X \rightarrow \mathbb{R} \cup \{+\infty\}$ is defined by

$$\eta_n(x) = \langle F(x), z_n \rangle + \frac{\alpha_n}{2} \langle e_n, z_n \rangle \|x - x_n\|^2.$$

Since the solution of (12)–(13) is not altered through multiplication of z_n by positive scalars, we can assume without loss of generality that $\|z_n\| = 1$ for all $n \in \mathbb{N}$. Note that by definition, $\Omega_n \subset \text{dom}(\eta_n) = \text{dom}(F)$, so that $\emptyset \neq \text{dom}(I_{\Omega_n}) \subset \text{dom}(\eta_n)$. According to [18, Theorem 3.23], it follows that x_{n+1} satisfies the first order optimality conditions for problem (12)–(13); i.e., there exists $u_n \in X$ such that

$$(14) \quad u_n \in \partial \eta_n(x_{n+1})$$

and

$$(15) \quad 0 \leq \langle u_n, x - x_{n+1} \rangle$$

for all $x \in \Omega_n$. Now define $\psi_n : X \rightarrow \mathbb{R} \cup \{+\infty\}$ as

$$(16) \quad \psi_n(x) = \langle F(x), z_n \rangle.$$

In view of (11) and (14), we have

$$(17) \quad u_n = v_n + \alpha_n \langle e_n, z_n \rangle (x_{n+1} - x_n)$$

for some

$$(18) \quad v_n \in \partial \psi_n(x_{n+1}).$$

Now take $x^* \in E$. By definition of E , x^* belongs to Ω_n for all $n \in \mathbb{N}$. Combining (15) with $x = x^*$ and (17), we have

$$(19) \quad \begin{aligned} 0 &\leq \langle v_n, x^* - x_{n+1} \rangle + \alpha_n \langle e_n, z_n \rangle \langle x_{n+1} - x_n, x^* - x_{n+1} \rangle \\ &\leq \langle F(x^*) - F(x_{n+1}), z_n \rangle + \alpha_n \langle e_n, z_n \rangle \langle x_{n+1} - x_n, x^* - x_{n+1} \rangle \\ &\leq \alpha_n \langle e_n, z_n \rangle \langle x_{n+1} - x_n, x^* - x_{n+1} \rangle, \end{aligned}$$

using (16) and (18) in the second inequality and the facts that x^* belongs to E and z_n belongs to $C^+ \setminus \{0\}$ in the third; it is clear that $F(x^*) - F(x_{n+1}) \preceq_C 0$ and therefore $\langle F(x^*) - F(x_{n+1}), z_n \rangle \leq 0$.

Now define $\gamma_n = \alpha_n \langle e_n, z_n \rangle$. Note that $\gamma_n > 0$, because $\alpha_n > 0$, e_n belongs to $\text{int}(C)$, and z_n belongs to $C^+ \setminus \{0\}$. By (19),

$$0 \leq \gamma_n \langle x_{n+1} - x_n, x^* - x_{n+1} \rangle = \gamma_n \left(\|x_n - x^*\|^2 - \|x_{n+1} - x^*\|^2 - \|x_n - x_{n+1}\|^2 \right),$$

implying, since $\gamma_n > 0$,

$$(20) \quad \|x_{n+1} - x^*\|^2 \leq \|x_n - x^*\|^2 - \|x_n - x_{n+1}\|^2.$$

Step 3. Boundedness of the sequence and proximity of consecutive iterates. This step is identical to the corresponding one in the convergence proof of the scalar-valued method. We present it for the sake of completeness. It follows from (20) that, for all $x^* \in E$, $\{\|x_n - x^*\|^2\}$ is a nonnegative and nonincreasing sequence, and hence is convergent. Rewriting (20) as

$$(21) \quad \|x_n - x_{n+1}\|^2 \leq \|x_n - x^*\|^2 - \|x_{n+1} - x^*\|^2,$$

and observing that the right-hand side of (21) converges to 0 as $n \rightarrow \infty$ because $\{\|x_n - x^*\|^2\}$ is convergent, we conclude that

$$(22) \quad \lim_{n \rightarrow \infty} \|x_n - x_{n+1}\| = 0.$$

It follows from (20) that

$$(23) \quad \|x_n - x^*\| \leq \|x_0 - x^*\|$$

for all $x^* \in E$. Since E is nonempty, we conclude that $\{x_n\}$ is bounded.

Step 4. Optimality of the weak cluster points of $\{x_n\}$. Since $\{x_n\}$ is bounded, it has weak cluster points. We will prove next that all of them are weakly efficient solutions of VOP. Let $\hat{x}$ be a weak cluster point of $\{x_n\}$ and let $\{x_{k_n}\}$ be a sequence weakly convergent to it.

Define $\psi_z : X \rightarrow \mathbb{R}$ as $\psi_z(x) = \langle F(x), z \rangle$. We claim that

$$(24) \quad \psi_z(\hat{x}) \leq \psi_z(x_n)$$

for all $z \in C^+$ and all $n \in \mathbb{N}$. Since F is positively lower semicontinuous and C -convex, ψ_z is lower semicontinuous and convex so that $\psi_z(\hat{x}) \leq \lim_{n \rightarrow \infty} \psi_z(x_{k_n})$. Since $F(x_{n+1}) \preceq_C F(x_n)$ for all n , we have $\psi_z(x_{n+1}) \leq \psi_z(x_n)$ for all n so that $\lim_{n \rightarrow \infty} \psi_z(x_{k_n}) = \inf\{\psi_z(x_n)\}$. The claim follows from the last two statements. It follows easily from (24) that

$$(25) \quad F(\hat{x}) \preceq_C F(x_n)$$

for all $n \in \mathbb{N}$.

Assume that $\hat{x}$ is not weakly efficient for VOP, i.e., that there exists $\bar{x} \in X$ such that $F(\bar{x}) \prec_C F(\hat{x})$. Take z_n as chosen before (12). Since $\|z_n\| = 1$ for all n , there exists, by virtue of the Bourbaki–Alaoglu theorem, a weak* cluster point of $\{z_n\}$, say $\bar{z}$, which is a weak* limit of the subsequence $\{z_{j_n}\}$. We claim now that C^+ is weakly* closed. Observe that $C^+ = \bigcap_{y \in C} \{z \in Y^* : \langle y, z \rangle \geq 0\}$. Since the linear forms $z \mapsto \langle y, z \rangle$ are weakly* continuous for all $y \in Y$, we have written C^+ as an intersection of weakly* closed sets, establishing the claim. It follows that $\bar{z}$ belongs to C^+ . Thus

$$\begin{aligned} \langle F(\bar{x}) - F(\hat{x}), z_{j_n} \rangle &\geq \langle F(\bar{x}) - F(x_{j_n+1}), z_{j_n} \rangle = \psi_{j_n}(\bar{x}) - \psi_{j_n}(x_{j_n+1}) \\ &\geq \langle v_{j_n}, \bar{x} - x_{j_n+1} \rangle = \langle u_{j_n}, \bar{x} - x_{j_n+1} \rangle - \gamma_n \langle x_{j_n+1} - x_{j_n}, \bar{x} - x_{j_n+1} \rangle \\ (26) \quad &\geq -\gamma_n \langle x_{j_n+1} - x_{j_n}, \bar{x} - x_{j_n+1} \rangle \geq -\gamma_n \|x_{j_n+1} - x_{j_n}\| \|\bar{x} - x_{j_n+1}\|, \end{aligned}$$

using (24) in the first inequality, (18) in the second, (17) in the second, and (15) in the third, together with the fact that $\bar{x}$ belongs to Ω_n for all $n \in \mathbb{N}$, because $F(\bar{x}) \prec_C F(\hat{x}) \preceq_C F(x_n)$, in view of (25).

Now we will take limits in the extreme expressions of (26). Regarding the rightmost one, since $\{\alpha_n\}$ is bounded, and $\|z_n\| = \|e_n\| = 1$, we have that $\{\gamma_n\}$ is bounded. Note that $\{x_n\}$ is bounded so that $\{\|\bar{x} - x_{j_n+1}\|\}$ is also bounded. Finally, $\lim_{n \rightarrow \infty} \|x_{j_n} - x_{j_n+1}\| = 0$ by (22). We conclude that the limit of the rightmost expression in (26) as $n \rightarrow \infty$ vanishes and, therefore, since $\bar{z}$ is the weak* limit of $\{z_{j_n}\}$, we get from (26) that

$$(27) \quad \langle \bar{F}(\bar{x}) - F(\hat{x}), \bar{z} \rangle \geq 0.$$

Now we claim that $\bar{z} \neq 0$. Take $e \in \text{int}(C)$. It follows from Lemma 2.2 of [2] that $\langle e, z_n \rangle \geq d(e, Y \setminus C) > 0$ for all $n \in \mathbb{N}$. Since $\bar{z}$ is the weak* limit of $\{z_{j_n}\}$, we conclude that $\langle e, \bar{z} \rangle > 0$, establishing the claim. Since $\bar{z} \neq 0$, it is clear that (27) contradicts the fact that $\bar{z}$ belongs to C^+ and the assumption that $F(\bar{x}) \prec_C F(\hat{x})$. Thus, such an assumption is false, and $\hat{x}$ is indeed weakly efficient for VOP.

Step 5. Uniqueness of the weak cluster point of $\{x_n\}$. This part of the proof, presented also for the sake of completeness, is again identical to the scalar-valued case, as in [20], and it uses Brézis's uniqueness argument. Suppose that both $\hat{x}$ and $\tilde{x}$ are weak cluster points of $\{x_n\}$. With the same argument as in (25), both $\hat{x}$ and $\tilde{x}$ belong to E so that both $\{\|\hat{x} - x_n\|\}$ and $\{\|\tilde{x} - x_n\|\}$ converge, as shown at the beginning of Step 3; i.e., there exist $\hat{\beta}, \tilde{\beta} \in \mathbb{R}$ such that

$$(28) \quad \lim_{n \rightarrow \infty} \|x_n - \hat{x}\| = \hat{\beta}, \quad \lim_{n \rightarrow \infty} \|x_n - \tilde{x}\| = \tilde{\beta}.$$

Since

$$\|x_n - \hat{x}\|^2 = \|x_n - \tilde{x}\|^2 + 2\langle x_n - \tilde{x}, \tilde{x} - \hat{x} \rangle + \|\tilde{x} - \hat{x}\|^2,$$

we conclude from (28) that

$$(29) \quad \lim_{n \rightarrow \infty} \langle x_n - \tilde{x}, \tilde{x} - \hat{x} \rangle = \frac{1}{2} (\hat{\beta}^2 - \tilde{\beta}^2 - \|\tilde{x} - \hat{x}\|^2).$$

The left-hand side of (29) vanishes, because $\tilde{x}$ is a weak cluster point of $\{x_n\}$, so that

$$(30) \quad \hat{\beta}^2 - \tilde{\beta}^2 = \|\tilde{x} - \hat{x}\|^2.$$

Reversing the roles of $\hat{x}$ and $\tilde{x}$, a similar reasoning leads to $\tilde{\beta}^2 - \hat{\beta}^2 = \|\tilde{x} - \hat{x}\|^2$, which, combined with (30), gives $\|\tilde{x} - \hat{x}\| = 0$, i.e., $\tilde{x} = \hat{x}$, establishing the uniqueness of the

weak cluster points of $\{x_n\}$. It follows that $\{x_n\}$ is weakly convergent to a weakly efficient solution of VOP and the proof is complete. $\square$

The next remark allows for the possibility of checking the stopping rule by solving a scalar program.

Remark 2. Let $e \in \text{int}(C)$ be fixed, and let x_n be the n th iterate of a sequence generated by **ALG1**. Then $x_n \in C\text{-ARGMIN}_w\{F(x) \mid x \in X\}$ if and only if

$$\inf\{t \in \mathbb{R} \mid \exists x \in X : F(x) \preceq_C F(x_n) + te\} = 0.$$

The proof of the above remark is obvious when using the definition of a weakly efficient solution and the continuity at 0 of the map $t \mapsto F(x_n) - te - F(x)$ for each fixed $x \in \text{dom}(F)$.

The next result is related to the possibility of replacing the stopping rule of **ALG1** by the following, which is easier to check: after computing x_{n+1} , the algorithm stops if $x_{n+1} = x_n$ (i.e., we set $x_{n+p} = x_n$ for all $p \geq 0$ in such a case). Under this new rule, if the algorithm stops at iteration n , then x_n is a weakly Pareto solution of the problem. On the other hand, this rule might fail to recognize weakly Pareto solutions: the sequence might hit such a solution at iteration n and proceed with the iterative process, eventually generating a whole sequence of different weakly Pareto solutions. However, this situation is not possible if x_n is a Pareto solution (not just a weakly Pareto one), in which case, as shown in the next proposition, the following iterate x_{n+1} coincides with x_n .

PROPOSITION 3.2. *Let x_n be the n th iterate of a sequence generated by Algorithm **ALG1**.*

1. *If $x_n \in C\text{-ARGMIN}_w\{F(x) + \frac{\alpha_n}{2}\|x - x_n\|^2 e_n \mid x \in \Omega_n\}$, then $x_n \in C\text{-ARGMIN}_w\{F(x) \mid x \in X\}$.*
2. *If $x_n \in C\text{-ARGMIN}_w\{F(x) \mid x \in X\}$, then $C\text{-ARGMIN}_w\{F(x) + \frac{\alpha_n}{2}\|x - x_n\|^2 e_n \mid x \in \Omega_n\} = \{x_n\}$.*

Proof. 1. Suppose that $x_n \in C\text{-ARGMIN}_w\{F(x) + \frac{\alpha_n}{2}\|x - x_n\|^2 e_n \mid x \in \Omega_n\}$, but x_n does not belong to $C\text{-ARGMIN}_w\{F(x) \mid x \in X\}$. Then, there exists some $\bar{x} \in X$ such that $F(\bar{x}) \prec_C F(x_n)$, i.e., there exists some $c \in \text{int}(C)$ such that $F(\bar{x}) = F(x_n) - c$. For $t \in [0, 1]$, define $x_t = (1 - t)\bar{x} + tx_n$. By the C -convexity of F ,

$$(31) \quad F(x_t) \preceq_C (1 - t)F(\bar{x}) + tF(x_n) = F(x_n) - (1 - t)c.$$

Since x_n solves the vector optimization problem in (10) (with x_n instead of x_{n+1}), by Theorem 2.1 there exists $z_n \in C^+ \setminus \{0\}$ such that

$$(32) \quad \varphi_n(x_n) = \min_{x \in \Omega_n} \varphi_n(x) \leq \varphi_n(x_t),$$

where φ_n is as in (11), with z_n substituting for z . In view of (31) and (32), and calling as before $\gamma_n = \alpha_n \langle e_n, z_n \rangle$, we have

$$(33) \quad \begin{aligned} \langle F(x_n), z_n \rangle &\leq \langle F(x_t), z_n \rangle + \frac{\gamma_n}{2} \|x_t - x_n\|^2 = \langle F(x_t), z_n \rangle + (1 - t)^2 \frac{\gamma_n}{2} \|\bar{x} - x_n\|^2 \\ &\leq \langle F(x_n), z_n \rangle - (1 - t)\langle c, z_n \rangle + (1 - t)^2 \frac{\gamma_n}{2} \|\bar{x} - x_n\|^2. \end{aligned}$$

It follows from (33) that

$$(34) \quad \langle c, z_n \rangle \leq (1 - t) \frac{\gamma_n}{2} \|\bar{x} - x_n\|^2$$

for all $t \in [0, 1)$. Letting $t \rightarrow 1^-$, we obtain from (34) that $\langle c, z_n \rangle \leq 0$, which is a contradiction.

2. Suppose now that $x_n \in C\text{-ARGMIN}\{F(x) \mid x \in X\}$, i.e., that x_n is a Pareto solution of the original problem. It is obvious that $x_n \in C\text{-ARGMIN}_w\{F(x) + \frac{\alpha_n}{2}\|x - x_n\|^2 e_n \mid x \in \Omega_n\}$. Let $\bar{x} \in C\text{-ARGMIN}_w\{F(x) + \frac{\alpha_n}{2}\|x - x_n\|^2 e_n \mid x \in \Omega_n\}$. Since $\bar{x} \in \Omega_n$ and x_n is a Pareto solution, we must have $F(\bar{x}) = F(x_n)$. On the other hand, by Theorem 2.1, there exists $z \in C^+ \setminus \{0\}$ such that

$$\begin{aligned} \langle F(\bar{x}), z \rangle + \frac{\alpha_n}{2} \langle e_n, z \rangle \|\bar{x} - x_n\|^2 &= \min_{x \in \Omega_n} \left\{ \langle F(x), z \rangle + \frac{\alpha_n}{2} \langle e_n, z \rangle \|x - x_n\|^2 \right\} \\ &\leq \langle F(x_n), z \rangle = \langle F(\bar{x}), z \rangle. \end{aligned}$$

It follows immediately that $\bar{x} = x_n$. $\square$

We discuss next a few issues related to assumption (A), i.e., the C -completeness hypothesis. First, note that, in the scalar minimization problems ($Y = \mathbb{R}$, $C = [0, +\infty)$), when the set of minimizers is not empty, every section $(F(x_0) - C) \cap F(X)$ is C -complete. We also have the following observation (see [12, Lemma 3.5, p. 47]).

Remark 3. Each of the following is a sufficient condition for assumption (A):

1. The set $(F(x_0) - C) \cap F(X)$ is compact.
2. The set $(F(x_0) - C) \cap F(X)$ is weakly compact.
3. The set $(F(x_0) - C) \cap F(X)$ is closed, has a lower bound (i.e., there exists some $a \in X$ such that, for all $x \in X$ verifying $F(x) \preceq_C F(x_0)$, we have $F(a) \preceq_C F(x)$) and the cone C has the Daniell property (i.e., any decreasing net having a lower bound converges to its infimum).

In practical situations, assumption (A) could be difficult to verify. Thus, it is interesting to establish some convergence results, albeit weaker than Theorem 3.1, for situations in which this assumption fails to hold. This is the content of the following two propositions.

PROPOSITION 3.3. *Let $F : X \rightarrow Y \cup \{+\infty_C\}$ be a proper C -convex map, positively lower semicontinuous. If the sequence $\{x_n\}$ generated by algorithm **ALG1** (without assuming that (A) holds) has a weak cluster point, then it is weakly convergent to a weakly efficient solution of the VOP.*

Proof. By assumption, there exists a weakly convergent subsequence $\{x_{k_n}\}$ of $\{x_n\}$ whose weak limit is some $\hat{x} \in X$. Since the sequence $\{F(x_n)\}$ is C -nonincreasing, we have that, for all $z \in C^+$, the sequence $\{\langle F(x_{k_n}), z \rangle\}$ is nonincreasing and, since the scalar convex function $\langle F(\cdot), z \rangle$ is weakly lower semicontinuous, we obtain that $\langle F(\hat{x}), z \rangle$ is a lower bound for $\{\langle F(x_{k_n}), z \rangle\}$. Since this result holds for all $z \in C^+$, we get that $F(\hat{x}) \preceq_C F(x_{k_n})$ for all $n \in \mathbb{N}$, implying that $F(\hat{x}) \preceq_C F(x_n)$. From this point on, the argument continues as in the proof of Theorem 3.1 after (25). $\square$

PROPOSITION 3.4. *Let $F : X \rightarrow Y \cup \{+\infty_C\}$ be a proper C -convex map, positively lower semicontinuous. Assume that $\dim Y < \infty$ and that $\text{int}(C^+) \neq \emptyset$. Let $\{x_n\}$ be a sequence generated by Algorithm **ALG1**. Suppose that $\{x_n\}$ is unbounded and that the set of finite cluster points of $\{F(x_n)\}$ is nonempty. Then the sequence $\{F(x_n)\}$ converges to an element of the weakly infimal set $\text{INF}_w(F(X))$ of the VOP.*

Proof. Note first that our assumption implies that $\{x_n\}$ is not weakly convergent.

Consider a subsequence $\{F(x_{k_n})\}$ convergent to some $\bar{y} \in Y$. We want to prove that the whole sequence $\{F(x_n)\}$ converges to $\bar{y}$. Note that $\bar{y}$ belongs to $\overline{F(X)}$. For every $z \in C^+$, the sequence $\{\langle F(x_n), z \rangle\}$ is convergent to $\langle \bar{y}, z \rangle$ because it is nonincreasing and admits a subsequence converging to $\langle \bar{y}, z \rangle$. On the other hand, since $\{F(x_n)\}$ is C -nonincreasing, and C is closed, it follows that $\bar{y}$ is a C -lower

bound for this sequence. Next we take some $z \in \text{int}(C^+)$, and, using Lemma 2.7 of [3], we get

$$(35) \quad \langle F(x_n) - \bar{y}, z \rangle \geq d(z, Y^* \setminus C^+) \|F(x_n) - \bar{y}\|$$

for all $n \in \mathbb{N}$. Since the left-hand side of (35) goes to zero as $n \rightarrow \infty$, and $d(z, Y^* \setminus C^+) > 0$, we obtain that $\{F(x_n)\}$ is convergent to $\bar{y}$.

If $\bar{y} \notin \text{INF}_w(F(X))$, then there exists $\bar{x} \in X$ with $F(\bar{x}) \prec_C \bar{y}$. Using the same argument as in the proof of Proposition 3.3, we get that $\bar{y} \preceq_C F(x_n)$ for all $n \in \mathbb{N}$. Then, we can establish weak convergence of $\{x_n\}$ to a weakly efficient solution of VOP, with the same argument as in the proof of Theorem 3.1 from (25) on. This contradicts the assumption that $\{x_n\}$ is unbounded, establishing the result. $\square$

4. Convergence to efficient (rather than weakly efficient) solutions. In real-life applications it is often the case that only efficient solutions (instead of weakly efficient ones) are of interest (see, e.g., section 2.3 in [11]). The fact that our algorithm is only guaranteed to converge to weakly efficient solutions of VOP could be seen as a weakness of the method, but we can overcome this drawback by taking advantage of the wide choice that **ALG1** offers in its iterative step, where any weakly efficient point of the regularized map in (10) is admissible as the next iterate x_{n+1} . We show next that, under a not too demanding assumption on C (which entails existence of efficient solutions), a judicious choice of $x_{n+1} \in C\text{-ARGMIN}_w\{F(x) + \alpha_n/2 \|x - x_n\|^2 \mid x \in \Omega_n\}$ guarantees that the generated sequence $\{x_n\}$ converges indeed to an efficient solution of VOP. We present next some required notation and the aforementioned assumption on C .

Let us consider, for any real $\delta \geq 0$, the set

$$(36) \quad K_\delta = \{z \in Y^* : \langle x, z \rangle \geq \delta \|x\| \|z\| \ \forall x \in C\}.$$

Note that $K_0 = C^+$, $K_0 \setminus \{0\} = C_s^+$, and $0 \leq \delta_1 < \delta_2 \implies K_{\delta_2} \subset K_{\delta_1}$. In particular $K_\delta \setminus \{0\} \subset C_s^+$ for any $\delta \geq 0$.

The required assumption is the following:

(H1) There exists $\delta > 0$ such that $K_\delta \neq \emptyset$.

Remark 4. It follows from Lemma 2.7 in [3] that, for each $\delta > 0$,

$$\left\{ z \in Y^* \setminus \{0\} \mid d\left(\frac{z}{\|z\|}, Y^* \setminus C^+\right) \geq \delta \right\} \subset K_\delta.$$

Thus, if the interior of C^+ is nonempty, then K_{δ_0} is nonempty for some $\delta_0 > 0$, and therefore $K_\delta \neq \emptyset$ for all $\delta \in [0, \delta_0]$.

Proposition 1.10 in [12] implies that $\text{int } C^+ \neq \emptyset$ whenever C has a *convex weakly compact base*; i.e., there exists a convex weakly compact set $B \subset C$ such that for any $c \in C \setminus \{0\}$, there are unique $b \in B$, $t > 0$ such that $c = tb$.

We consider now a specific instance of **ALG1**, designed for guaranteeing convergence of $\{x_n\}$ to an efficient solution of VOP, assuming that the cone C is such that (H1) holds. First, we modify the stopping rule in the obvious way, namely, we only stop at iteration n if x_n belongs to $C\text{-ARGMIN}\{F(x) \mid x \in X\}$. For the iterative step, we specialize to some degree the choice of x_{n+1} : we use (H1) in order to fix a value of δ such that $K_\delta \neq \emptyset$, we select exogenously some $z_n \in K_\delta$, and finally we take x_{n+1} as

$$(37) \quad x_{n+1} = \text{argmin}_{x \in \Omega_n} \left\langle F(x) + \frac{\alpha_n}{2} \|x - x_n\|^2 e_{n, z_n} \right\rangle.$$

We remark that (37) is a legitimate option within the framework of **ALG1**: by Theorem 2.1, if z_n belongs to C_s^+ , then x_{n+1} , as defined by (37), is properly efficient, and hence weakly efficient, for the problem C -MINIMIZE $\{F(x) + \alpha_n/2 \|x - x_n\|^2 \mid x \in \Omega_n\}$, as required by (10). In order to ensure convergence to an efficient solution of VOP, we demand a little more of z_n : it must belong to K_δ , rather than just to C_s^+ .

We establish next the convergence properties of **ALG1** under the specific form of the iterative step given by (37).

THEOREM 4.1. *Let $F : X \rightarrow Y \cup \{+\infty_C\}$ be a proper C -convex map, positively lower semicontinuous. Under assumptions (A) and (H1), any sequence $\{x_n\}$ generated by algorithm **ALG1** with x_{n+1} selected according to (37) is convergent, with respect to the weak topology of X , to an efficient solution of problem VOP.*

Proof. We follow the proof of Theorem 3.1. Steps 1, 2, and 3 hold without changes. We proceed to establish that the weak limit points of $\{x_n\}$ are efficient for VOP.

The result is obvious if the stopping rule applies at some finite iteration. Assume that this is not the case, and let $\hat{x}$ be a weak limit point of $\{x_n\}$. If $\hat{x}$ is not efficient for VOP, there exists $\bar{x} \in X$ such that

$$(38) \quad F(\bar{x}) \preceq_C F(\hat{x}), \quad F(\bar{x}) \neq F(\hat{x}).$$

Since $\|z_n\| = 1$ for all n , there exists, by virtue of the Bourbaki–Alaoglu theorem, a weak* cluster point of $\{z_n\}$, say $\bar{z}$, which is a weak* limit of the subsequence $\{z_{j_n}\}$. Since C^+ is weakly* closed (see the proof of Theorem 3.1) we have that $\bar{z} \in C^+$. Note that, in view of (25), $\bar{x} \in \cap_{n \geq 0} \Omega_n$ so that (26) and (27) hold with the same arguments as in the proof of Theorem 3.1.

Now we claim that $\bar{z} \in C_s^+$, and this is the only point where we invoke the specific choice of x_{n+1} given by (37), allowed by (H1). Take $e \in C$. Since z_n belongs to K_δ for all n , we have $\langle e, z_n \rangle \geq \delta \|e\|$. Note that $z \mapsto \langle e, z \rangle$ is continuous in the weak* topology of X so that, taking the limit as $n \rightarrow +\infty$, we obtain that $\langle e, \bar{z} \rangle \geq \delta \|e\|$, establishing the claim.

Observe that $\bar{z} \in C_s^+$ and (38) contradict (27). It follows that $\hat{x}$ is efficient for VOP.

The proof of Step 5 is identical to the corresponding proof in Theorem 3.1. $\square$

5. An inexact proximal algorithm. Now we present the inexact version of our algorithm, to be called **ALG2**. It requires an exogenous relative error tolerance $\sigma \in [0, 1)$ and three exogenous sequences: the ones required by **ALG1**, i.e., a bounded sequence of positive real numbers $\{\alpha_n\}$, and a sequence $\{e_n\} \subset \text{int}(C)$ such that $\|e_n\| = 1$ for all n , and now also a sequence $\{z_n\} \subset C^+$ such that $\|z_n\| = 1$ for all $n \in \mathbb{N}$. The method generates a sequence $\{x_n\} \subset X$ in the following way:

Initialization: Choose $x_0 \in \text{dom}(F)$.

Stopping rule: Given x_n , if $x_n \in C\text{-ARGMIN}_w\{F(x) \mid x \in X\}$, then $x_{n+p} = x_n$ for all $p \geq 1$.

Iterative step: Given x_n , if $x_n \notin C\text{-ARGMIN}_w\{F(x) \mid x \in X\}$, then let $\Omega_n = \{x \in X : F(x) \preceq_C F(x_n)\}$ and define $f_n(x) = \langle F(x), z_n \rangle + I_{\Omega_n}(x)$. Take as x_{n+1} any vector $x \in X$ such that there exists $\varepsilon_n \in \mathbb{R}_+$ satisfying

$$(39) \quad 0 \in \partial_{\varepsilon_n} f_n(x) + \alpha_n \langle e_n, z_n \rangle (x - x_n),$$

$$(40) \quad \varepsilon_n \leq \sigma \frac{\alpha_n}{2} \langle e_n, z_n \rangle \|x - x_n\|^2.$$

Remark 5. We mention now that the difference between the presentation of the iterative steps in **ALG1** and **ALG2** is not substantial. The subproblems of **ALG1** require finding weakly efficient points for a regularized vector-valued function; the subproblems of **ALG2** demand ε -subgradients of a scalar-valued one, which could be seen, in light of Theorem 2.1, as approximate weakly efficient points of the vector-valued one. We chose the presentation above in order to avoid the possibly cumbersome tasks of defining, for vector-valued maps, either some kind of approximate weakly efficient points or some kind of ε -subgradients. As discussed in Remark 9 of section 6, the subproblems of **ALG2**, despite its scalar-valued presentation, in some cases are more suitably solved by algorithms specifically devised for vector optimization.

The convergence result for **ALG2** is the following.

THEOREM 5.1. *Let $F : X \rightarrow Y \cup \{+\infty_C\}$ be a proper C -convex map, positively lower semicontinuous. Under assumption (A), introduced just before the statement of Theorem 3.1, any sequence $\{x_n\}$ generated by algorithm **ALG2** is convergent, with respect to the weak topology of X , to a weakly efficient point for VOP.*

Proof. *Step 1. Existence of the iterates.* According to our definition given by (5), we have $\partial f(x) = \partial_0 f(x) \subset \partial f_\varepsilon(x)$ for all convex $f : X \rightarrow \mathbb{R} \cup \{+\infty\}$, all $x \in X$, and all $\varepsilon \in \mathbb{R}_+$. Thus, assuming some x_n exists, the strongly convex function

$$\bar{f}_n(x) := f_n(x) + \frac{\alpha_n}{2} \langle e_n, z_n \rangle \|x - x_n\|^2$$

has a strongly convex subdifferential, which has a zero by virtue of Minty's theorem. Such a zero satisfies the inclusion in (39), with $\varepsilon_n = 0$, which trivially satisfies (40).

Step 2. Fejér convergence to the set of lower bounds of the initial section. Again let $\gamma_n = \alpha_n \langle e_n, z_n \rangle$. In view of (39),

$$(41) \quad \gamma_n(x_n - x_{n+1}) \in \partial_{\varepsilon_n} f_n(x_{n+1}).$$

Take any $x^* \in E$. Then, in view of (41) and the definition of ∂_{ε_n} , we have

$$(42) \quad -\varepsilon_n + \gamma \langle x_n - x_{n+1}, x^* - x_{n+1} \rangle \leq f_n(x^*) - f_n(x_{n+1}).$$

Note that, since x^* belongs to $E \subset \Omega_n$, we have that $x^* \in \Omega_n$, so that

$$(43) \quad f_n(x^*) = \langle F(x^*), z_n \rangle,$$

and also

$$(44) \quad \langle F(x^*) - F(x_{n+1}), z_n \rangle \leq 0,$$

because $F(x^*) \preceq_C F(x_{n+1})$ and $z_n \in C^+$. Note also that, in view of the definitions of f_n and I_{Ω_n} , we have $\partial_{\varepsilon_n} f_n(x) = \emptyset$ for all $x \notin \Omega_n$. Thus x_{n+1} belongs to Ω_n , i.e.,

$$(45) \quad f_n(x_{n+1}) = \langle F(x_{n+1}), z_n \rangle.$$

Combining (42)–(45), we get

$$(46) \quad -\varepsilon_n + \gamma_n \langle x_n - x_{n+1}, x^* - x_{n+1} \rangle \leq \langle F(x^*) - F(x_{n+1}), z_n \rangle \leq 0.$$

It follows from (46) that

$$(47) \quad 0 \leq \varepsilon_n + \gamma_n \langle x_{n+1} - x_n, x^* - x_{n+1} \rangle = \varepsilon_n + \gamma_n (\|x_n - x^*\|^2 - \|x_{n+1} - x^*\|^2 - \|x_n - x_{n+1}\|^2).$$

Combining (47) and (40), we get

$$(48) \quad 0 \leq \gamma_n \left(\|x_n - x^*\|^2 - \|x_{n+1} - x^*\|^2 - (1 - \sigma) \|x_n - x_{n+1}\|^2 \right).$$

Since $0 \leq \sigma < 1$ and $\gamma_n > 0$, it follows from (48) that $\|x_{n+1} - x^*\| \leq \|x_n - x^*\|$, so that $\{\|x_n - x^*\|\}$ is nonincreasing and nonnegative, hence convergent, and also that

$$(49) \quad (1 - \sigma) \|x_n - x_{n+1}\|^2 \leq \|x_n - x^*\|^2 - \|x_{n+1} - x^*\|^2.$$

Since the right-hand side of (49) converges to 0 as $n \rightarrow \infty$ and $1 - \sigma > 0$, we conclude from (49) that $\lim_{n \rightarrow \infty} \|x_n - x_{n+1}\| = 0$. Note that we have obtained in this step exactly the same results as in Step 2 in the proof of Theorem 3.1.

Steps 3, 4, and 5. Boundedness of $\{x_n\}$ and optimality and uniqueness of its weak cluster point. The argument in Steps 3, 4, and 5 in the proof of Theorem 3.1 holds “verbatim” for this case. $\square$

The results in Propositions 3.3 and 3.4 can be extended “mutatis mutandis” to the sequences generated by **ALG2**. We refrain from doing this for the sake of conciseness.

6. Final remarks.

Remark 6. Under (H1), one can suitably adjust **ALG2** so that the generated sequence converges to an efficient, rather than a weakly efficient, solution of VOP, exactly as we did with **ALG1** in section 4. We omit the details.

Remark 7. The proximal point method is indeed a conceptual scheme rather than an implementable algorithm, in the sense that it merely transforms a given problem into a sequence of better behaved subproblems. Thus, the performance of the method depends essentially on the algorithm used to solve the subproblems. In this situation, it makes little sense to compare the proximal method with other methods for finding efficient points (weakly or not) in terms of computational efficiency, unless a specific procedure is chosen for solving the subproblems. In this paper, we refrain from discussing algorithms for the solution of the subproblems, and hence we skip discussions of implementation issues or comparisons with alternative approaches. Nevertheless, some hints in this direction are commented upon in Remark 9 below.

Remark 8. One could fix from the onset some $z \in C^+$, define $\psi : X \rightarrow \mathbb{R} \cup \{+\infty\}$ as $\psi(x) = \langle F(x), z \rangle$, and proceed to apply the proximal point method for scalar-valued functions to the minimization of ψ . Under C -convexity of F and assumption (A), ψ has minimizers and, taking into account the well-known convergence results for the scalar proximal point method (see, e.g., [20] for the exact method as in (1), and see [21] for the inexact method in (3)–(4)) the sequence generated by the proximal method would be weakly convergent to a minimizer of ψ , which would be weakly efficient for VOP, in view of Theorem 2.1. This is the well-known *scalarization* approach for solving VOP. On the other hand, as can be observed in the proof of Theorem 3.1 and in the statement of algorithm **ALG2**, our approach can be seen as choosing in each step a different scalarization through the selection of some $z_n \in C^+$. Thus, the novelty of the method in this paper with respect to the above-mentioned approach lies in the additional flexibility, allowing a different scalarization in each iteration. As mentioned in Remark 7, this paper does not deal with the computational performance of the method, but we intend to study, in the future, good strategies for the choice of the scalarization vectors z_n , from the point of view of actual implementations. In view of Remark 7 above, such choices will very likely be linked to the method chosen for solving the subproblems.

Remark 9. We mentioned in Remark 8 that our method can be seen as a procedure which chooses at each iteration a certain scalarization and minimizes a regularized scalar-valued function. We emphasize now that this is just a way of visualizing the procedure, but it does not entail, even in the case of the inexact version **ALG2**, that the subproblems should actually be treated as scalar minimization problems. In fact, it could be inconvenient to do so, as we discuss next. The presence of the term I_{Ω_n} in the definition of f_n demands that $F(x_{n+1}) \preceq_C F(x_n)$. This requirement could be difficult to comply with if we intend to solve the subproblem by minimizing $\tilde{f}(x) = f_n(x) + (\gamma_n/2) \|x - x_n\|^2$. On the other hand, there exist methods for vector minimization which work directly with the order $\preceq_C$, and which are thus able to deal without trouble with constraints of the type $F(x) \preceq_C F(x_n)$. For instance, we mention the method introduced in [10] for finding weakly efficient points of differentiable maps $G : \mathbb{R}^p \rightarrow \mathbb{R}^m$ with respect to the partial order induced by a cone $C \subset \mathbb{R}^m$. This method, which extends the steepest descent method for scalar-valued minimization, generates a sequence $\{x_j\}$, which converges to a weakly efficient point for G , with the decreasing property $G(x_{j+1}) \preceq_C G(x_j)$ for all j . It is easy to check that if we are in the n th subproblem of either **ALG1** or **ALG2**, and if we apply a method with these properties to $G(x) = F(x) + (\alpha_n/2) \|x - x_n\|^2 e_n$, then the limit x^* of the generated sequence will satisfy $F(x^*) \preceq_C F(x_n)$ because of the decreasing property, and also (40) with $\varepsilon_n = 0$, because x^* is weakly efficient for G . Thus, such an x^* can be taken as the next iterate x_{n+1} of our methods in both cases. In fact, for Algorithm **ALG2** in the finite dimensional and differentiable setting, the method in [10] has an even better feature: applied to the same G as above, after a finite number of iterations it will generate an iterate $\bar{y}$ with the property that $\bar{y} - \theta J_G(\bar{y})^t z_n$ (where $J_G(x)$ is the Jacobian matrix of G at x , and θ is some positive number) satisfies (39)–(40) for some $\varepsilon_n > 0$, and thus can be taken as the next iterate x_{n+1} . Since the method in [10] is an explicit gradient-type scheme, this approach provides a truly finite procedure for the subproblems in **ALG2** in the smooth and finite dimensional case.

Acknowledgments. The first author wishes to express his gratitude to Professor A. N. Iusem for his hospitality during the author's visit to the Instituto de Matemática Pura e Aplicada.

REFERENCES

- [1] H. BENKER, A. HAMEL, AND C. TAMMER, *An algorithm for vectorial control approximation problems*, in Multiple Criteria Decision Making, G. Fandel and T. Gal, eds., Springer, Berlin, 1997, pp. 3–12.
- [2] S. BOLINTINÉANU, *Approximate efficiency and scalar stationarity in unbounded nonsmooth convex vector optimization problems*, J. Optim. Theory Appl., 106 (2000), pp. 265–296.
- [3] S. BOLINTINÉANU, *Vector variational principles; ε -efficiency and scalar stationarity*, J. Convex Anal., 8 (2001), pp. 71–85.
- [4] R. S. BURACHIK, A. N. IUSEM, AND B. F. SVAITER, *Enlargement of monotone operators with applications to variational inequalities*, Set-Valued Anal., 5 (1997), pp. 159–180.
- [5] R. S. BURACHIK, S. SCHEIMBERG, AND B. F. SVAITER, *Robustness of the hybrid extragradient proximal point algorithm*, J. Optim. Theory Appl., 111 (2001), pp. 117–136.
- [6] D. BUTNARIU AND A. N. IUSEM, *Totally Convex Functions for Fixed Point Computation and Infinite Dimensional Optimization*, Kluwer, Dordrecht, 2000.
- [7] J. FLIEGE AND B. F. SVAITER, *Steepest descent methods for multicriteria optimization*, Math. Methods Oper. Res., 51 (2000), pp. 479–494.
- [8] A. GÖPFERT, H. RIAHI, C. TAMMER, AND C. ZALINESCU, *Variational Methods in Partially Ordered Spaces*, Springer, New York, 2003.
- [9] L. M. GRAÑA DRUMMOND AND A. N. IUSEM, *A projected gradient method for vector optimization problems*, Comput. Optim. Appl., 28 (2004), pp. 5–30.

- [10] L. M. GRAÑA DRUMMOND AND B. F. SVAITER, *A steepest descent method for vector optimization*, J. Comput. Appl. Math., 175 (2005), pp. 395–414.
- [11] J. JAHN, *Theory of vector maximization: Various concepts of efficient solutions*, in Multicriteria Decision Making. Advances in MCDM Models, Algorithms, Theory, and Applications, T. Gal, T. J. Stewart, and T. Hanne, eds., Kluwer, Boston, 1999, pp. 2-1–2-32.
- [12] T. D. LUC, *Theory of Vector Optimization*, Lecture Notes in Econom. and Math. Syst. 319, Springer, Berlin, 1989.
- [13] B. MARTINET, *Régularisation d'inéquations variationnelles par approximations successives*, Rev. Française de Inform. Recherche Opér., 2 (1970), pp. 154–159.
- [14] B. MARTINET, *Algorithmes pour la Résolution de Problèmes d'optimisation et Minimax*, Thèse d'état, Université de Grenoble, 1972.
- [15] K. MIETTINEN AND M. M. MÄKELÄ, *Interactive bundle-based methods for non-differentiable multiobjective optimization: NIMBUS*, Optimization, 34 (1995), pp. 231–246.
- [16] J. MOREAU, *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.
- [17] R. PATZ, J. SPITZNER, AND C. TAMMER, *Decision support for location problems in town planning*, Internat. Trans. Oper. Res., 9 (2002), pp. 261–278.
- [18] R. R. PHELPS, *Convex Functions, Monotone Operators and Differentiability*, Lecture Notes in Math. 1364, Springer, Berlin, 1988.
- [19] R. T. ROCKAFELLAR, *A dual approach to solving nonlinear programming problems by unconstrained optimization*, Math. Programming, 5 (1973), pp. 354–373.
- [20] R. T. ROCKAFELLAR, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
- [21] M. V. SOLODOV AND B. F. SVAITER, *An inexact hybrid extragradient-proximal point algorithm using the enlargement of a maximal monotone operator*, Set-Valued Anal., 7 (1999), pp. 323–345.

SOLVING MULTICOMMODITY FLOW PROBLEMS BY AN APPROXIMATION SCHEME*

JORGE VILLAVICENCIO†

Abstract. We study an approximation scheme to solve minimum cost multicommodity flow problems to a relative accuracy of $\varepsilon \in (0, 1]$. The proposed scheme, which we shall call Algorithm $\mathcal{A}$, is a bisection-based procedure, so it maintains an interval that contains the optimal value. At each iteration, Algorithm $\mathcal{A}$ defines a block angular sharing problem that is solved to a relative accuracy of $O(\varepsilon)$. The computed solution defines the current approximation to an optimal solution and it is used by Algorithm $\mathcal{A}$ to throw away half of the current interval. It is shown that when Algorithm $\mathcal{A}$ no longer shrinks the current interval, the current solution solves the minimum cost multicommodity flow problem to the given accuracy ε . To compute approximate solutions to the sharing problems that Algorithm $\mathcal{A}$ defines, we propose using Algorithm $\mathcal{L}$ from [J. Villavicencio and M. D. Grigoriadis, *Approximate Lagrangian decomposition with a modified Karmarkar logarithmic potential*, in Network Optimization, Lecture Notes in Econom. and Math. Systems 450, Springer, Berlin, 1997, pp. 471–485], but we replace its fixed small step size by one computed using inexact line searches. This modification allows large steps in Algorithm $\mathcal{L}$. We show that the coordination complexities of Algorithm $\mathcal{L}$ and Algorithm $\mathcal{L}$ with inexact line searches are the same if these algorithms are applied to the block angular sharing problems defined by Algorithm $\mathcal{A}$. This result is also true for block angular sharing problems with convex blocks and with coupling constraints given by nonnegative linear functions.

Key words. multicommodity flow, line search, resource sharing problem, approximation algorithm, structured optimization, Lagrangian decomposition

AMS subject classifications. 65K05, 68W25, 68Q25, 90C08, 90C35

DOI. 10.1137/S1052623401391907

1. Introduction. A practical problem in large-scale structured optimization is the *minimum cost multicommodity flow* problem [1], defined as follows:

$$(1.1) \quad \begin{aligned} z^* &= \min \sum_{k=1}^K (c^k)^T x^k, \\ \sum_{k=1}^K x^k &\leq d, \\ Nx^k &= b^k, \quad k = 1, \dots, K, \\ 0 \leq x^k &\leq u^k, \quad k = 1, \dots, K, \end{aligned}$$

where K is the number of commodities, N is the node-arc incidence matrix of the given n -node M -arc directed network, d is the M -vector of arc capacities, and for each commodity $k = 1, \dots, K$, x^k is the M -vector of arc flows, c^k is an M -vector of arc cost coefficients, which we shall assume are nonnegative, b^k is a given exogenous flow vector which has one positive and one or more negative components so that $\sum_i b_i^k = 0$, and u^k is the upper bound vector on the arc flows. Each column of N , corresponding to an arc (i, j) , has two nonzeros, a $+1$ as its i th, and a -1 as its j th element.

*Received by the editors July 5, 2001; accepted for publication (in revised form) July 26, 2004; published electronically June 3, 2005. This research was supported by Fondo Nacional de Ciencia y Tecnología under grant Fondecyt No. 1990434.

<http://www.siam.org/journals/siopt/15-4/39190.html>

†Facultad de Matemáticas, Pontificia Universidad Católica de Chile, Casilla 306, Correo 22, Santiago, Chile (jvillavi@puc.cl).

We shall denote $x = (x^1, \dots, x^K)$, $z(x) = \sum_k (c^k)^T x^k$, $f(x) = \sum_k x^k$, and $B = B^1 \times B^2 \times \dots \times B^K$, where $B^k = \{x^k \mid Nx^k = b^k, 0 \leq x^k \leq u^k\}$ for each $k = 1, \dots, K$. With this notation, problem (1.1) can be written as

$$(\mathcal{P}) \quad \begin{aligned} z^* &= \min z(x) \\ \text{s.t. } f(x) &\leq d, \\ x &\in B, \end{aligned}$$

where $z(x)$ and $f(x)$ are nonnegative for $x \in B$. We shall denote by x^* an optimal solution of problem $(\mathcal{P})$. Observe that, without loss of generality, we can assume that $d > 0$.

For an M -vector of prices $p \geq 0$, we define the dual function ω associated with problem $(\mathcal{P})$ by

$$\begin{aligned} \omega(p) &= \min z(x) + p^T(f(x) - d) \\ \text{s.t. } x &\in B. \end{aligned}$$

Also, we define the Lagrangian dual problem

$$(\mathcal{D}) \quad \begin{aligned} z^* &= \max \omega(p) \\ \text{s.t. } p &\geq 0. \end{aligned}$$

For problems $(\mathcal{P})$ and $(\mathcal{D})$, Lagrangian duality [4] states that the optimal values of these two problems are equal.

The approximation variant of problem $(\mathcal{P})$ calls for computing an ε -approximate solution, where $\varepsilon \in (0, 1]$ is a given *relative accuracy*; i.e.,

$$(\mathcal{P}_\varepsilon) \quad \text{find } x \in B \text{ such that } z(x) \leq (1 + \varepsilon)z^* \text{ and } f(x) \leq (1 + \varepsilon)d.$$

Problem $(\mathcal{P})$ can be solved using a bisection search to find the smallest z for which there is an $x \in B$ that satisfies the constraints $z(x) \leq z$ and $f(x) \leq d$. Thus, the optimization problem can be reduced to a sequence of sharing problems, each of which can be formulated as a problem of the form

$$(1.2) \quad \begin{aligned} \lambda_z^* &= \min \lambda \\ \text{s.t. } z(x) &\leq \lambda z, \\ f(x) &\leq \lambda d, \\ x &\in B. \end{aligned}$$

There are several general [12], [13], [20], [24], [25] and even specialized [5], [14], [15] approximation schemes which can be applied to problem (1.2). Given $z \in \mathbb{R}$ we can use such an approximation scheme to compute λ , an ε -approximate solution of problem (1.2), i.e., $\lambda \leq (1 + \varepsilon)\lambda_z^*$. If $\lambda > 1 + \varepsilon$, then $z^* > z$, so z defines a lower bound for z^* . On the other hand, if $\lambda \leq 1$, then $z^* \leq z$, so z defines an upper bound for z^* . This observation can be used to define a bisection search scheme that approximately solves problem (1.2), in order to decide how to halve the length of the current interval [10], [15], [20]. Such a scheme halts when $1 < \lambda \leq 1 + \varepsilon$ so that it computes $z \in \mathbb{R}$ and $x \in B$, which satisfy $z(x) \leq (1 + \varepsilon)z$ and $f(x) \leq (1 + \varepsilon)d$. At this point, the above stopping condition is not strong enough to guarantee that the z computed so far is a good estimation of z^* since, when choosing z large enough, λ is determined only by the constraints $f(x) \leq \lambda d$. As a consequence of this, even though λ is such that $1 < \lambda \leq 1 + \varepsilon$, it may not be related to z , and thus we cannot infer any information

about the closeness of z to z^* . To illustrate this claim, in the appendix we exhibit an instance of problem $(\mathcal{P})$ that admits an ε -approximate solution λ of (1.2) such that $1 < \lambda \leq 1 + \varepsilon$ for all $z \geq \rho$ and some ρ . Therefore, z can be as far away as we want from the optimal value z^* . This example shows that the previous scheme is not correct and it clearly highlights that a bisection-based approximation scheme for problem $(\mathcal{P})$ has to be carefully designed.

In this work, we shall propose a bisection-based approximation scheme which reduces the computation of an $O(\varepsilon)$ -approximate solution to problem $(\mathcal{P})$ to a sequence of sharing problems, each of which is a relaxation of (1.2), and we shall show that this scheme is correct, i.e., it computes an $O(\varepsilon)$ -approximate solution to problem $(\mathcal{P})$.

Even though our proposed approximation scheme is presented to approximately solve minimum cost multicommodity flow problems, it is more general and can be used to approximately solve problem $(\mathcal{P})$ when it is defined by different functions $z(x)$ and $f(x)$ and set B from those presented here, but its implementation requires a solver capable of computing approximate solutions to each of the sharing problems it defines. If problem $(\mathcal{P})$ is convex and the vector function $f(x)$ is nonnegative over B , the sharing problems can be approximately solved using one of the algorithms from [12], [13], [24], [25], all of which use a fixed small step size. However, in [26] numerical experiments performed over a wide set of concurrent flow test problems showed that Algorithm $\mathcal{L}$ in [25] with the fixed small step length replaced by one computed using inexact line searches performs very well. Thus, Algorithm $\mathcal{L}$ with inexact line searches is a good choice among those algorithms able to solve the sharing problems generated if $(\mathcal{P})$ is a multicommodity flow problem.

The previous analysis in [25] implicitly bounds all step lengths from above by a fixed small constant and it cannot directly accommodate large steps. The steps computed by an inexact line search might be larger than the fixed small step size and, under this circumstance, the previous analysis does not cover these inexact step sizes. On the other hand, we shall show that an appropriately chosen inexact line search will always reject the fixed step length, so the fixed-size step is outside the set of steps considered by the inexact line-search. Here we provide a complexity analysis which allows step lengths determined by inexact line searches and we show that the iteration bound of the new algorithm matches our previous iteration bound. But, based on the numerical experiments performed in [26] with Algorithm $\mathcal{L}$ with inexact line searches, and more general experience with inexact line searches [18], [22], in practice one might obtain far better overall reductions in the potential, a claim which has to be verified with numerical experimentation not performed here.

We observe that, although our complexity result for inexact line searches is presented as a way to improve the performance of our approximation scheme, it can be stated independently of it. In addition, our approximation scheme does not depend on our results for inexact line searches.

The paper is organized as follows. In section 2 we present our approximation scheme for problem $(\mathcal{P})$ and show its correctness. In section 3 we review the elements required to implement our approximation scheme and, in particular, we review Algorithm $\mathcal{L}$. In section 4 we discuss line searches. In section 5 we derive a bound on the potential function reduction obtained at each iteration of our proposed scheme, when an inexact line search based on the *Goldstein–Armijo* or *Wolfe* rule is used to compute the step size (Lemma 5.4). Then, in Theorem 5.6 we show how this reduction translates into the iteration bound

$$N_\varepsilon = O(M(\varepsilon^{-2} \ln \varepsilon^{-1} + \ln M)).$$

2. The approximation scheme. Our approximation scheme is based on bisection search and, for a given $\varepsilon > 0$ and z , it defines the problem

$$(2.1) \quad \begin{aligned} \lambda_z &= \min \lambda \\ \text{s.t. } z(x) &\leq \lambda z, \\ f(x) &\leq \lambda(1 + \varepsilon)d, \\ x &\in B, \end{aligned}$$

which is a perturbation of problem (1.2). Our scheme requires only the computation of an $(\varepsilon/3)$ -approximate solution to problem (2.1), i.e., a pair (x, λ) with $x \in B$ such that

$$(2.2) \quad \lambda \leq (1 + \varepsilon/3)\lambda_z, \quad z(x) \leq \lambda z, \quad \text{and} \quad f(x) \leq \lambda(1 + \varepsilon)d.$$

Working with such an approximate solution helps us to keep the computational costs involved in the overall process low. However, to estimate the optimal value z^* , within some desired error tolerance, using just approximate solutions to problem (2.1) a new scheme is required. We present such a scheme in the following algorithm, whose input consists of the functions z and f , the vector d , the set B , a relative error tolerance $\varepsilon \in (0, 1]$, and $\mu \in (0, 1]$, a tolerance we introduce to deal with the case $z^* = 0$. We assume μ to be small enough so that when $z^* = 0$, any $x \in B$ such that $z(x) \leq \mu$ and $f(x) \leq (1 + \varepsilon)d$ can be taken as a good approximation to x^* . The algorithm output is $x \in B$, which is either an ε -approximate solution to problem $(\mathcal{P})$ or satisfies $z(x) \leq \mu$ and $f(x) \leq (1 + \varepsilon)d$.

ALGORITHM $\mathcal{A}(z, f, d, B, \mu, \varepsilon, x)$.

Let $\underline{z} = \mu/(1 + \varepsilon)$, $\bar{z} = \infty$, and $z = \mu$.

Compute x, λ an $(\varepsilon/3)$ -approximate solution of (2.1).

if $\lambda \leq (1 + \varepsilon/3)/(1 + \varepsilon)$ **stop**.

repeat

if $\bar{z} = \infty$ **then** $z = 2\underline{z}$

else $z = (\bar{z} + \underline{z})/2$.

Compute x, λ an $(\varepsilon/3)$ -approximate solution of (2.1).

if $1 - \varepsilon/3 < \lambda \leq 1 + \varepsilon/2$ **stop**.

if $\lambda \leq 1 - \varepsilon/3$ **then** $\bar{z} = z(1 - \varepsilon/3)$

else $\underline{z} = z(1 + \varepsilon/2)/(1 + \varepsilon/3)$.

end

Observe that Algorithm $\mathcal{A}$ also can be applied to optimization problems defined by other kind of functions such as $z(x)$ and $f(x)$ and set B .

If $\bar{z} = \infty$, the setting $z = 2\underline{z}$ is quite arbitrary and can be replaced by any other which generates a strictly increasing sequence of z -values that goes to infinity. Also, if $\bar{z} < \infty$, any $z \in (\underline{z}, \bar{z})$ can be used instead of the middle point of the interval $(\underline{z}, \bar{z})$.

In what follows, referring to problem (2.1) and for $x \in B$, we shall denote

$$\lambda(x) \doteq \max \left\{ \frac{z(x)}{z}, \frac{f_1(x)}{(1 + \varepsilon)d_1}, \dots, \frac{f_M(x)}{(1 + \varepsilon)d_M} \right\},$$

where $f_1(x), \dots, f_M(x)$ are the components of the M -vector $f(x)$.

To analyze Algorithm $\mathcal{A}$, we shall assume that at least one of the constraints $f(x) \leq d$ is tight at $x = x^*$, with x^* an optimal solution to problem $(\mathcal{P})$. Provided that this assumption holds, we have that

$$\max \left\{ \frac{f_1(x^*)}{(1 + \varepsilon)d_1}, \dots, \frac{f_M(x^*)}{(1 + \varepsilon)d_M} \right\} = \frac{1}{1 + \varepsilon}$$

and then

$$(2.3) \quad \lambda(x^*) = \max \left\{ \frac{1}{1+\varepsilon}, \frac{z^*}{z} \right\}.$$

Our next observation about Algorithm $\mathcal{A}$ is that, at every iteration $\underline{z} \leq z^*$. Initially the algorithm sets $z = \mu$ and computes $x \in B$ and λ such that $z(x) \leq \lambda\mu$ and $f(x) \leq \lambda(1+\varepsilon)d$. If $\lambda \leq (1+\varepsilon/3)/(1+\varepsilon)$, $z(x)$ is an approximation to z^* and the algorithm correctly stops. On the other hand, if $\lambda > (1+\varepsilon/3)/(1+\varepsilon)$, since λ is an $(\varepsilon/3)$ -approximate solution of (2.1), we have

$$(1+\varepsilon/3)\lambda_z \geq \lambda > \frac{1+\varepsilon/3}{1+\varepsilon},$$

and then $\lambda_z > 1/(1+\varepsilon)$. Thus, $\lambda(x^*) > 1/(1+\varepsilon)$ and consequently $\underline{z} = \mu/(1+\varepsilon) \leq z^*$. In the following iterations Algorithm $\mathcal{A}$ updates $\underline{z}$ only if $\lambda > 1+\varepsilon/2$. Since λ is an $(\varepsilon/3)$ -approximate solution to problem (2.1),

$$(1+\varepsilon/3)\lambda_z \geq \lambda > 1+\varepsilon/2.$$

Therefore,

$$(2.4) \quad \lambda(x^*) \geq \lambda_z > \frac{1+\varepsilon/2}{1+\varepsilon/3} > 1.$$

Thus, $\lambda(x^*) > 1$ and by (2.3) we see that $\lambda(x^*) = z^*/z$, which together with (2.4) implies that

$$(2.5) \quad \frac{z^*}{z} > \frac{1+\varepsilon/2}{1+\varepsilon/3} > 1.$$

Inequality (2.5) shows that Algorithm $\mathcal{A}$ updates $\underline{z}$ in such a way that $z^* \geq \underline{z}$ is always satisfied. Since $z^* \geq \underline{z}$ at the beginning of Algorithm $\mathcal{A}$, we can conclude that this inequality will hold for all iterations of Algorithm $\mathcal{A}$.

Next, we shall show that Algorithm $\mathcal{A}$ performs a finite number of iterations. To see this, we first consider those iterations for which $\bar{z} = \infty$. Given that at each of these iterations $\underline{z}$ is doubled and $\underline{z} \leq z^*$, we can conclude that the number of these iterations is at most $O(\log(z^*/\mu))$.

Now we study the iterations of Algorithm $\mathcal{A}$ for which $\bar{z}$ is finite. Once Algorithm $\mathcal{A}$ sets $\bar{z}$ to a finite value, this value is $\bar{z} = 2\underline{z}$, where $\underline{z}$ is the current value defined by Algorithm $\mathcal{A}$. Thus, after ℓ updates of $\bar{z}$, the corresponding value $\bar{z}'$ is such that

$$\underline{z} \leq \bar{z}' \leq \bar{z}(1-\varepsilon/3)^\ell \leq 2\underline{z}(1-\varepsilon/3)^\ell,$$

which shows that the number of times $\bar{z}$ is updated is bounded by $O(1/|\log(1-\varepsilon)|)$. Similarly, after ℓ updates of $\underline{z}$ the corresponding value $\underline{z}'$ is such that

$$2\underline{z} = \bar{z} \geq \underline{z}' \geq \left(\frac{1+\varepsilon/2}{1+\varepsilon/3} \right)^\ell \underline{z} = \left(1 + \frac{\varepsilon/6}{1+\varepsilon/3} \right)^\ell \underline{z} \geq (1+\varepsilon/8)^\ell \underline{z},$$

which shows that the number of times $\underline{z}$ is updated is bounded by $O(1/\log(1+\varepsilon))$.

In the previous paragraphs we showed that Algorithm $\mathcal{A}$ performs a finite number of iterations and that this number is bounded by a logarithmic factor depending on the data that define the problem. Next, we shall show that Algorithm $\mathcal{A}$ is correct.

LEMMA 2.1. *Algorithm $\mathcal{A}$ halts with $x \in B$, which satisfies either $z(x) \leq \mu$ and $f(x) \leq (1 + \varepsilon)d$, or*

$$z(x) \leq (1 + 2\varepsilon)z^* \quad \text{and} \quad f(x) \leq (1 + 2\varepsilon)d.$$

Proof. If Algorithm $\mathcal{A}$ halts at its first stop, then it correctly halts with $x \in B$ such that $z(x) \leq \mu$ and $f(x) \leq (1 + \varepsilon)d$. Otherwise, Algorithm $\mathcal{A}$ halts with $x \in B$ and λ such that $1 - \varepsilon/3 < \lambda \leq 1 + \varepsilon/2$. Furthermore, λ is an $(\varepsilon/3)$ -approximate solution of (2.1), i.e., $\lambda \leq (1 + \varepsilon/3)\lambda_z$. These estimates imply that

$$\lambda_z \geq \frac{\lambda}{1 + \varepsilon/3} > \frac{1 - \varepsilon/3}{1 + \varepsilon/3} > \frac{1}{1 + \varepsilon}.$$

The above inequality, together with $\lambda(x^*) \geq \lambda_z$, implies that

$$\lambda(x^*) > \frac{1}{1 + \varepsilon}.$$

From (2.3) we can see that $\lambda(x^*) = z^*/z$ so that $z < (1 + \varepsilon)z^*$. Since the pair $x \in B$ and λ satisfies (2.2), we have

$$z(x) \leq (1 + \varepsilon/2)z < (1 + \varepsilon/2)(1 + \varepsilon)z^* \leq (1 + 2\varepsilon)z^*$$

and

$$f(x) \leq (1 + \varepsilon/2)(1 + \varepsilon)d \leq (1 + 2\varepsilon)d. \quad \square$$

3. Implementation of Algorithm $\mathcal{A}$. To implement Algorithm $\mathcal{A}$, we must have an algorithm which approximately solves the block angular min-max resource sharing problem (2.1). Algorithm $\mathcal{L}$, defined in [25], has shown very good practical performance in approximately solving sharing problems when applied to a wide set of concurrent flow test problems [26]. In our implementation of Algorithm $\mathcal{A}$ we shall use Algorithm $\mathcal{L}$ to approximately solve problem (2.1). Therefore, we shall summarize those aspects of Algorithm $\mathcal{L}$ which are relevant to our implementation of Algorithm $\mathcal{A}$. For simplicity, we shall denote $M^+ = M + 1$ and $f_{M^+}(x) = z(x)$ for all $x \in B$. Algorithm $\mathcal{L}$ assumes that in (2.1) the functions f_m , $m = 1, \dots, M^+$, are nonnegative and convex and that $z = 1$ and $d = e$, where e is the vector with all its components equal to 1. Since $d > 0$ and $z > 0$, by scaling the constraints in (2.1) if necessary we can assume that $z = 1$ and $d = e$ as required by Algorithm $\mathcal{L}$.

Algorithm $\mathcal{L}$ is based on minimizing the *potential function*

$$(3.1) \quad \Phi_t(\theta, x) = \ln \theta - \frac{t}{M^+} \sum_{m=1}^{M^+} \ln(\theta - f_m(x))$$

over $x \in B$, where $t > 0$ is a small fixed parameter. Since $\Phi_t(\cdot)$ has the barrier property, i.e., $\Phi_t(\theta, x) \rightarrow \infty$ for $\theta \rightarrow \lambda(x)$ or for $\theta \rightarrow \infty$, a solution of (2.1) can be obtained by solving a sequence of convex programs of the form $\min\{\Phi_t(\theta, x) \mid x \in B, \theta \in \mathbb{R}_+\}$ for some decreasing sequence of values for t . It is easy to see from the optimality conditions of such problems that the minimum value of θ can be determined for a fixed t and $x \in B$ by computing the (unique) root $\theta(x)$ of the equation

$$(3.2) \quad \frac{t}{M^+} \sum_{m=1}^{M^+} \frac{\theta}{\theta - f_m(x)} = 1.$$

It is shown (Lemmas 1 and 2 in [25]) that the implicit function $\theta(x)$ approximates $\lambda(x)$, and the potential function $\Phi_t(\theta(x), x)$ approximates $\ln \lambda(x)$, as follows:

$$(1-t)\theta(x) \leq \lambda(x) \leq (1-t/M^+)\theta(x),$$

$$(1-t)\ln \lambda(x) \leq \Phi_t(\theta(x), x) \leq (1-t)\ln \lambda(x) + t\ln(e/t).$$

For sufficiently small values of the parameter t , the above observations permit us to replace our original task of minimizing the piecewise function $\lambda(x)$ by the minimization of the smooth potential $\Phi_t(\theta(x), x)$ over $x \in B$.

A typical iteration of our algorithm starts with the current iterate $x \in B$ for which a *coordinator* computes the unique solution $\theta(x)$ of (3.2) and directly determines the vector $p \in \mathbb{R}^{M^+}$ of Lagrange multipliers or “prices” for the constraints from the definition

$$(3.3) \quad p_m \doteq p_m(x) = \frac{t}{M^+} \frac{\theta(x)}{\theta(x) - f_m(x)}, \quad m = 1, \dots, M^+,$$

where $p \in P \doteq \{p \in \mathbb{R}^{M^+} \mid e^T p = 1, p \geq 0\}$ since $\theta(x)$ is the root of (3.2). This p is then passed to an ε' -approximate block solver (with $\varepsilon' = O(\varepsilon)$) which computes an ε' -approximate block minimizer $\hat{x}$ such that $p^T f(\hat{x}) \leq (1+\varepsilon') \min\{p^T f(y) \mid y \in B\}$. This $\hat{x}$, which we denote by $\mathcal{ABS}(p, \varepsilon')$, is returned to the coordinator, which then computes a new iterate $x \in B$ and checks a termination criterion based on the relative error in the duality gap. The algorithm is summarized below. For notational convenience we shall henceforth denote $\theta \doteq \theta(x)$, $f_m \doteq f_m(x)$, and $\hat{f}_m \doteq f_m(\hat{x})$.

ALGORITHM $\mathcal{L}(f, B, \varepsilon)$.

begin

Initialize: $\sigma = 2$, $t = \sigma/10$, $p = e/M^+$, and $x := \mathcal{ABS}(p, t)$.

while $\sigma > \varepsilon$ **do**

begin $\{\sigma$ -scaling phase $\}$

repeat $\{\text{coordination step}\}$

Compute θ from (3.2) and $p \in P$ from (3.3).

$\hat{x} := \mathcal{ABS}(p, t)$.

$x := (1 - \tau)x + \tau\hat{x}$ for a small, fixed step length $\tau \in (0, 1]$.

until $p^T(f - \hat{f})/\theta \leq \sigma/2$

$\sigma := \sigma/2$, $t := \sigma/10$.

end

end

The algorithm is invoked with the desired relative solution accuracy $\varepsilon \in (0, 1]$ and proceeds in $O(\log \varepsilon^{-1})$ “ σ -scaling phases” by reducing the current relative accuracy σ from its initial value down to the given ε . It is shown in [25] that the *coordination complexity* of Algorithm $\mathcal{L}$, i.e., the number of iterations needed to compute an ε -approximate $x \in B$, is

$$N_\varepsilon = O\left(M\left(\varepsilon^{-2} \ln \varepsilon^{-1} + \ln M\right)\right).$$

As a practical matter, since an $O(\sigma)$ -approximate solution of (2.1) and of its Lagrangian dual is obtained at each σ -scaling phase, the process can be modulated, monitored, and terminated after any scaling phase, even before the originally specified relative accuracy ε is reached.

The core of Algorithm $\mathcal{L}$ is a coordination step, which consists of the computation of θ , a call to a block solver, and the computation of a step length. The computation

of $\theta(x)$ is independent of the complexity for evaluating $f(x)$, but in order to bound the error in each component of p by $O(\varepsilon)$, it is necessary to compute $\theta(x)$ to a binary length of $O(\ln M/\varepsilon)$ bits. Once $f(x)$ has been computed, the evaluation of $\theta(x)$ involves $O(\ln M/\varepsilon)$ bisections of the interval $[\lambda(x)/(1-t/M), \lambda(x)/(1-t)]$. Each such bisection requires $O(M)$ arithmetic operations. Thus, ignoring the cost of computing $f(x)$, each coordination step runs in $O(M \ln(M/\varepsilon))$ sequential or $O(\ln M \ln(M/\varepsilon))$ parallel time on an $M/\ln M$ -processor EREW PRAM.

In our case, the block subproblems that Algorithm $\mathcal{L}$ defines are single commodity minimum cost network flow problems, so that the block solver can be selected from the many efficient algorithms for solving minimum cost network flow problems [1]. For example, these problems can be solved by using the algorithm of Goldberg and Tarjan [9] in $\tilde{O}(nM)$ time, where $\tilde{O}$ means we have suppressed the logarithmic terms.

Algorithm $\mathcal{L}$ uses the small fixed step length

$$(3.4) \quad \tau = \frac{3t\sigma}{32M^+},$$

although its performance can be improved using alternative forms to define this step length. Next, we will address the issue in which the coordinator computes a step length based on inexact line searches.

4. Inexact line searches. Inexact step sizes have long been known to give good practical results. We shall show that when they are used inside Algorithm $\mathcal{L}$ to compute its step length, the coordination complexity bound [25] of the algorithm is kept. The choice of the step length can range from just a fixed step length, as in the analysis in [25], to one which minimizes the objective function along a search direction. Several algorithms, such as golden section, Fibonacci search, bisection search, and Newton's method, all of which have been widely studied in nonlinear programming, can be used for this purpose. Such line search algorithms are called *exact* since they can compute a (local) minimum of the objective function over the search direction to an arbitrary accuracy. They are generally inefficient in practice, particularly for problems with expensive function and/or gradient evaluations.

Faster *inexact* or *approximate* line searches, first studied in [3], [11], [27], and others, are commonly used. They replace the stopping rules which are based on the optimality conditions and used by exact line searches, by weaker termination criteria. These criteria consist of a *sufficient decrease rule* which prevents large steps while allowing strict decreases in the function value along the selected search direction, and a *large step rule* which ensures sufficiently large decreases in the function value. As its name suggests, the second rule tends to exclude the arbitrarily small step lengths admitted by the first rule. These two rules (implicitly) define an interval of "acceptable" step lengths, say, $\tau \in [\tau_L, \tau_R]$. Line search can be implemented by a variety of techniques, including bisection search. Most methods start with a larger initial search interval and then iterate until a trial step length τ , satisfying the above two rules, is encountered.

Popular inexact line search methods use a *sufficient decrease rule* of the form

$$(4.1) \quad \phi(\tau) \leq \phi(0) + \alpha_1 \phi'(0)\tau \quad \text{for } 0 < \alpha_1 < 1$$

and use variants of the following two *large-step rules*:

$$(4.2a) \quad \phi(\tau) \geq \phi(0) + \alpha_2 \phi'(0)\tau \quad \text{for } \alpha_1 < \alpha_2 < 1,$$

$$(4.2b) \quad \phi'(\tau) \geq \alpha_2 \phi'(0) \quad \text{for } \alpha_1 < \alpha_2 < 1,$$

where $\phi(\tau)$ is the function value along the search direction and the scalars α_1, α_2 are fixed parameters, independently selected for each method.

We shall refer to the pair (4.1), (4.2a) as the Goldstein–Armijo step-size rule (GA rule) and refer to the pair (4.1), (4.2b) as the Wolfe rule (W rule). Methods based on the former require no derivative evaluations except at $\tau = 0$, but those based on the W rule may require them at every trial point. Among the many inexact line search methods proposed heretofore, the earliest seem to be by Polak [21] using the GA rule, by Lemaréchal [17] and Powell [23] using the W rule, and by Fletcher [6], who uses the following variant of the W rule:

$$(4.3) \quad |\phi'(\tau)| \leq \eta |\phi'(0)| \quad \text{for } 0 < \alpha_1 < \eta < 1,$$

often referred to as the strong Wolfe conditions. Gill et al. [8] suggested another large-step rule that requires the derivative only at $\tau = 0$. Moré and Thunberg's line search algorithm [18] uses (4.1) along with rule (4.3). Potra and Shi [22] generalize the W rule so that the GA and Gill et al. large-step rules are included.

Inexact line searches have recently been specialized for use with the logarithmic barrier and potential functions. This use was mainly motivated by interior point methods for linear programming. Murray and Wright [19] used interpolating functions specialized to the logarithmic barrier to compute a W-type step length. Anstreicher [2] proved that Karmarkar's $O(\sqrt{n}L)$ iteration bound [16] is retained if the fixed step size of that algorithm is replaced by an inexact line search based on rule (4.1) and a large-step rule, which is the fixed potential reduction needed in the complexity analysis of that algorithm. For additional details on inexact line search algorithms and their analyses, see, e.g., Bazaraa, Sherali, and Shetty [4], Fletcher [6], Gill, Murray, and Wright [7], and Lemaréchal [17].

5. Analysis of Algorithm $\mathcal{L}$ with inexact line searches. In this section we present our new analysis of Algorithm $\mathcal{L}$ in which we have replaced the fixed step τ used in [25] by one which satisfies either the GA rule (4.1), (4.2a) or the W rule (4.1), (4.2b). The analysis shows that the coordination complexities of Algorithm $\mathcal{L}$ and Algorithm $\mathcal{L}$ with the step sizes computed with the GA or W rule are the same. The analysis also points out that choosing appropriately α_2 in the GA or W rule causes the fixed step size (3.4) to be rejected by both rules, so a new analysis for Algorithm $\mathcal{L}$ with inexact line searches is required. These two analyses obtain their respective coordination complexities showing that a large enough decrease of the potential function value is obtained at each iteration. However, they follow different strategies to show this. The analysis given in [25] bounds from above the potential reduction and then replaces the small fixed step size in this bound to obtain the desired potential reduction. Note that it is crucial to the validity of this analysis that the step size be small. On the other hand, the analysis given here focuses on showing that the step size is large enough so that the desired potential reduction is obtained from (4.1).

We state without proof the following observations from [25].

LEMMA 5.1. $p^T f = \theta(1 - t)$ for any $x \in B$.

LEMMA 5.2. $p^T(f + \hat{f}) \leq 2\theta$ for any $x \in B$.

LEMMA 5.3. $(1 - t) \ln \lambda(x) \leq \Phi_t(\theta(x), x) \leq (1 - t) \ln \lambda(x) + t \ln(e/t)$ for $x \in B$.

We start our analysis by restating the potential function in a form more convenient for our purposes. We denote by $\phi_t(\tau)$ the potential along the search direction $x - \hat{x}$ provided by Algorithm $\mathcal{L}$, i.e., $\phi_t(\tau) \doteq \Phi(\theta, (1 - \tau)x + \tau\hat{x})$. By the linearity of the

f_m , we can write $\phi_t(\tau)$ as follows:

$$\begin{aligned}\phi_t(\tau) &= \ln \theta - \frac{t}{M^+} \sum_{m=1}^{M^+} \ln(\theta - f_m + \tau(f_m - \hat{f}_m)) \\ &= \ln \theta - \frac{t}{M^+} \sum_{m=1}^{M^+} \ln \left((\theta - f_m) \left(1 + \tau \frac{f_m - \hat{f}_m}{\theta - f_m} \right) \right).\end{aligned}$$

By (3.1), this translates into

$$\phi(\tau) = \phi_t(0) - \frac{t}{M^+} \sum_{m=1}^{M^+} \ln \left(1 + \tau \frac{f_m - \hat{f}_m}{\theta - f_m} \right),$$

and the definition (3.3) of the prices $p_m \in P$ gives

$$\phi_t(\tau) = \phi_t(0) - \frac{t}{M^+} \sum_{m=1}^{M^+} \ln \left(1 + \tau \frac{M^+}{t} \frac{p_m(f_m - \hat{f}_m)}{\theta} \right).$$

For notational convenience, we introduce

$$(5.1) \quad u_m \doteq \frac{M^+}{t} \frac{p_m(f_m - \hat{f}_m)}{\theta}$$

and then we have that

$$(5.2) \quad \phi_t(\tau) = \phi_t(0) - \frac{t}{M^+} \sum_{m=1}^{M^+} \ln(1 + \tau u_m).$$

The derivative of the potential is the expression

$$(5.3) \quad \phi'_t(\tau) = -\frac{t}{M^+} \sum_{m=1}^{M^+} \frac{u_m}{1 + \tau u_m},$$

$$(5.4) \quad \phi'_t(0) = -\frac{t}{M^+} \sum_{m=1}^{M^+} u_m = -\frac{p^T(f - \hat{f})}{\theta}.$$

Note that by the nonnegativity of the f_m and by Lemma 5.2, $p_m|f_m - \hat{f}_m| \leq p_m(f_m + \hat{f}_m) \leq p^T(f + \hat{f}) \leq 2\theta$. This inequality and (5.1) provide the estimate

$$(5.5) \quad |u_m| = \frac{M^+}{t} \frac{p_m|f_m - \hat{f}_m|}{\theta} \leq \frac{2M^+}{t} \quad \text{for all } m = 1, \dots, M^+.$$

Also, by the nonnegativity of f_m , $m = 1, \dots, M^+$,

$$\sum_{m=1}^{M^+} u_m^2 \leq \left(\frac{M^+}{t} \right)^2 \sum_{m=1}^{M^+} \left(\frac{p_m(f_m + \hat{f}_m)}{\theta} \right)^2 \leq \left(\frac{M^+}{t} \right)^2 \left(\frac{p^T(f + \hat{f})}{\theta} \right)^2$$

and then, by Lemma 5.2,

$$(5.6) \quad \sum_{m=1}^{M^+} u_m^2 \leq 4 \left(\frac{M^+}{t} \right)^2.$$

The stopping criterion of Algorithm $\mathcal{L}$ guarantees that $\phi'_t(0) \leq -\sigma/2$ so that it is always possible to compute a descent step $\tau > 0$. However, for our logarithmic potential, τ must also satisfy

$$(5.7) \quad \tau \leq \tau_{max} = \min\{1, \bar{\tau}\},$$

where $\bar{\tau}$ is the singularity of the function $\phi_t(\cdot)$, and this restriction may well conflict with the large-step rule (4.2a) or (4.2b). Nevertheless, if the sufficient decrease rule (4.1) is satisfied for $\tau = \tau_{max}$, then $\tau_{max} = 1$ since $\phi_t(\bar{\tau}) = \infty$. In this case, the unit step τ_{max} gives a decrease of at least $\alpha_1\sigma/2$ in the value of the potential ϕ_t because (4.1) reduces to

$$(5.8) \quad \phi_t(\tau) \leq \phi_t(0) + \alpha_1\phi'_t(0) \leq \phi_t(0) - \alpha_1\sigma/2.$$

On the other hand, if τ_{max} does not satisfy (4.1), the algorithm in [17] finds $\tau < \tau_{max}$ which satisfies condition (4.1) and either (4.2a) or (4.2b). In view of (5.8), we shall be concerned only with the latter general case.

LEMMA 5.4. *Suppose that $\tau > 0$ satisfies the “sufficient decrease” rule (4.1) and either the GA (4.2a) or W (4.2b) large-step rule for some fixed $0 < \alpha_1 < \alpha_2 < 1$. Then*

$$\phi_t(\tau) \leq \phi_t(0) - \frac{3t\alpha_1(1-\alpha_2)}{32M^+} \left(\frac{p^T(\hat{f}-f)}{\theta} \right)^2.$$

Moreover, for α_2 adequately chosen, the fixed-size step (3.4) is not in the set of possible steps which either rule considers.

Proof. By (5.4), it is easy to see that condition (4.1) proves the claim for any

$$(5.9) \quad \tau \geq \frac{3t(1-\alpha_2)}{32M^+} \frac{p^T(f-\hat{f})}{\theta}.$$

We consider two cases. The first is that τ satisfies (4.2a) and the second is that τ satisfies (4.2b).

Case 1. We shall first prove that if τ satisfies (4.2a) and $|\tau u_m| \leq 3/16$, $m = 1, \dots, M^+$, then

$$(5.10) \quad \tau \geq \frac{3t(1-\alpha_2)}{8M^+} \frac{p^T(f-\hat{f})}{\theta}.$$

From this statement we can prove (5.9) since, assuming the contrary, i.e.,

$$\tau < \frac{3t(1-\alpha_2)}{32M^+} \frac{p^T(f-\hat{f})}{\theta},$$

we immediately have from (5.5) that

$$|\tau u_m| \leq \frac{3(1-\alpha_2)p^T(f-\hat{f})}{16\theta} \leq \frac{3(1-\alpha_2)p^T f}{16\theta}.$$

Since $p^T f = (1-t)\theta$ by Lemma 5.1, we have that $|\tau u_m| \leq 3/16$ for $m = 1, \dots, M^+$, and then (5.10) holds and thus furnishes the desired contradiction.

Now, we prove (5.10). We first derive an upper bound on $\phi_t(\tau)$ from (5.2):

$$(5.11) \quad \begin{aligned} \phi_t(\tau) &= \phi_t(0) - \frac{t}{M^+} \sum_{m=1}^{M^+} \ln(1 + \tau u_m) \\ &\leq \phi_t(0) - \frac{t\tau}{M^+} \sum_{m=1}^{M^+} u_m + \frac{2t\tau^2}{3M^+} \sum_{m=1}^{M^+} u_m^2, \end{aligned}$$

where the second line follows from the elementary inequality (5.14) of Proposition 5.5 below. Second, we consider that replacing (5.6) in (5.11) gives

$$\phi_t(\tau) \leq \phi_t(0) - \frac{\tau p^T(f - \hat{f})}{\theta} + \frac{8\tau^2 M^+}{3t}.$$

Next, we combine the above inequality and the GA large-step rule (4.2a) to write

$$\phi_t(0) + \alpha_2 \phi'_t(0)\tau \leq \phi_t(0) - \frac{\tau p^T(f - \hat{f})}{\theta} + \frac{8\tau^2 M^+}{3t},$$

which by (5.4) simplifies to $(1 - \alpha_2)p^T(f - \hat{f})/\theta \leq 8\tau M^+/3t$, or (5.10).

Observe that the fixed step length (3.4) of Algorithm $\mathcal{L}$ is such that

$$|\tau u_m| \leq \frac{3t\sigma}{32M^+} \frac{2M^+}{t} = \frac{3\sigma}{16} \leq \frac{3}{16}.$$

Thus, if we assume that the fixed step length satisfies (4.2a), then inequality (5.10) holds. Since $p^T(f - \hat{f})/\theta > \sigma/2$ by the stopping condition of Algorithm $\mathcal{L}$, we have

$$\tau \geq 2(1 - \alpha_2) \frac{3t\sigma}{32M^+}.$$

Then, for $\alpha_2 < 1/2$ we can conclude that

$$\tau > \frac{3t\sigma}{32M^+},$$

which contradicts (3.4); i.e., for $\alpha_2 < 1/2$, the fixed step size (3.4) defined by Algorithm $\mathcal{L}$ is never accepted by the GA rule.

Case 2. In this case, we shall first show that any τ satisfying (4.2b) and $|\tau u_m| \leq 3/16$, $m = 1, \dots, M^+$, is such that

$$(5.12) \quad \tau \geq \frac{13t(1 - \alpha_2)}{64M^+} \frac{p^T(f - \hat{f})}{\theta}.$$

To prove (5.12), we first derive an upper bound on $\phi'_t(\tau)$ from (5.3) and the elementary inequality (5.15) of Proposition 5.5 below:

$$\phi'_t(\tau) = -\frac{t}{\tau M^+} \sum_{m=1}^{M^+} \frac{\tau u_m}{1 + \tau u_m} \leq -\frac{t}{\tau M^+} \left(\tau \sum_{m=1}^{M^+} u_m - \frac{16\tau^2}{13} \sum_{m=1}^{M^+} u_m^2 \right),$$

which combined with the bound (5.6) gives

$$(5.13) \quad \phi'_t(\tau) \leq \frac{-p^T(f - \hat{f})}{\theta} + \frac{64M^+\tau}{13t}.$$

Next, we combine the W large-step rule (4.2b), (5.4), and (5.13) to write

$$-\alpha_2 \frac{p^T(f - \hat{f})}{\theta} \leq \phi'_t(\tau) \leq -\frac{p^T(f - \hat{f})}{\theta} + \frac{64M^+\tau}{13t},$$

which yields the bound (5.12).

Now we prove that all τ satisfying (4.2b) are such that

$$\tau \geq \frac{3t(1 - \alpha_2)}{32M^+} \frac{p^T(f - \hat{f})}{\theta}.$$

Assuming the contrary, by (5.5) we have

$$|\tau u_m| < \frac{3(1 - \alpha_2)}{16} \frac{p^T(f - \hat{f})}{\theta},$$

which by Lemma 5.1 translates into $|\tau u_m| \leq 3/16$. Then inequality (5.12) holds and thus furnishes the desired contradiction.

Observe that the fixed step length (3.4) of Algorithm $\mathcal{L}$ is such that

$$|\tau u_m| \leq \frac{3t\sigma}{32M^+} \frac{2M^+}{t} = \frac{3\sigma}{16} \leq \frac{3}{16}.$$

Thus, if we assume that the fixed step length satisfies (4.2b), then inequality (5.12) holds. Since $p^T(f - \hat{f})/\theta > \sigma/2$ by the stopping condition of Algorithm $\mathcal{L}$, we have

$$\tau \geq \frac{13(1 - \alpha_2)}{12} \frac{3t\sigma}{32M^+}.$$

Then, for $\alpha_2 < 1/13$ we can conclude that

$$\tau > \frac{3t\sigma}{32M^+},$$

which contradicts (3.4); i.e., for $\alpha_2 < 1/13$, the fixed step size (3.4) defined by Algorithm $\mathcal{L}$ is never accepted by the W rule.

Hence, either of the two large-step rules is sufficient to guarantee the claimed decrease in the potential. In addition, for α_2 adequately chosen the fixed step size of Algorithm $\mathcal{L}$ will be rejected by both large-step rules. $\square$

We now provide simple proofs for the elementary inequalities we used in the previous analysis.

PROPOSITION 5.5. *For all $|w| \leq 3/16$, the following are valid inequalities:*

$$(5.14) \quad \ln(1 + w) \geq w - 2w^2/3,$$

$$(5.15) \quad w/(1 + w) \geq w - 16w^2/13.$$

Proof. First, we prove (5.14). Expanding this function yields

$$(5.16) \quad \ln(1 + w) = w - w^2/2 + (1 + \chi)^{-3}w^3/3,$$

with $0 < |\chi| < |w|$. Since $|w| \leq 3/16$ and $(1 + \chi)^{-3}$ is decreasing, $(1 + \chi)^{-3}w \geq -|w|(1 - |\chi|)^{-3} \geq -3(16/13)^3/16 \geq -1/2$. Substituting this inequality in (5.16) yields (5.14).

Second, we prove (5.15). We have

$$(5.17) \quad \frac{w}{1+w} - w = -\frac{w^2}{1+w}.$$

Since $w \geq -|w| \geq -3/16$, $1+w \geq 13/16$, which upon substitution in (5.17) provides (5.15). $\square$

Next, we provide an overall coordination complexity bound to approximately solve problem (2.1).

THEOREM 5.6. *For a given relative accuracy $\varepsilon \in (0, 1]$, an ε -approximate solution to problem (2.1) can be computed using Algorithm $\mathcal{L}$ with inexact line searches in a total of*

$$N_\varepsilon = O(M(\varepsilon^{-2} \ln \varepsilon^{-1} + \ln M))$$

coordination steps.

Proof. We shall first bound the potential reduction obtained by Algorithm $\mathcal{L}$ when it moves from iterate x to the next iterate $x' = (1 - \tau)x + \tau\hat{x}$. Since $\tau < \tau_{max}$, we have that $\theta - f_m(x') \geq \theta - f_m + \tau(f_m - \hat{f}_m) > 0$. Thus, $\lambda(x') < \theta$ and then

$$\Phi(\theta(x'), x') = \min_{\theta} \Phi(\theta, x') \leq \Phi(\theta(x), x') = \phi_t(\tau).$$

This inequality, Lemma 5.4, and the stopping criterion of Algorithm $\mathcal{L}$ provide

$$(5.18) \quad \Phi(\theta(x'), x') \leq \Phi(\theta(x), x) - \frac{3t\alpha_1(1 - \alpha_2)\sigma^2}{128M^+}.$$

Next, we obtain a bound on the number of coordination steps N_σ performed within a single σ -scaling phase of Algorithm $\mathcal{L}$ with inexact line searches and then obtain a combined bound for all σ -scaling phases.

Let x, x'' denote the starting and final iterates, respectively, of a σ -scaling phase. For these two iterates, (5.18) provides

$$\Phi_t(\theta(x''), x'') \leq \Phi_t(\theta(x), x) - N_\sigma \frac{3t\alpha_1(1 - \alpha_2)\sigma^2}{128M^+}.$$

Now, $(1 - t)\ln\lambda_z^* \leq \Phi_t(\theta(x''), x'')$ by the left inequality of Lemma 5.3 and $\Phi_t(\theta(x), x) \leq (1 - t)\ln\lambda(x) + t\ln(e/t)$ by the right inequality. Hence,

$$\begin{aligned} N_\sigma \frac{3t\alpha_1(1 - \alpha_2)\sigma^2}{128M^+} &\leq (1 - t)\ln(\lambda(x)/\lambda_z^*) + t\ln(e/t), \\ N_\sigma &\leq \frac{43M^+}{\alpha_1(1 - \alpha_2)} \left(\frac{\ln(\lambda(x)/\lambda_z^*)}{t\sigma^2} + \frac{\ln(e/t)}{\sigma^2} \right). \end{aligned}$$

From this point, we can follow the analysis given in [25] to first bound N_σ from above by

$$N_\sigma = O(M\sigma^{-2} \ln \sigma^{-1}),$$

from where the claimed bound is obtained summing N_σ over all σ -scaling phases. $\square$

Appendix. Example. We define $\bar{\nu}$ as

$$(A.1) \quad \begin{aligned} \bar{\nu} &= \min \nu \\ \text{s.t. } f(x) &\leq \nu d, \\ x &\in B, \end{aligned}$$

and we consider the problem

$$(\bar{P}) \quad \begin{aligned} z^* &= \min z(x) \\ \text{s.t. } f(x) &\leq (1 + \varepsilon/3)\bar{\nu}d, \\ x &\in B, \end{aligned}$$

to which we apply the scheme given in the introduction. Problem (1.2) for this instance is

$$(A.2) \quad \begin{aligned} \lambda_z^* &= \min \lambda \\ \text{s.t. } z(x) &\leq \lambda z, \\ f(x) &\leq \lambda(1 + \varepsilon/3)\bar{\nu}d, \\ x &\in B. \end{aligned}$$

Let $x \in B$ and λ be a feasible solution of (A.2). Then, x and $\lambda(1 + \varepsilon/3)\bar{\nu}$ give a feasible solution of (A.1) so that $\bar{\nu} \leq \lambda(1 + \varepsilon/3)\bar{\nu}$. Thus, $\lambda \geq 1/(1 + \varepsilon/3)$ and consequently $\lambda_z^* \geq 1/(1 + \varepsilon/3)$ for all z .

In order to set up this example, we assume that $z(x)$ is bounded from above by α over B and we choose $z \geq (1 + \varepsilon/3)\alpha$. Let $\bar{x} \in B$ be an optimal solution of (A.1). Then, $\bar{x}$ and $1/(1 + \varepsilon/3)$ give a feasible solution of (A.2) so that $\lambda_z^* \leq 1/(1 + \varepsilon/3)$. Thus, we can conclude that $\lambda_z^* = 1/(1 + \varepsilon/3)$ for all $z \geq (1 + \varepsilon/3)\alpha$.

Now, we take $z \geq (1 + \varepsilon/3)\alpha$, $\lambda = (1 + \varepsilon)\lambda_z^*$, and a pair (x, ν) , which is an ε -approximate solution of (A.1). We have that $f(x) \leq \nu d \leq (1 + \varepsilon)\bar{\nu}d$ and then

$$f(x) \leq \frac{1 + \varepsilon}{1 + \varepsilon/3}(1 + \varepsilon/3)\bar{\nu}d = \lambda(1 + \varepsilon/3)\bar{\nu}d.$$

Also,

$$z(x) \leq \alpha \leq (1 + \varepsilon)\alpha = \frac{1 + \varepsilon}{1 + \varepsilon/3}(1 + \varepsilon/3)\alpha \leq \frac{1 + \varepsilon}{1 + \varepsilon/3}z = \lambda z.$$

Thus, for all $z \geq (1 + \varepsilon/3)\alpha$, the pair (x, λ) with $\lambda = (1 + \varepsilon)\lambda_z^*$ is an ε -approximate solution of (A.2) such that

$$1 < \lambda = \frac{1 + \varepsilon}{1 + \varepsilon/3} \leq 1 + \varepsilon.$$

Acknowledgment. The author gratefully acknowledges the valuable comments offered by two anonymous referees.

REFERENCES

- [1] R. K. AHUJA, T. L. MAGNANTI, AND J. B. ORLIN, *Network Flows*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
- [2] K. M. ANSTREICHER, *A standard form variant, and safeguarded linesearch, for the modified Karmarkar algorithm*, Math. Program., 47 (1990), pp. 337–351.
- [3] L. ARMIJO, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific J. Math., 16 (1966), pp. 1–3.
- [4] M. S. BAZARAA, H. D. SHERALI, AND C. M. SHETTY, *Nonlinear Programming: Theory and Algorithms*, John Wiley & Sons, New York, 1993.
- [5] L. K. FLEISCHER, *Approximating fractional multicommodity flow independent of the number of commodities*, SIAM J. Discrete Math., 13 (2000), pp. 505–520.
- [6] R. FLETCHER, *Practical Methods of Optimization*, 2nd ed., John Wiley, New York, 1987.
- [7] P. E. GILL, W. MURRAY, AND M. H. WRIGHT, *Practical Optimization*, Academic Press, New York, 1981.
- [8] P. E. GILL, W. MURRAY, M. A. SAUNDERS, AND M. H. WRIGHT, *A note on a sufficient-decrease criterion for a non-derivative step-length procedure*, Math. Program., 23 (1982), pp. 349–352.
- [9] A. V. GOLDBERG AND R. E. TARJAN, *Solving minimum cost flow problems by successive approximation*, Math. Oper. Res., 15 (1990), pp. 430–466.
- [10] A. V. GOLDBERG, J. D. OLDHAM, S. PLOTKIN, AND C. STEIN, *An Implementation of a Combinatorial Approximation Algorithm for Minimum-Cost Multicommodity Flow*, Lecture Notes in Comput. Sci. 1412, Springer, Berlin, 1998, pp. 338–352.
- [11] A. A. GOLDSTEIN, *Constructive Real Analysis*, Harper, New York, 1967.
- [12] M. D. GRIGORIADIS AND L. G. KHACHIYAN, *Fast approximation schemes for convex programs with many blocks and coupling constraints*, SIAM J. Optim., 4 (1994), pp. 86–107.
- [13] M. D. GRIGORIADIS AND L. G. KHACHIYAN, *Coordination complexity of parallel price-directive decomposition*, Math. Oper. Res., 21 (1996), pp. 321–340.
- [14] M. D. GRIGORIADIS AND L. G. KHACHIYAN, *Approximate minimum-cost multicommodity flows in $\tilde{O}(KNM\epsilon^{-2})$ time*, Math. Program., 75 (1996), pp. 477–482.
- [15] D. KARGER AND S. PLOTKIN, *Adding multiple cost constraints to combinatorial optimization problems with applications to multicommodity flows*, in Proceedings of the 27th Annual ACM Symposium on Theory of Computing, ACM, New York, 1995, pp. 18–25.
- [16] N. KARMARKAR, *A new polynomial time algorithm for linear programming*, Combinatorica, 4 (1986), pp. 373–395.
- [17] C. LEMARÉCHAL, *A View of Line Searches*, Lecture Notes in Control and Inform. Sci., Springer-Verlag, New York, 1981, pp. 59–78.
- [18] J. J. MORE AND D. J. THUENTE, *Line search algorithms with guaranteed sufficient decrease*, ACM Trans. Math. Software, 20 (1994), pp. 286–307.
- [19] W. MURRAY AND M. H. WRIGHT, *Line search procedure for the logarithmic barrier function*, SIAM J. Optim., 4 (1994), pp. 229–246.
- [20] S. A. PLOTKIN, D. B. SHMOYS, AND E. TARDOS, *Fast approximation algorithms for fractional packing and covering problems*, Math. Oper. Res., 20 (1995), pp. 257–301.
- [21] E. POLAK, *Computational Methods in Optimization*, Academic Press, New York, 1971.
- [22] F. A. POTRA AND Y. SHI, *Efficient line search algorithm for unconstrained optimization*, J. Optim. Theory Appl., 85 (1995), pp. 677–704.
- [23] M. J. D. POWELL, *A View of Unconstrained Optimization*, Report CSS 14, Atomic Energy Research Establishment, Harwell, UK, 1975.
- [24] J. VILLAVICENCIO AND M. D. GRIGORIADIS, *Approximate structured optimization by cyclic block-coordinate descent*, in Applied Mathematics and Parallel Computing. Festschrift for Klaus Ritter, H. Fischer, B. Riedmüller, and S. Schäffler, eds., Physica-Verlag, Heidelberg, 1996, pp. 359–371.
- [25] J. VILLAVICENCIO AND M. D. GRIGORIADIS, *Approximate Lagrangian decomposition with a modified Karmarkar logarithmic potential*, in Network Optimization, Lectures Notes in Econom. and Math. Systems 450, Springer, Berlin, 1997, pp. 471–485.
- [26] J. VILLAVICENCIO, *Approximation Algorithms for Min-Max and Max-Min Structured Optimization*, Ph.D. thesis, RUTCOR, Rutgers University, New Brunswick, NJ, 1999.
- [27] P. WOLFE, *Convergence conditions for ascent methods*, SIAM Rev., 11 (1969), pp. 226–235.

SCALARIZING FUNCTIONS FOR GENERATING THE WEAKLY EFFICIENT SOLUTION SET IN CONVEX MULTIOBJECTIVE PROBLEMS*

DINH THE LUC[†], THAI QUYNH PHONG[‡], AND MICHEL VOLLE[†]

Abstract. A method that uses normal directions of convex sets is proposed to construct a sequence of scalarizing functions which generates all weakly efficient solutions of a convex multi-objective optimization problem. The convergence of the method is proven and some small-scale numerical examples are given.

Key words. convex multiobjective problem, efficient solution, normal direction, scalarization

AMS subject classification. 90C31

DOI. 10.1137/040603097

1. Introduction. Let $\mathbb{R}^n$ denote the n -dimensional Euclidean space and $\mathbb{R}_+^n$ its positive orthant. For two vectors a and b of $\mathbb{R}^n$, we shall write $a \geq b$ if $a - b \in \mathbb{R}_+^n$ and $a > b$ if $a - b \in \text{int}\mathbb{R}_+^n$, where $\text{int}\mathbb{R}_+^n$ is the interior of $\mathbb{R}_+^n$. Given a nonempty set $A \subseteq \mathbb{R}^n$, a point $x \in A$ is said to be a weakly efficient (or weakly Pareto) point of A if there is no $y \in A$ verifying $y > x$. The set of all weakly efficient points of A is denoted by $WMax(A)$. Let $f : \mathbb{R}^m \rightarrow \mathbb{R}^n$ be a vector function and $X \subseteq \mathbb{R}^m$ a nonempty set. The multiobjective optimization problem associated with f and X is written as

$$\begin{aligned} \text{(VP)} \quad & \text{Max } f(x) \\ & \text{s.t. } x \in X, \end{aligned}$$

which amounts to finding a point $x_0 \in X$, called a weakly efficient solution, such that $f(x_0) \in WMax(f(X))$. The set of all weakly efficient solutions of problem (VP) is denoted by $WS(f, X)$.

In this paper we are interested in finding the whole solution set $WS(f, X)$. This problem is important in applications, especially in multicriteria decision making where the decision maker wishes to have in hand all nondominated (efficient) solutions in order to make his decision based on his own judgment or on eventually postoptimal analysis of these solutions (see [4], [19], [21] for more details on multicriteria decision making). To date there exists a large variety of methods for solving problem (VP), starting with the classical ones (the weighting method, the ϵ -constraint method, the goal programming method, etc.) to more recent ones, such as the outer approximation method, the normal-boundary intersection method, and the normal cone method (see [1], [2], [3], [6], [9], [10], [11], [15], [16], [17], [18], [21] and the references therein). Most of these methods allow us to find one or a few solutions which are then presented to the decision maker, who decides to accept it as a final solution or to continue the search for a more preferred one. Some of them are aimed at generating the whole

*Received by the editors January 12, 2004; accepted for publication (in revised form) November 15, 2004; published electronically June 30, 2005.

<http://www.siam.org/journals/siopt/15-4/60309.html>

[†]University of Avignon, 33 rue Louis Pasteur, Avignon, France (dtluc@univ-avignon.fr, michel.volle@univ-avignon.fr).

[‡]University of Danang, Mathematics, 17 Le Duan, Danang, Vietnam (phong.thaiquynh@univ-avignon.fr).

solution set, but except for the case of linear problems of medium size (see [2], [10], [22]), their numerical implementation remains very hard work. To our knowledge, very few papers exist which fully describe numerical algorithms for generating the set $WS(f, X)$ for nonlinear problems. Among them we find the papers [6] and [16] quite interesting. The work [16] deals with bicriteria problems and uses the weighting method to trace the efficient curve, while [6] treats a general case when the problem is nonlinear and even not necessarily convex. We will discuss the latter paper in more detail in subsection 5.4. The method that we are going to develop in the present paper is based on the following theoretical fact (see [12], [13], [14]): there exists a so-called scalarizing function $g : \mathbb{R}^n \rightarrow \mathbb{R}$ which is continuous and weakly monotone in the sense that $g(a) > g(b)$ for $a > b$, such that the solution set of (VP) is exactly the solution set $S(g \circ f, X)$ of the scalar problem

$$(P_g) \quad \begin{array}{ll} \max & g \circ f(x) \\ \text{s.t.} & x \in X. \end{array}$$

This result, being useful in the study of the solution structure and solution stability of (VP), remains difficult to implement numerically even in the case of linear problems. In our paper, to find the solution set of (VP) we shall construct a sequence of scalarizing functions g_k which converge to g , and by solving scalar problem (P_{g_k}) we obtain an approximation of the set $WS(f, X)$. The construction of g_k is based on an outer approximation technique for convex sets which exploits normal directions to the outcome set $f(X)$.

The paper is organized as follows. In section 2, we briefly recall the definition of normal directions to convex sets and express optimality conditions of weakly efficient points in terms of normal directions. In section 3 a method of outwardly approximating convex sets by polytopes is presented. Section 4 is devoted to the construction of a sequence of scalarizing functions. In section 5, the results of the previous sections are applied to construct scalarizing functions which solve problem (VP) when it is convex. The convergence of the method is proven and its use and implementation are addressed. Some small-scale numerical examples are provided in section 6 and a final remark is given in the last section.

2. Normal directions and weakly efficient points. Let $A \subseteq \mathbb{R}^n$ be a nonempty convex and closed set and let $a \in A$. A vector $v \in \mathbb{R}^n$ is said to be a normal direction to A at a if

$$\langle v, x - a \rangle \leq 0 \quad \text{for all } x \in A.$$

It is known that if $y \in \mathbb{R}^n$ and if $a \in A$ is the projection of y on A , that is,

$$\|y - a\| = \inf_{x \in A} \|y - x\|,$$

then the vector $y - a$ is a normal direction to A at a . The set of all normal directions to A at a forms a closed convex cone which is denoted by $N(A, a)$ and called the normal cone to A at a . Optimality conditions at a weakly efficient point of a convex set can be expressed in terms of normal directions as follows.

LEMMA 2.1. *Let A be a nonempty closed convex set. Then a vector $a \in A$ is a weakly efficient point of A if and only if $N(A, a) \cap \mathbb{R}_+^n \neq \{0\}$.*

Proof. This lemma is immediate from [14, p. 91, Theorem 2.10]. $\square$

We shall make use of the following notation: for $A \subseteq \mathbb{R}_+^n$, $A^\diamond = (A - \mathbb{R}_+^n) \cap \mathbb{R}_+^n$. It follows that $WMax(A) = WMax(A^\diamond) \cap A$.

PROPOSITION 2.2. *Let $A \subseteq \mathbb{R}_+^n$ be convex and compact with $A = A^\diamond$. Then*

- (i) *the projection of any point of $\mathbb{R}_+^n \setminus A$ on A is a weakly efficient point of A ;*
- (ii) *for $a \in A$ and $v \in \mathbb{R}_+^n, v \neq 0$, the point $a + t_0 v$, where $t_0 = \sup\{t \geq 0 : a + tv \in A\}$, is a weakly efficient point of A .*

Proof. Let $y \in \mathbb{R}_+^n \setminus A$ and let $a \in A$ be the projection of y on A . Then $y - a \in N(A, a)$ and $y - a \neq 0$. We show that $y - a \geq 0$. Indeed, if not, say the first component of this vector is strictly negative, $(y - a)_1 < 0$; then the vector b , which is obtained from a by replacing the first component by the corresponding component of y , belongs to A because $A = A^\diamond$. Moreover,

$$\|y - b\| = \left(\sum_{i=2}^n (y_i - a_i)^2 \right)^{1/2} < \left(\sum_{i=1}^n (y_i - a_i)^2 \right)^{1/2} = \|y - a\|.$$

This contradicts the choice of a . By this, $y - a \in (N(A, a) \cap \mathbb{R}_+^n) \setminus \{0\}$, and in view of Lemma 2.1, a is a weakly efficient point of A .

Now, let $a \in A, v \in \mathbb{R}_+^n, v \neq 0$ and let $t_0 = \max\{t : a + tv \in A\}$. If $a + t_0 v$ is not a weakly efficient point of A , then there is some $b \in A$ such that $a + t_0 v \in b - \text{int}\mathbb{R}_+^n$. There is also some $\epsilon > 0$ such that $a + (t_0 + \epsilon)v \in b - \text{int}\mathbb{R}_+^n$. Then $a + (t_0 + \epsilon)v \in A$, which contradicts the choice of t_0 . $\square$

3. Outer approximation of a convex set. Given two nonempty subsets A_1 and A_2 of $\mathbb{R}^n$, the Hausdorff distance between them is defined by

$$h(A_1, A_2) = \inf\{t > 0 : A_1 \subseteq A_2 + tB_n, A_2 \subseteq A_1 + tB_n\},$$

where B_n denotes the closed unit ball in $\mathbb{R}^n$. A sequence of compact sets $\{A_k\}_{k=1}^\infty \subset \mathbb{R}^n$ is said to converge to a compact set A if $\lim_{k \rightarrow \infty} h(A_k, A) = 0$.

When A is a convex, compact set and M is a polytope (a bounded polyhedral convex set) that contains A , the Hausdorff distance between A and M can easily be calculated.

LEMMA 3.1. *Let $v_1, \dots, v_k$ be the vertices of a polytope M and let $A \subseteq M$ be nonempty, convex, and compact. Then*

$$h(A, M) = \max\{d(v_i, A) : i = 1, \dots, k\},$$

where $d(v_i, A)$ is the distance from v_i to A .

Proof. It follows from the definition that $v_i \in A + d(v_i, A)B_n$ for $i = 1, \dots, k$. Denote by $d = \max\{d(v_i, A) : i = 1, \dots, k\}$. Then $M = \text{conv}\{v_1, \dots, v_k\} \subseteq A + dB_n$, where $\text{conv}\{v_1, \dots, v_k\}$ stands for the convex hull of $v_1, \dots, v_k$. This implies that $h(A, M) \leq d$. Moreover, for a vertex v_i with $d(v_i, A) = d$, one has $v_i \notin A + tB_n$ whenever $t < d$. Consequently, $h(A, M) = d$ and the proof is complete. $\square$

We shall now present a method of constructing a sequence of polytopes which outwardly approximate a given convex and compact set. See [7] for more on approximations of convex sets. Let M be any polytope that contains a convex, compact set A . By setting $M_1 = M$, we suppose that $M_2, \dots, M_k$ have already been constructed which verify the inclusion $A \subseteq M_{i+1} \subseteq M_i$ for $i = 1, \dots, k-1$. Let $\{v_1, \dots, v_l\}$ be the set of all vertices of M_k which do not belong to A . Let a_j be the projections of v_j on A , $1 \leq j \leq l$. We define

$$M_{k+1} = M_k \cap \{x \in \mathbb{R}^n : \langle v_j - a_j, x \rangle \leq \langle v_j - a_j, a_j \rangle, 1 \leq j \leq l\}.$$

Here are some properties of the polytopes M_k .

PROPOSITION 3.2. *The sequence of polytopes $\{M_k\}_{k=1}^\infty$ verifies the following properties:*

- (i) $A \subseteq M_{k+1} \subseteq M_k$ for $k = 1, 2, \dots$.
- (ii) $\lim_{k \rightarrow \infty} M_k = A$.
- (iii) *When A is a polytope, there is $k \geq 1$ such that $M_k = A$.*

Proof. The inclusions of the first assertion are obvious. Let us prove (ii). Denote by p_k a vertex of M_k such that $d(p_k, A) = h(A, M_k)$ according to Lemma 3.1. Since $M_k \subseteq M$, without loss of generality one may assume that $\{p_k\}_{k=1}^\infty$ converges to some limit p_0 . If this limit belongs to A , we are done. Otherwise, let a_k and a_0 be the projections of p_k and p_0 , respectively, on A . By continuity of projection $\{a_k\}_{k=1}^\infty$ converges to a_0 . Hence, for k sufficiently large we have

$$\langle p_k - a_k, p_{k+1} - a_k \rangle \geq \|p_0 - a_0\|^2 / 2 > 0,$$

which implies that

$$\langle p_k - a_k, p_{k+1} \rangle > \langle p_k - a_k, a_k \rangle$$

and yields a contradiction $p_{k+1} \notin M_{k+1}$.

Assume now that A is a polytope. By passing to a space of less dimension, one may suppose that A has a nonempty interior. By the construction, if some vertex of M_k does not belong to A , then the face of A that contains the projection of this vertex on A as a relative interior point must be in the boundary of M_{k+1} and outside the boundary of M_k . Since the number of faces of the boundary of A is finite, there is some k such that the boundary of A is included in the boundary of M_k . The set M_k being a polytope, we conclude $A = M_k$. $\square$

4. Scalarizing functions. Let $\mathcal{A}$ denote the family of all compact sets A in $\mathbb{R}_+^n$ with nonempty interior and with $A^\diamond$ convex. Let Λ denote the standard n -simplex consisting of vectors $\lambda = (\lambda_1, \dots, \lambda_n) \in \mathbb{R}_+^n$ with $\sum_{i=1}^n \lambda_i = 1$. Elements of this set are sometimes called weight vectors which reflect the degree of importance of each component of f in the scalarized problem (P_g) , where $g(y) = \langle \lambda, y \rangle$. For $A \in \mathcal{A}$ we also denote $e_{A^\diamond} := \max\{t > 0 : (t, \dots, t) \in A^\diamond\}$ and define

$$s_A(\lambda) = \max_{a \in A} \langle \lambda, a \rangle,$$

$$g_A(y) = \max_{\lambda \in \Lambda} \frac{\langle \lambda, y \rangle}{s_A(\lambda)}.$$

We note that the first function is the support function of the set A and the second one is well defined because A is compact with nonempty interior in $\mathbb{R}_+^n$, which implies that the denominator $s_A(\lambda)$ is strictly positive.

PROPOSITION 4.1. *Let $A, A_1, A_2 \in \mathcal{A}$. The following assertions hold:*

- (i) g_A is convex, homogeneous, continuous, and weakly monotone.
- (ii) $g_A = g_{A^\diamond}$.
- (iii) $y \in A^\diamond$ if and only if $y \in \mathbb{R}_+^n$ and $g_A(y) \leq 1$.
- (iv) $g_{A_1}(y) \leq g_{A_2}(y)$ for every $y \in \mathbb{R}_+^n$ if and only if $A_2^\diamond \subseteq A_1^\diamond$.
- (v) If $A_1 \subseteq A_2 + \varepsilon B_n$, then for each y , $g_{A_1}(y) \geq g_{A_2}(y) - \varepsilon \|y\| / (e_{A_2^\diamond})^2$. In particular, if $\{A_k\}_{k=1}^\infty \subset \mathcal{A}$ converges to $A \in \mathcal{A}$, then

$$\lim_{k \rightarrow \infty} g_{A_k}(y) = g_A(y) \text{ for each } y \in \mathbb{R}_+^n.$$

Proof. The first assertion is obvious because g_A is a max-function of a compact family of linear, weakly monotone functions. The second assertion follows from the fact that for each $\lambda \in \Lambda$, one has $s_A(\lambda) = s_{A^\diamond}(\lambda)$. For the third assertion if $y \in A^\diamond$, then $y \in \mathbb{R}_+^n$ and there is some $x \in A$ with $y \leq x$. Hence $\langle \lambda, y \rangle \leq s_A(\lambda)$ for each $\lambda \in \Lambda$ and $g_A(y) \leq 1$. Conversely, if $y \in \mathbb{R}_+^n$ and $y \notin A^\diamond$, then by separation there is some $\lambda \in \Lambda$ such that $\langle \lambda, y \rangle > s_{A^\diamond}(\lambda)$. This and (ii) show that $g_A(y) > 1$. For the fourth assertion, assume that $A_2^\diamond \subseteq A_1^\diamond$. Then $s_{A_1^\diamond}(\lambda) \geq s_{A_2^\diamond}(\lambda)$. Hence for each $y \in \mathbb{R}_+^n$, one has

$$g_{A_2^\diamond}(y) \geq g_{A_1^\diamond}(y),$$

which by (ii) implies

$$g_{A_2}(y) \geq g_{A_1}(y).$$

Conversely, assume $A_2^\diamond \not\subseteq A_1^\diamond$ and let $y \in A_2^\diamond \setminus A_1^\diamond$. There is $\lambda \in \Lambda$ such that

$$\langle \lambda, y \rangle > \langle \lambda, a \rangle \text{ for all } a \in A_1^\diamond.$$

Since A_1 is compact, we derive $g_{A_1^\diamond}(y) > 1$, which together with (ii) and (iii) yields $g_{A_1^\diamond}(y) > g_{A_2^\diamond}(y)$.

We proceed to proving (v). Note that for each $\lambda \in \Lambda$, $s_{A_2+\varepsilon B_n}(\lambda) = s_{A_2}(\lambda) + \varepsilon \|\lambda\|$ and $e_{A_2^\diamond} \leq s_{A_2}(\lambda)$. Consequently,

$$\begin{aligned} \frac{\langle \lambda, y \rangle}{s_{A_1}(\lambda)} &\geq \frac{\langle \lambda, y \rangle}{s_{A_2+\varepsilon B_n}(\lambda)} \geq \frac{\langle \lambda, y \rangle}{s_{A_2}(\lambda) + \varepsilon} \\ &\geq \frac{\langle \lambda, y \rangle}{s_{A_2}(\lambda)} - \frac{\varepsilon \langle \lambda, y \rangle}{s_{A_2}(\lambda)(s_{A_2}(\lambda) + \varepsilon)} \geq \frac{\langle \lambda, y \rangle}{s_{A_2}(\lambda)} - \frac{\varepsilon \|y\|}{(e_{A_2^\diamond})^2}, \end{aligned}$$

which implies

$$g_{A_1}(y) \geq g_{A_2}(y) - \varepsilon \|y\| / (e_{A_2^\diamond})^2.$$

Finally, if $h(A_k, A) \rightarrow 0$ as $k \rightarrow \infty$, then for every $\varepsilon > 0$, there is an index k_0 such that

$$A_k \subseteq A + \varepsilon B_n \subseteq A_k + 2\varepsilon B_n \text{ for } k \geq k_0.$$

We deduce for each $y \in \mathbb{R}^n$ that

$$g_{A_k}(y) \geq g_A(y) - \varepsilon \frac{\|y\|}{(e_{A^\diamond})^2} \geq g_{A_k}(y) - \varepsilon \frac{\|y\|}{(e_{A_k^\diamond})^2} - \varepsilon \frac{\|y\|}{(e_{A^\diamond})^2}.$$

It is evident that $e_{A_k^\diamond} \rightarrow e_{A^\diamond}$ as $k \rightarrow \infty$. The above inequalities yield $\lim_{k \rightarrow \infty} g_{A_k}(y) = g_A(y)$. $\square$

5. Application to problem (VP).

5.1. Construction of polytopes approaching $f(X)$. Throughout this section we shall make the following assumption:

(H) $f(X)$ is compact and contained in $\text{int}\mathbb{R}_+^n$ such that $[f(X)]^\diamond$ is convex.

The convexity of the set $[f(X)]^\diamond$ is guaranteed, for instance, when X is a convex, compact set and f is a concave function; the requirement $f(X) \subset \text{int}\mathbb{R}_+^n$ can be achieved by considering the function $f(\cdot) - a$ instead of f when f is bounded from below and $a = (a_1, \dots, a_n)$ is some vector with $f_i(x) > a_i$ for $i = 1, \dots, n$ and $x \in X$. Let us now apply the method developed in section 3 to construct a sequence of polytopes $\{A_k\}_{k=1}^\infty$ which approaches the set $f(X)$ (see also [2] for a similar construction). To this end, consider the scalar problems

$$\begin{aligned} \max \quad & f_i(x) \\ \text{s.t.} \quad & x \in X \end{aligned}$$

for $i = 1, \dots, n$. Denote by q the vector $(q_1, \dots, q_n)$, where $q_1, \dots, q_n$ are upper bounds of optimal values of these problems. Set

$$A_1 = [0, q_1] \times \dots \times [0, q_n].$$

This is a polytope which contains $[f(X)]^\diamond$. Assuming A_k with $k \geq 1$ has already been defined, we now construct A_{k+1} as follows. Denote by V_k the set of all vertices of A_k . For each $v \in V_k$ solve the problem

$$\begin{aligned} (P_v) \quad \min \quad & \|v - y\|^2 \\ \text{s.t.} \quad & y \in [f(X)]^\diamond, \end{aligned}$$

which actually finds the projection of v on $[f(X)]^\diamond$.

Let y_v be the optimal solution of this problem. Set

$$p_v := \frac{v - y_v}{\sum_{i=1}^n (v - y_v)_i},$$

where $(v - y_v)_i$ is the i th coordinate of the vector $(v - y_v)$, whenever the latter vector is nonzero. It follows from the proof of Proposition 2.2 that p_v is a normalized normal vector, i.e., $p_v \in N([f(X)]^\diamond, y_v) \cap \Lambda$. Define

$$A_{k+1} := A_k \cap \{y \in \mathbb{R}^n : \langle p_v, y \rangle \leq \langle p_v, y_v \rangle, v \in V_k\}.$$

Some properties of A_k are collected in the next result.

PROPOSITION 5.1. *Under assumption (H) the following assertions hold:*

(i) *The set A_k is a polytope and verifies*

$$[f(X)]^\diamond \subseteq A_{k+1} \subseteq A_k.$$

(ii) *If, for some $k \geq 1$, $V_k \subseteq [f(X)]^\diamond$, then $A_k = [f(X)]^\diamond$.*

(iii) *If $[f(X)]^\diamond$ is a polytope, then there is some $k \geq 1$ such that $A_k = [f(X)]^\diamond$.*

(iv) $\lim_{k \rightarrow \infty} A_k = [f(X)]^\diamond$.

Proof. The first and the two last assertions are immediate from Proposition 3.2. For (ii), if all vertices of A_k belong to $[f(X)]^\diamond$, then, the latter set being convex, we have the inclusion $A_k \subseteq [f(X)]^\diamond$, and by (i) equality follows. $\square$

Let us illustrate the approach of this subsection by the following simple example (see Figure 5.1). Let X be the positive quarter of the unit disc in the plan $\mathbb{R}^2$ and let f be the identity function from $\mathbb{R}^2$ to itself. Then the maximal value q_i of f_i on X is equal to one, $i = 1, 2$. The first polytope A_1 approaching $f(X)$ is the unit square $[0, 1] \times [0, 1]$ whose vertices are $v_0 = (0, 0)$, $v_1 = (1, 0)$, $v_2 = (0, 1)$, and $v_3 = (1, 1)$. Only

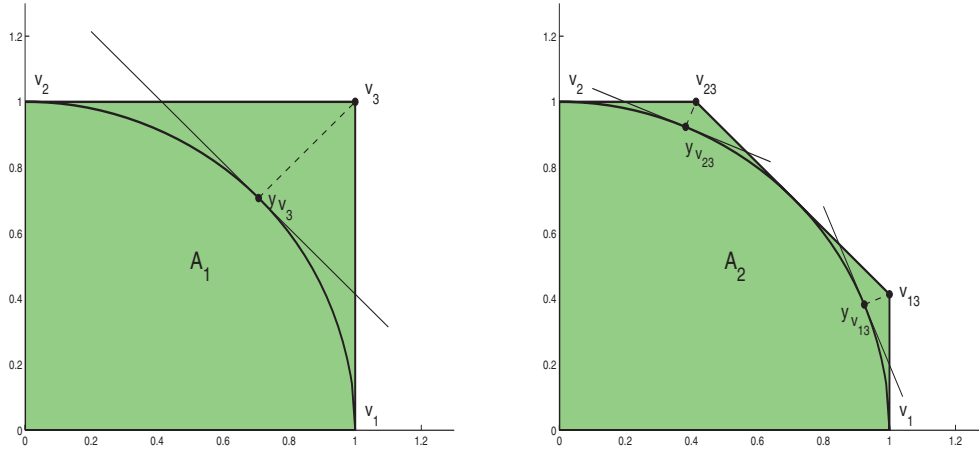


FIG. 5.1.

the vertex v_3 of A_1 is outside of $f(X)$. Its projection on $f(X)$ is $y_{v_3} = (1/\sqrt{2}, 1/\sqrt{2})$. The new polytope A_2 approaching $f(X)$ is given by the intersection of A_1 with the half-plan

$$\frac{1}{\sqrt{2}}x_1 + \frac{1}{\sqrt{2}}x_2 \leq 1.$$

Its vertices are v_0, v_1, v_2 of the polytope A_1 and two new ones $v_{13} = (1, \sqrt{2} - 1)$ and $v_{23} = (\sqrt{2} - 1, 1)$.

5.2. Construction of scalarizing functions for problem (VP). We return to the question of finding the solution set $WS(f, X)$ of (VP).

COROLLARY 5.2. *Under assumption (H), we have*

$$WS(f, X) = S(g_{f(X)} \circ f, X).$$

Proof. According to Proposition 4.1, the function $g_{f(X)}$ is weakly monotone. It follows that

$$S(g_{f(X)} \circ f, X) \subseteq WS(f, X).$$

For the converse inclusion, let x be a weakly efficient solution of (VP). By scalarization, there is some $\lambda \in \Lambda$ such that $x \in S(\lambda \circ f, X)$. This in particular implies that

$$g_{f(X)}(f(x)) \geq 1.$$

On the other hand, for any $x' \in X$ one has

$$g_{f(X)}(f(x')) \leq 1.$$

Consequently, x maximizes $g_{f(X)} \circ f$ over X . The proof is complete. $\square$

Since the scalarizing function $g_{f(X)}$ is difficult to be computed because its determination is equivalent to the solution of problem (VP), we wish to use the sequence of polytopes that we have already obtained in the first subsection to construct a sequence of scalarizing functions that approach the function $g_{f(X)}$.

Denote by N_1 the set of the canonical basis vectors $e_1 = (1, 0, \dots, 0), \dots, e_n = (0, \dots, 1)$. Assuming N_k has already been determined, we form N_{k+1} by adding to N_k the vectors $p_v, v \in V_k$, that are obtained by solving (P_v) . Then the set A_k is the solution set to the system of linear inequalities:

$$\begin{aligned} \langle p, y \rangle &\leq \rho_p, \quad p \in N_k, \\ -\langle e_j, y \rangle &\leq 0, \quad j = 1, \dots, n, \end{aligned}$$

where

$$\rho_p = \begin{cases} \langle p, q \rangle & \text{if } p \in N_1, \\ \langle p, y_v \rangle & \text{if } p = p_v, v \in V_i, i = 2, \dots, k-1. \end{cases}$$

For each $y \in \mathbb{R}^n$ and $k = 1, 2, \dots$ we define

$$g_k(y) = \max_{\lambda \in N_k} \frac{\langle \lambda, y \rangle}{s_{V_k}(\lambda)}$$

and remember that

$$g_{A_k}(y) = \max_{\lambda \in \Lambda} \frac{\langle \lambda, y \rangle}{s_{A_k}(\lambda)}.$$

A link between these functions is seen as follows.

LEMMA 5.3. *For every $y \in \mathbb{R}^n$ one has $g_k(y) \leq g_{A_k}(y)$, and equality holds when $y \in \mathbb{R}_+^n$.*

Proof. Because $N_k \subseteq \Lambda$ and $s_{V_k}(\lambda) = s_{A_k}(\lambda)$, the inequality of the lemma is immediate. Further, assume that $y \geq 0$. We have $\langle \lambda, y \rangle \geq 0$ for every $\lambda \in \Lambda$. Denote $t = g_k(y)$. If $t = 0$, then $y = 0$ and $g_{A_k}(y) = 0$. If $t > 0$, one has

$$\langle \lambda, y/t \rangle \leq s_{V_k}(\lambda),$$

which implies that $y/t \in A_k$. Since A_k is a convex set, we deduce for each $\lambda \in \Lambda$ that

$$\langle \lambda, y/t \rangle \leq s_{A_k}(\lambda) = s_{V_k}(\lambda).$$

Consequently, $g_{A_k}(y) \leq t$, and equality follows. $\square$

COROLLARY 5.4. *Under assumption (H), for every $y \in \mathbb{R}_+^n$ one has*

$$\lim_{k \rightarrow \infty} g_k(y) = g_{f(X)}(y).$$

Proof. This combines the above lemma with Propositions 4.1 and 5.1. $\square$

With the help of scalarizing functions g_k we obtain the following scalar problem:

$$\begin{aligned} (P_k) \quad & \max \quad g_k \circ f(x) \\ & \text{s.t. } x \in X. \end{aligned}$$

The following result gives us a relationship between the solution set of (P_k) and the solution set of (VP).

THEOREM 5.5. *Under assumption (H), the following assertions hold:*

- (i) $S(g_k \circ f, X) = \{x \in X : g_k(f(x)) = 1\}$.
- (ii) $S(g_k \circ f, X) \subseteq WS(f, X)$.
- (iii) *If $[f(X)]^\diamond$ is a polyhedron, then there is some $k \geq 1$ such that $S(g_k \circ f, X) = WS(g, X)$.*
- (iv) $h(A_k, [f(X)]^\diamond) = \max\{\|v - y_v\| : v \in V_k\}$.
- (v) $\lim_{k \rightarrow \infty} \max\{\|v - y_v\| : v \in V_k\} = 0$.
- (vi) $h(WMax(A_k), WMax([f(X)]^\diamond)) = \max\{\|v - y_v\| : v \in V_k\}$.

Proof. For the first assertion, let $x \in X$. Then as $f(x) \in f(X) \subseteq A_k$, we obtain

$$g_k(f(x)) = \max_{\lambda \in N_k} \frac{\langle \lambda, f(x) \rangle}{s_{V_k}(\lambda)} \leq 1.$$

By choosing x to maximize f_1 over X , we see that $g_k(f(x)) = 1$. Hence the optimal value of problem (P_k) is exactly 1. The second assertion is obtained from Proposition 4.1(i). The third assertion is deduced from Proposition 5.1. The fourth assertion follows from Lemma 3.1. The fifth assertion follows from the fourth assertion and from Proposition 5.1.

Let us prove (vi). Let $a \in WMax(A_k)$. If $a \in [f(X)]^\diamond$, then it is a weakly efficient point of this latter set. If $a \notin [f(X)]^\diamond$, then by Proposition 2.2, the projection y of a on $[f(X)]^\diamond$ is a weakly efficient point of $[f(X)]^\diamond$. Moreover, $\|a - y\| \leq h(A_k, [f(X)]^\diamond)$. Hence

$$WMax(A_k) \subseteq WMax([f(X)]^\diamond) + h(A_k, [f(X)]^\diamond)B_n.$$

Conversely, let $y \in WMax([f(X)]^\diamond)$. If it is a weakly efficient point of A_k , then $d(y, WMax(A_k)) = 0$. If not, then, as $y \in A_k$, there is some $a \in A_k$ such that $y \in a - \text{int}\mathbb{R}_+^n$. Since $[f(X)]^\diamond$ is convex, there is a normal direction, say v , to this set at y such that $v \in \mathbb{R}_+^n, v \neq 0$. In view of Proposition 2.2, there is $t_0 > 0$ such that $y + t_0v$ is a weakly efficient point of A_k .

It follows that y is the projection of $y + t_0v$ on $[f(X)]^\diamond$ and hence $d(y, WMax(A_k)) \leq \|(y + t_0v) - y\| \leq h(A_k, [f(X)]^\diamond)$. By this we derive

$$WMax([f(X)]^\diamond) \subseteq WMax(A_k) + h(A_k, [f(X)]^\diamond)B_n.$$

Consequently,

$$h(WMax(A_k), WMax([f(X)]^\diamond)) \leq \max\{\|v - y_v\| : v \in V_k\}.$$

On the other hand, since $WMax(A_k)$ and $WMax([f(X)]^\diamond)$ are in the boundaries of A_k and $WMax([f(X)]^\diamond)$, respectively, we also have

$$h(WMax(A_k), WMax([f(X)]^\diamond)) \geq h(A_k, [f(X)]^\diamond).$$

This and assertion (iv) complete the proof. $\square$

It is important to note that problem (P_k) is equivalent to solving problems (P_λ) with $\lambda \in N_k$, and the solution set of (P_k) is exactly the solution set $S(N_k \circ f, X) := \cup_{\lambda \in N_k} S(\lambda \circ f, X)$.

5.3. Solving (VP). The results of the two above subsections lead to the following method for solving problem (VP). In the remainder of this section we assume that X is compact and f is continuous.

Step 1 (initialization). Choose a small $\epsilon \geq 0$. Find a_i and $b_i \in R$ such that $a_i \geq f_i(x) \geq b_i$ for each $x \in X$ and $i = 1, \dots, n$. Set $k = 1$ and

$$\begin{aligned} f(x) &= f(x) - (b_1, \dots, b_n), \\ q &= (a_1, \dots, a_n) - (b_1, \dots, b_n), \\ N_k &= \{e_1, \dots, e_n\}, \\ \rho_p &= \langle p, q \rangle \text{ for } p \in N_k, \\ V_k &= \{v = (v_1, \dots, v_n) : v_i \in \{0, q_i\}, i = 1, \dots, n\}. \end{aligned}$$

Step 2. Solve problem (P_v) ,

$$\min_{y \in [f(X)]^\diamond} \|v - y\|^2,$$

and let y_v be the optimal solution for each $v \in V_k$.

Step 3. Set $V_k^+ = \{v \in V_k : \|v - y_v\| > \epsilon\}$.

If $V_k^+ = \emptyset$, then stop. The set

$$S_\epsilon := \cup_{p \in N_k} \{x \in X : \langle p, f(x) \rangle \geq \rho_p - \epsilon\}$$

is taken as the optimal solution set, while the set

$$\bar{S}_\epsilon := \cup_{p \in N_k} \{y \in f(X) : \langle p, y \rangle \geq \rho_p - \epsilon\}$$

is taken as the optimal value set of (VP).

Otherwise, go to the next step.

Step 4. For each $v \in V_k^+$ set

$$\begin{aligned} p_v &= (v - y_v) / \sum_{i=1}^n (v - y_v)_i, \\ \rho_{p_v} &= \langle p_v, y_v \rangle, \\ N_{k+1} &= N_k \cup \{p_v : v \in V_k^+\}. \end{aligned}$$

Find the vertex set V_{k+1} of the polytope A_{k+1} , which is given by the system of linear inequalities

$$\langle p, y \rangle \leq \rho_p, \quad p \in N_{k+1}, \quad \text{and } y \geq 0.$$

Set $k = k + 1$ and return to Step 2.

The main characteristics of the above method are summarized as follows.

COROLLARY 5.6. Assume (H) and that X is a nonempty compact set. Then

- (i) for a given $\epsilon > 0$, the algorithm terminates after a finite number of iterations;
- (ii) at the final iteration (when $V_k^+ = \emptyset$), one has $WS(f, X) \subseteq S_\epsilon$ and

$$\lim_{\epsilon \rightarrow 0} h(S_\epsilon, WS(f, X)) = 0;$$

- (iii) when problem (VP) is linear, with $\epsilon = 0$ the algorithm terminates after a finite number of iterations and produces the entire set of weakly efficient solutions.

Proof. According to Theorem 5.5(v), for k sufficiently large, V_k^+ is empty, and hence the algorithm terminates at this iteration. To prove (ii), assume that $V_k^+ = \emptyset$. It follows from Theorem 5.5 that $h(WMin(A_k), WMin([f(X)]^\diamond)) \leq \epsilon$. By the construction of A_k one has $WMin(A_k) = \cup_{p \in N_k} \{y \in A_k : \langle p, y \rangle = \rho_p\}$. Consequently, if $x \in WS(f, X)$, then $f(x) \in WMin([f(X)]^\diamond)$ and there exists $a \in WMin(A_k)$ such that $\|f(x) - a\| \leq \epsilon$. Let $p \in N_k$ be such that $\langle p, a \rangle = \rho_p$. Then $\langle p, f(x) \rangle = \langle p, a \rangle + \langle p, f(x) - a \rangle \geq \rho_p - \epsilon$, which shows that $x \in S_\epsilon$. To prove the limit of (ii), suppose to the contrary that for some sequence of positive numbers ϵ_n converging to 0, there exist $x_n \in S_{\epsilon_n}$ with the distance $d(x_n, WS(f, X)) \geq \epsilon_0 > 0$. Let $p_n \in \Lambda$ verify $\langle p_n, f(x_n) \rangle \geq \rho_{p_n} - \epsilon_n$. Without loss of generality we may assume that p_n converges to some $p_0 \in \Lambda$ and that x_n converges to $x_0 \in X$. Since f is continuous, and by the definition of ρ_p , we derive $\rho_{p_0} = \max_{x \in X} \langle p_0, f(x) \rangle$. Consequently, $\langle p_0, f(x_0) \rangle = \max_{x \in X} \langle p_0, f(x) \rangle$. By this, $x_0 \in WS(f, X)$, which is a contradiction with the inequality $d(x_0, WS(f, X)) \geq \epsilon_0$ by the continuity of the distance function. The last assertion is obtained from Theorem 5.5. $\square$

5.4. Advantages of the algorithm. The properties given in Corollary 5.6 show certain advantages of the algorithm that we described in the previous subsection over some existing ones. They mainly concern the obtainment of the exact solution set after a finite number of iterations for linear problems and the convergence to the solution set for convex problems.

In the known weighting method (see [21]) the stopping criterion depends uniquely on the number of weight vectors chosen from the weight set; thus the behavior of the data of the problem is not exploited, so that even when the problem is linear, it is not guaranteed that the solution set can be obtained after a finite number of iterations. To see this let us consider problem (VP) in which $X = [0, 1] \subset \mathbb{R}$ and $f(x) = (x, -\pi x) \in \mathbb{R}^2$ for $x \in X$. The weight set is often generated as follows. For $k \geq 1$, set $\Delta_k = \{(i/k, (k-i)/k) : i = 0, 1, \dots, k\}$. The solution set of scalarized problems using the weight set Δ_k is the set $S_k = \cup_{\lambda \in \Delta_k} S(\lambda \circ f, X)$. Direct calculation shows that for each k , S_k consists of two end points $x = 0$ and $x = 1$ only. Thus, it never approaches the solution set of (VP) which consists of the whole set X despite the fact that the Hausdorff distance from the weight sets Δ_k to the standard simplex $\Lambda = \{(x_1, x_2) \in \mathbb{R}^2 : x_1, x_2 \geq 0, x_1 + x_2 = 1\}$ goes to 0 as k tends to ∞ .

A similar situation can occur when applying the normal-boundary intersection method of [6] to the convex case (note that this method was originally developed to solve nonconvex problems). This can be seen by the following example. Consider problem (VP) with X being the ball touching the coordinate axes in $\mathbb{R}^3$, say $X = \{x = (x_1, x_2, x_3) \in \mathbb{R}^3 : (x_1 - 1)^2 + (x_2 - 1)^2 + (x_3 - 1)^2 \leq 2\}$, and $f(x) = -x$. This problem is equivalent to the problem of minimizing $F = -f$ on the set X , which was denoted by (MOP) in [6]. The so-called CHIM simplex (the convex hull of individual minima) is the convex hull of three points $(0, \sqrt{2}, \sqrt{2})$, $(\sqrt{2}, 0, \sqrt{2})$, and $(\sqrt{2}, \sqrt{2}, 0)$. The normal-boundary intersection method generates solutions of (MOP) by choosing a finite subset, say A , of the CHIM simplex and finding for each $a \in A$ those $x \in X$ such that $F(x) = a + t_0(1, 1, 1)$ with $t_0 = \min\{t : a + t(1, 1, 1) \in F(X)\}$. It is clear that all the solutions of (MOP) within a distance less than $(2 - \sqrt{2})/\sqrt{6}$ from the solutions $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$ are never obtained by this method regardless of what set A is chosen.

6. Numerical examples. In the implementation of the method we presented in the previous section, the following observations may be useful. Problem (P_v) , which is repeatedly used in Step 2, can be written as

$$\begin{aligned} \min \quad & \|v - y\|^2 \\ \text{s.t.} \quad & f(x) - y \geq 0, \\ & x \in X \text{ and } y \geq 0. \end{aligned}$$

When X is a convex set and f is a concave function, this problem is a convex mathematical programming problem with a quadratic objective function and can be solved by many existing methods (see also [8] for scalarization using a norm in a general setting). In Step 4, the information of the previous iterations can be used to determine the vertex set V_{k+1} of A_{k+1} . It consists of some elements of V_k^+ and some new vertices obtained by the cuts by the hyperplanes $\langle p_v, y \rangle = \rho_{p_v}, v \in V_k^+$.

Below, our solution method is illustrated by some small-scale examples. We first consider a linear multiobjective programming problem of [10]. For this problem with the zero tolerance the algorithm terminates after a finite number of steps and yields the same result as that of [10]. The second example is a convex biobjective programming

problem with two variables. The low dimensionality of the variable is chosen for the sake of easy visualization. With a positive tolerance the program terminates after a finite number of iterations and generates an approximation of the weakly efficient solutions. Preliminary account of computational experience is given in the last subsection after having tested several linear problems that have pseudorandomly generated data. Our program has been written in MATLAB Version 6.1. One of the most important procedures in coding the algorithm is the finding of the vertex set V_{k+1} in Step 4. We use the method of [5] for this purpose, which seems to be simple and quite efficient for enumerating vertices of polytopes (see also [20]). A more complete picture of computational aspects of the algorithm needs further investigations in order to handle problems of larger size.

6.1. Example 1. Consider the following linear programming problem (see [9]):

$$\begin{aligned} \text{Max} \quad & \begin{bmatrix} 1 & 3 & -2 & 0 & 1 \\ 3 & -1 & 0 & 3 & 1 \\ 1 & 0 & 2 & 0 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ . \\ . \\ . \\ x_5 \end{bmatrix} \\ \text{s.t.} \quad & \begin{bmatrix} -2 & -4 & 0 & 0 & -3 \\ 0 & 0 & -2 & -5 & -4 \\ -5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -2 & 0 \\ -5 & -5 & -2 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ . \\ . \\ . \\ x_5 \end{bmatrix} \geq \begin{bmatrix} -27 \\ -35 \\ -26 \\ -24 \\ -36 \end{bmatrix}, \\ & x_1, \dots, x_5 \geq 0. \end{aligned}$$

We set $\epsilon = 0$. After five iterations the algorithm terminates and the result contains 64 weakly efficient faces (this number corresponds to the power of N_5), which give exactly 4 maximal weakly efficient faces of dimension 3, 31 weakly efficient edges, and 17 weakly efficient vertices as obtained in [10]. The maximal number of vertices to be stored is 271.

6.2. Example 2. Consider the following biobjective programming problem:

$$\begin{aligned} \text{Max} \quad & -(-0.5(x_1 - 1)^2 - (x_2 - 5)^2, -(x_1 - x_2)^2) \\ \text{s.t.} \quad & x_1 \geq 0, x_2 \geq 0 \text{ and } x_1^2 + x_2^2 \leq 4. \end{aligned}$$

We choose $\epsilon = 0.003$ and find $a_1 = -9.3756$, $a_2 = 0$, $b_1 = -26.5$, and $b_2 = -4.5$. At the initialization the vertex set V_1 , except for the zero, consists of three elements,

$$v_1 = (0, 4.5), \quad v_2 = (17.1244, 4.5), \quad v_3 = (17.1244, 0).$$

By solving problems (P_{v_i}) , we obtain three vectors,

$$y_1 = (0, 4.5), \quad y_2 = (16.2785, 3.6132), \quad y_3 = (17.1214, 0).$$

Thus the set V_1^+ consists of vector y_2 only. The cut that uses this vector generates two new vertices,

$$v_5 = (17.1244, 2.8064), \quad v_6 = (15.3489, 4.5000).$$

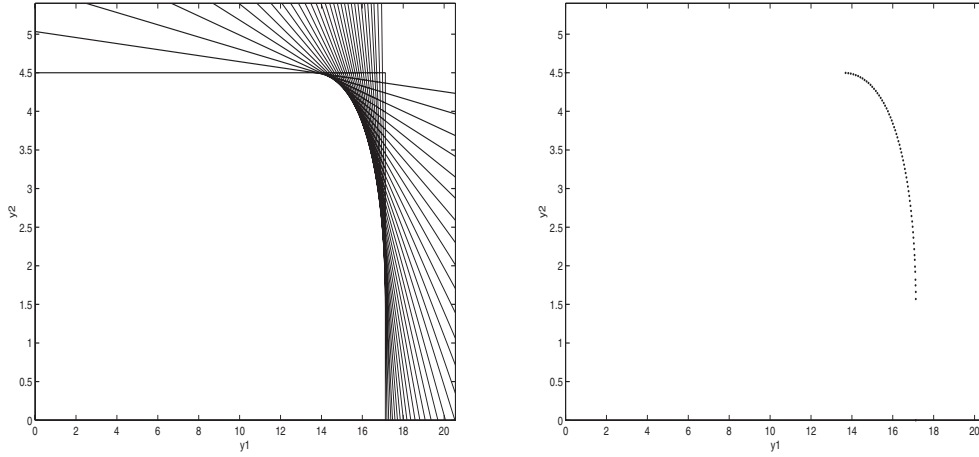


FIG. 6.1. On the left, the straight lines present the cuts described in Step 4, and on the right, the points obtained by the projection (Step 2) give an approximation of the weakly efficient values of the problem in the outcome space.

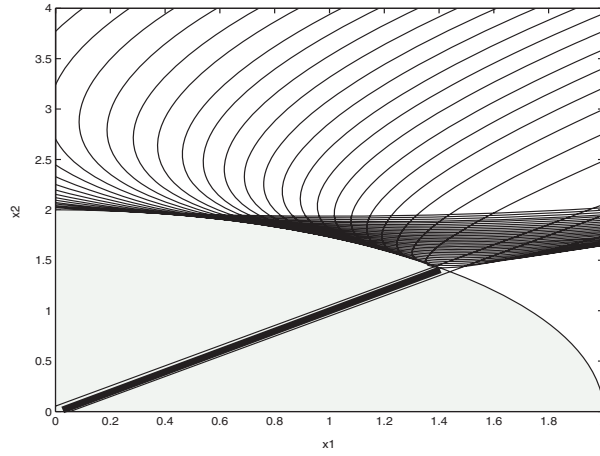


FIG. 6.2. The set S_ϵ obtained in Step 3 is an approximation of the weakly efficient solution set in the decision space.

In this way, the vertex set at the second iteration, except for the zero, is given by $V_2 = \{v_1, v_3, v_5, v_6\}$. The algorithm terminates after 5 iterations.

Figure 6.1 illustrates the cuts performed in the outcome space and the projections of the vertices of A_k on $f(X)$ which present an approximation of the weakly efficient values of the problem. Figure 6.2 shows the set S_ϵ (Step 3) in the decision space which is an approximation of the weakly efficient solutions of the problem. This set consists of the intersections of the regions bounded by elliptic upper level sets,

$$\langle p, f(x) \rangle \geq \rho_p - \epsilon,$$

and the feasible region (the first quarter disc),

$$\{(x_1, x_2) \in \mathbb{R}^2 : 0 \leq x_1 \leq x_2, x_1^2 + x_2^2 = 4\}.$$

TABLE 6.1

Problem size			Average results		
n	p	m	Number of weak. eff. faces	Max. number of vertices	CPU time
3	10	10	17.6	27	8.888
3	10	20	29	42	39.248
3	20	10	19.8	33.8	11.678
3	20	20	54.2	90.4	99.464
4	10	10	37.6	73.2	37.076
4	10	20	46.4	93.6	119.708
4	20	10	73.8	197	142.278
4	20	20	117.6	292	486.076
5	10	10	34.4	185.8	170.026
5	10	20	72	675	2071.456

Observe that S_ϵ includes also that portion of the small strip between two parallel lines in the first quarter disc, which is obtained from the elliptic upper level set with $p = (0, 1)$. When ϵ tends to 0, the set S_ϵ collapses to the weakly efficient solution set $\{(x_1, x_2) \in \mathbb{R}^2 : 0 \leq x_1 = x_2 \leq \sqrt{2}\} \cup \{(x_1, x_2) \in \mathbb{R}^2 : 0 \leq x_1 \leq x_2, x_1^2 + x_2^2 = 4\}$.

6.3. Randomly generated linear problems. We have applied our code to 50 linear problems in which the entries of the objective functions (an $(n \times m)$ -matrix) and the entries of the constraints (a $(p \times m)$ -matrix A and a p -vector b) are pseudorandom numbers from uniform distribution such that the polytope $\{x \in \mathbb{R}^n : Ax \leq b, x \geq 0\}$ is nonempty. Table 6.1 indicates some results of our experiments. The numbers in columns 4 to 6 are the average results of five problems of each size. For the problem instances, we choose $\epsilon = 0.01$. The experiments were carried out on a SPARC II (4×400 MHz, 1G RAM).

Table 6.1 shows that the number of vertices and computing time increases rapidly when the number of objectives grows. This is essentially due to the large number of quadratic programs to be solved in Step 2.

7. Final remark. We have seen that the scalarizing functions g_k constructed in section 5.2 are quite useful in developing the algorithm (section 5.3) for finding the weakly efficient solution set of (VP) and in establishing the convergence of the algorithm. The outer approximation method we have chosen to construct the scalarizing functions has the distinct feature of not only approximating a convex set by a sequence of polytopes, as many other methods offer, but also of approximating the weakly efficient set of the given set. For nonlinear problems when $\epsilon > 0$ is sufficiently small the set S_ϵ can be regarded as an approximation of the weakly efficient solution set. We note, however, that even in (scalar) mathematical programming, finding all optimal solutions of a problem is not easy, and most of the available algorithms allow one to obtain one or a few optimal solutions only. Therefore, the determination of the set S_ϵ is computationally expensive. A close inspection of the algorithm reveals that S_ϵ contains a subset of $x_v \in X$, solutions to problem (P_v) (Step 2), which are weakly efficient solutions of (VP). In terminology of multicriteria decision making [21], the solution y_v of Step 2 is the best compromise solution of the set $[f(X)]^\diamond$ from the target v . In this sense the set of those x_v with $v \in \cup_{i=1}^k V_i$ represents the best compromise portion of the solution set of (VP) with respect to the target set formed by positive vertices of approximating polytopes of $[f(X)]^\diamond$ and can be served as an approximation of the solution set of the problem. By this, when an interaction is necessary, the decision maker can take part by choosing the target among the elements of

the set V_i at Step 2 and stop the search once the compromise solution corresponding to his chosen target is reached.

To close up this remark we emphasize that our method is developed for convex problems, or more precisely for problems which satisfy hypothesis (H) of subsection 5.1. When this hypothesis is not fulfilled, only a part of the solution set, the image of which lies in the convex hull of $f(X)$, is obtained. In such a situation a new kind of approximation of the nonconvex set $f(X)$ is needed, instead of approximation by convex polytopes.

REFERENCES

- [1] H. P. BENSON, *Vector maximization with two objective functions*, J. Optim. Theory Appl., 28 (1979), pp. 253–257.
- [2] H. P. BENSON, *An outer approximation algorithm for generating all efficient extreme points in the outcome set of a multiple objective linear programming problem*, J. Global Optim., 13 (1998), pp. 1–24.
- [3] H. P. BENSON AND E. SUN, *Outcome space partition of the weight set in multiobjective linear programming*, J. Optim. Theory Appl., 105 (2000), pp. 17–36.
- [4] V. CHANKONG AND Y. Y. HAİMES, *Multiobjective Decision Making: Theory and Methodology*, North-Holland, New York, 1983.
- [5] P. C. CHEN, P. HANSEN, AND B. JAUMARD, *On-line and off-line vertex enumeration by adjacency lists*, Oper. Res. Lett., 10 (1991), pp. 403–409.
- [6] I. DAS AND J. E. DENNIS, *Normal-boundary intersection: A new method for generating the Pareto surface in nonlinear multicriteria optimization problems*, SIAM J. Optim., 8 (1998), pp. 631–657.
- [7] R. HORST AND H. TUY, *Global Optimization*, Springer-Verlag, Berlin, 1996.
- [8] J. JAHN, *Scalarization in vector optimization*, Math. Programming, 29 (1984), pp. 203–218.
- [9] J. JAHN, *Vector Optimization: Theory, Applications, and Extensions*, Springer-Verlag, Berlin, 2004.
- [10] N. T. B. KİM AND D. T. LUC, *Normal cones to a polyhedral convex set and generating efficient faces in linear multiobjective programming*, Acta Math. Vietnam., 25 (2000), pp. 101–124.
- [11] J. G. LIN, *Three methods for determining Pareto-optimal solutions of multiple-objective problems*, in Large-Scale Systems, Y. C. Ho and S. K. Mitter, eds., Plenum Press, New York, 1975, pp. 117–138.
- [12] D. T. LUC, *On scalarizing method in vector optimization*, in Large-Scale Modelling and Interactive Decision Analysis, G. Fandel, M. Grauer, A. Kurzhanski, and A. P. Wiersbicki, eds., Lecture Notes in Econom. and Math. Systems 273, Springer-Verlag, Berlin, 1986, pp. 46–51.
- [13] D. T. LUC, *Scalarization of vector optimization problems*, J. Optim. Theory Appl., 55 (1987), pp. 85–102.
- [14] D. T. LUC, *Theory of Vector Optimization*, Lecture Notes in Econom. and Math. Systems 319, Springer-Verlag, Berlin, 1989.
- [15] K. MIETTINEN, *Nonlinear Multiobjective Optimization*, Kluwer Academic, Boston, 1999.
- [16] J. RAKOWSKA, R. T. HAFKA, AND L. T. WATSON, *Tracing the efficient curve for multi-objective control-structure optimization*, Comput. Systems Engrg., 2 (1991), pp. 461–471.
- [17] Y. SAWARAGI, H. NAKAYAMA, AND T. TANINO, *Theory of Multiobjective Optimization*, Academic Press, New York, 1985.
- [18] R. E. STEUER, *Multiple-Criteria Optimization: Theory, Computation, and Application*, John Wiley and Sons, New York, 1986.
- [19] T. J. STEWART AND R. C. VAN DEN HONERT, EDS., *Trends in Multicriteria Decision Making*, Lecture Notes in Econom. and Math. Systems 465, Springer-Verlag, Berlin, 1997.
- [20] T. V. THİEU, B. T. TAM, AND V. T. BAN, *An outer approximation method for globally minimizing a concave function over a compact convex set*, Acta Math. Vietnam., 8 (1983), pp. 3–34.
- [21] P. L. YU, *Multiple-Criteria Decision Making: Concepts, Techniques and Extensions*, Plenum Press, New York, 1985.
- [22] M. ZELENY, *Linear Multiobjective Programming*, Springer-Verlag, New York, 1974.

SPACE MAPPING FOR OPTIMAL CONTROL OF PARTIAL DIFFERENTIAL EQUATIONS*

MICHAEL HINTERMÜLLER[†] AND LUÍS N. VICENTE[‡]

Abstract. Solving optimal control problems for nonlinear partial differential equations represents a significant numerical challenge due to the tremendous size and possible model difficulties (e.g., nonlinearities) of the discretized problems. In this paper, a novel space-mapping technique for solving the aforementioned problem class is introduced, analyzed, and tested. The advantage of the space-mapping approach compared to classical multigrid techniques lies in the flexibility of not only using grid coarsening as a model reduction but also employing (perhaps less nonlinear) surrogates. The space mapping is based on a regularization approach which, in contrast to other space-mapping techniques, results in a smooth mapping and, thus, avoids certain irregular situations at kinks. A new Broyden update formula for the sensitivities of the space map is also introduced. This quasi-Newton update is motivated by the usual secant condition combined with a secant condition resulting from differentiating the space-mapping surrogate. The overall algorithm employs a trust-region framework for global convergence. Issues involved in the computations are highlighted, and a report on a few illustrative numerical tests is given.

Key words. space mapping, optimal control of PDEs, simulation-based optimization, PDE-constrained optimization, trust regions, quasi-Newton methods

AMS subject classifications. 90C06, 90C26, 90C53

DOI. 10.1137/S105262340342907X

1. Introduction. Let us assume that we are interested in optimizing some objective related to physical phenomena simulated by a system of differential equations. We might be trying to determine unknown system parameters by matching observable data, or we might want to control properties of the system so that its state matches a given desired profile. Let us further assume that our goal is to minimize a smooth function $g : X \subset \mathbb{R}^n \rightarrow \mathbb{R}$, evaluated by accurately solving the discretized system of differential equations that models the underlying physical phenomena. Given the computational complexity involved in simulating the system, the model g might be expensive to evaluate, and some alternative smooth function $\hat{g} : \hat{X} \subset \mathbb{R}^{\hat{n}} \rightarrow \mathbb{R}$ is assumed available at a cheaper cost, by solving the system less accurately or by using some form of surrogate. We will call g the fine model and $\hat{g}$ the coarse model. Similarly, X and $\mathbb{R}^n$ will be called the fine domain and fine space, respectively, and $\hat{X}$ and $\mathbb{R}^{\hat{n}}$ the coarse domain and coarse space, respectively. We will assume that X and $\hat{X}$ are open domains.

The space-mapping technique provides an attractive framework to improve the use of the coarse model $\hat{g}$ as a surrogate for the optimization of the fine model g . The space-mapping surrogate is of the form $\hat{g} \circ P$, where P , the so-called space mapping,

*Received by the editors June 3, 2003; accepted for publication (in revised form) September 21, 2004; published electronically June 30, 2005.

<http://www.siam.org/journals/siopt/15-4/42907.html>

[†]Rice University, Department of Computational and Applied Mathematics, Houston, TX 77005 and Department of Mathematics, University of Graz, A-8010 Graz, Austria (michael.hintermueller@uni-graz.at). This author was supported by the Austrian science fund FWF under grant SFB “Optimierung und Kontrolle.”

[‡]Departamento de Matemática, Universidade de Coimbra, 3001-454 Coimbra, Portugal (lnv@matuc.pt). Support for this author was provided by Centro de Matemática da Universidade de Coimbra, by FCT under grant POCTI/35059/MAT/2000, by the European Union under grant IST-2000-26063, and by Fundação Calouste Gulbenkian.

attempts to match, in the coarse space, the fine model values and/or their responses.

The space-mapping technique was introduced first by Bandler et al. [5] in 1994. The idea of space mapping has been developed along different directions and generalized to a number of contexts. To overcome some of its inherent difficulties, techniques from nonlinear optimization have been incorporated. One of the problems lies in the information necessary to compute the sensitivities (or the Jacobian) of the space mapping which involves, among other things, (possibly expensive) gradient information of the fine model. Bandler et al. [6] suggested the use of the Broyden method to construct linear approximations for the space mapping. This space-mapping Broyden method, also named the aggressive space-mapping method, has been then enhanced by Bakr et al. [2] with the application of trust regions for globalization. These and other approaches are reviewed by Bakr et al. in the papers [3, 4]. See also [13, 26]. The reader is further referred to the special issue on surrogate modeling and space mapping that has been recently edited by Bandler and Madsen [7].

In this paper we are concerned with the application of the space-mapping technique to control problems for partial differential equations (PDEs). The problem under consideration is the following:

$$\begin{aligned} (1.1a) \quad & \text{minimize } \mathcal{J}(y, u) \quad \text{over } (y, u) \in W \times U \\ (1.1b) \quad & \text{subject to } A(y, u)y + C(y, u) = 0 \quad \text{in } \Omega \quad + \text{ boundary conditions,} \end{aligned}$$

where $\mathcal{J} : W \times U \rightarrow \mathbb{R}$ is a sufficiently smooth objective functional, $A(y, u)$ denotes a second order partial differential operator, $C(y, u)$ is a possibly nonlinear mapping, and Ω is a bounded domain in $\mathbb{R}^n$. Above, W , U are appropriate Hilbert spaces and y is referred to as the state variable. The variable u is the control variable. We assume that for every $u \in U$ the state equation in (1.1b) admits a (unique) solution $y = y(u) \in W$. Using $y(u)$ we can consider the reduced problem

$$(1.2) \quad \text{minimize } \mathcal{J}_{\text{red}}(u) = \mathcal{J}(y(u), u) \quad \text{over } u \in U$$

instead of (1.1).

One instance of the model problem (1.1) is given by

$$\begin{aligned} & \text{minimize } \frac{1}{2} \|y - y_d\|_{L^2(\Omega)}^2 + \frac{\delta}{2} \|u\|_{L^2(\Omega)}^2 \quad \text{over } (y, u) \in H_0^1(\Omega) \times L^2(\Omega) \\ & \text{subject to } -\Delta y + f(y) = u \quad \text{in } \Omega = (0, 1)^2, \end{aligned}$$

where $y_d \in L^2(\Omega)$, $\delta > 0$ is fixed, and f denotes a sufficiently smooth function; see, e.g., [18, 19]. Obviously, we have $A(y, u) = -\Delta$ and $C(y, u) = f(y) - u$. Here, $H_0^1(\Omega)$ and $L^2(\Omega)$ denote Sobolev and Lebesgue spaces; see, e.g., [12].

Output least squares formulations of parameter identification problems are further instances of the general model problem (1.1). In this case, one aims at determining a quantity u , which is not directly accessible to measurements, by fitting the measured data y_d . Frequently, the relation between u and y can be described by a semilinear elliptic PDE:

$$-\operatorname{div}(e(u)\nabla y) + C(y, u) = 0 \quad \text{in } \Omega, \quad y \in H_0^1(\Omega).$$

Hence, $A(y, u)y = -\operatorname{div}(e(u)\nabla y)$, where e is a possibly nonlinear mapping.

In [21, 22, 24] a multigrid approach (algorithm MG/OPT) to discretizations of minimization problems of type (1.1) is considered. In a two-grid approach, the algorithm combines a prescribed number of iterations of a minimization algorithm on the

fine grid with high accuracy solves of a slightly modified problem on the coarse grid. Prolongation and restriction operators achieve the transport of coarse grid solutions to the fine grid, and vice versa. Our space-mapping approach, however, allows more flexibility in the sense that not only may we consider a coarse grid discretization of the underlying optimization problem, but we can also use a surrogate model which is even simpler to solve than the discretized problem on the coarse grid. This is of particular importance in cases where the coarse grid approximation is still difficult to handle due to, e.g., problematic nonlinearities and/or model complexities. Also, the surrogate idea applies without grid coarsening. In fact, we may want to replace a difficult problem by an approximating simpler one on the same (fine) grid. In addition, the trust-region based approach that we use is globally convergent in the general nonconvex case, while MG/OPT, as outlined in [21, 22, 24], requires convexification to enforce descent. However, let us point out that in this paper we are primarily interested in the above-mentioned flexibility of space mapping of combining different models on different levels to obtain *good* approximations of fine model solutions quickly (when compared to, e.g., multigrid techniques), rather than in designing a new fast fine model solver. As it appears (see the discussion at the end of Example 2 in section 8), for the problem class considered here, the space-mapping technique exhibits the potential to be at the core of a fast and flexible fine model solver. This, however, requires a different algorithmic framework than the one introduced in the present paper and will be reported elsewhere.

The outline of the paper is as follows. In section 2 we introduce a new smooth space-mapping technique. The key idea is to utilize regularization techniques of Tikhonov type. This approach is a remedy to problematic nondifferentiability and nonuniqueness in space mapping. The motivation of our approach and its practical appearance differ significantly from the norm-penalization technique in [6]. Section 3 gives a comparison between a nonsmooth space mapping previously suggested by Vicente [27] and the smoothed one introduced in this paper. In section 4 the space-mapping algorithm of [2, 6] is outlined in our space-mapping context. It is based on a Broyden-type approximation of the sensitivities of the space mapping and on a trust-region-type globalization. The focus of section 5 is on a new Broyden update reflecting the approximation requirements induced by the sensitivities of the space mapping and by its use in the gradient of the space-mapping surrogate. The application of our new space-mapping approach to optimal control problems governed by PDEs is the core of section 6. In section 7 we consider computational aspects with respect to coarse and fine model derivatives. A report on numerical test runs including a comparison with nonlinear multigrid methods is given in section 8. We end this paper with some conclusions and prospects of future work.

Notation. Throughout we use D_f for the Jacobian matrix of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$. As usual, ∇ denotes the gradient operator. In the case where $m = 1$, we have $D_f(x)^\top = \nabla f(x)$. The partial derivatives of f with respect to x_i are denoted by $f_{x_i}(x)$ or $\nabla_{x_i} f(x)$, where the latter notation is typically used if $f : \mathbb{R}^n \rightarrow \mathbb{R}$.

2. A new smooth space mapping. We introduce in this paper a new definition of space mapping $P : X \rightarrow \hat{X}$ as follows:

$$(2.1) \quad P(x) = \operatorname{argmin} \left\{ \frac{\alpha}{2} \|\hat{r}(\hat{x}) - r(x)\|_M^2 + \frac{1}{2} [\hat{g}(\hat{x}) - g(x)]^2 \mid \hat{x} \in \hat{X} \right\},$$

where $\alpha > 0$ is a smoothing parameter whose role will become clear later. We assume that the argmin operator returns a single minimizer, in other words, that P is a

point-to-point mapping. Here r and $\hat{r}$ are some operators that map X and $\hat{X}$ into some common space $\mathbb{R}^p$ where the values of fine and coarse variables can be compared against each other. An illustration of r and $\hat{r}$ is given in section 6. In the definition of the space mapping, M is a $p \times p$ symmetric positive definite matrix and $\|\cdot\|_M$ is the ellipsoidal norm defined by $\|z\|_M = \|M^{\frac{1}{2}}z\|_2$.

The space mapping P defines a surrogate $g_P = \hat{g} \circ P$ for the fine model g . One of the aims of space mapping is to minimize the surrogate g_P instead of minimizing g ; i.e., an approximate solution to

$$\text{minimize } g(x) \quad \text{over } x \in X$$

is obtained by solving

$$\text{minimize } g_P(x) = (\hat{g} \circ P)(x) = \hat{g}(P(x)) \quad \text{over } x \in X.$$

The effect of the smoothing parameter α becomes clearer by taking a close look at the first order necessary conditions of problem (2.1):

$$(2.2) \quad \alpha D_{\hat{r}}^{\top}(\hat{x})M(\hat{r}(\hat{x}) - r(x)) + [\hat{g}(\hat{x}) - g(x)]\nabla\hat{g}(\hat{x}) = 0,$$

with $\hat{x} = P(x)$. Now, we differentiate (2.2) with respect to x :

$$\begin{aligned} & \alpha D_{\hat{r}}^{\top}(\hat{x})^{\top}M(D_{\hat{r}}(\hat{x})D_P(x) - D_r(x)) + \alpha D_{\hat{r}}^2(\hat{x}; M(\hat{r}(\hat{x}) - r(x)))D_P(x) \\ & + \nabla\hat{g}(\hat{x})(\nabla\hat{g}(\hat{x})^{\top}D_P(x) - \nabla g(x)^{\top}) + [\hat{g}(\hat{x}) - g(x)]\nabla^2\hat{g}(\hat{x})D_P(x) = 0, \end{aligned}$$

where $D_{\hat{r}}^2(\hat{x}; z)$ is the derivative of $D_{\hat{r}}(\hat{x})z$ with respect to $\hat{x}$. So, one obtains

$$G(x)D_P(x) = \alpha D_{\hat{r}}^{\top}(\hat{x})^{\top}MD_r(x) + \nabla\hat{g}(\hat{x})\nabla g(x)^{\top},$$

where

$$\begin{aligned} G(x) = & \alpha \left(D_{\hat{r}}^{\top}(\hat{x})^{\top}MD_{\hat{r}}(\hat{x}) + D_{\hat{r}}^2(\hat{x}; M(\hat{r}(\hat{x}) - r(x))) \right) \\ & + [\hat{g}(\hat{x}) - g(x)]\nabla^2\hat{g}(\hat{x}) + \nabla\hat{g}(\hat{x})\nabla\hat{g}(\hat{x})^{\top}, \end{aligned}$$

with $\hat{x} = P(x)$. The following theorem summarizes the basic smoothness properties of the surrogate $g_P = \hat{g} \circ P$.

THEOREM 2.1. *Let g , $\hat{g}$, r , and $\hat{r}$ be continuously differentiable functions in their domains. Assume that P is a well-defined point-to-point mapping from X to $\hat{X}$.*

- (1) *Then g_P is regular in X (i.e., g_P has one-sided directional derivatives in X).*
- (2) *In addition, let $\hat{g}$ and $\hat{r}$ be twice continuously differentiable in X . If α is such that $G(x)$ is uniformly nonsingular in X , then g_P is continuously differentiable in X .*

Proof. The fact stated in (1) comes directly from the properties of marginal or value functions (see, e.g., [25]). The proof of (2) lies in the informal derivation given before the theorem. $\square$

We point out that in practical applications, due to the existence of several local minima, (2.1) may not yield a point-to-point mapping P . In this case, one can employ additional selection criteria such as picking the local minimizer of (2.1) with least norm or closest to some reference value to get a single valued P .

3. Comparing smooth and nonsmooth approaches. Let us further study the smoothing effect of $\alpha > 0$ by comparing our new approach to the approach introduced by Vicente [27], where the space mapping $\bar{P}$ is defined as

$$(3.1) \quad \bar{P}(x) = \operatorname{argmin} \left\{ \frac{1}{2} [\hat{g}(\hat{x}) - g(x)]^2 \mid \hat{x} \in \hat{X} \right\}$$

if $S(x) = \{\hat{x} \in \hat{X} \mid \hat{g}(\hat{x}) = g(x)\}$ is empty, and as

$$(3.2) \quad \bar{P}(x) = \operatorname{argmin} \left\{ \frac{1}{2} \|\hat{x} - x\|_2^2 \mid \hat{g}(\hat{x}) = g(x) \mid \hat{x} \in \hat{X} \right\}$$

if $S(x)$ is nonempty. In this setting it is considered that $n = \hat{n} = p$ and that r and $\hat{r}$ are the identity operators. It is proved in [27] that if g and $\hat{g}$ are continuously differentiable functions and if $\bar{P}$ is point-to-point, then $g_{\bar{P}} = \hat{g} \circ \bar{P}$ is a regular function, i.e., a function that has one-sided directional derivatives. The first order necessary conditions for (3.2) imply, under the constraint qualification $\nabla \hat{g}(\bar{P}(x)) \neq 0$, the existence of a Lagrange multiplier $\lambda(x)$ such that

$$(3.3) \quad \hat{x} - x + \lambda(x) \nabla \hat{g}(\hat{x}) = 0,$$

with $\hat{x} = \bar{P}(x)$. Nondifferentiability can only occur on the boundary of the set $\{x \mid S(x) \neq \emptyset\}$. When approaching the boundary of $\{x \mid S(x) \neq \emptyset\}$ from its interior, a kink occurs when $\nabla \hat{g}(\bar{P}(x))$ is approaching zero and $\bar{P}(x)$ is not becoming close to x ; in these situations $|\lambda(x)|$ tends to $+\infty$.

The analogue of (2.1) in the setting considered in this section would be

$$(3.4) \quad P(x) = \operatorname{argmin} \left\{ \frac{\alpha}{2} \|\hat{x} - x\|_2^2 + \frac{1}{2} [\hat{g}(\hat{x}) - g(x)]^2 \mid \hat{x} \in \hat{X} \right\}.$$

A variation of this definition has been independently analyzed in [26]. In this case $G(x)$ would reduce to

$$G(x) = \alpha I + [\hat{g}(\hat{x}) - g(x)] \nabla^2 \hat{g}(\hat{x}) + \nabla \hat{g}(\hat{x}) \nabla \hat{g}(\hat{x})^\top,$$

with $\hat{x} = P(x)$. The smoothing role of α becomes more evident in this context. Moreover, condition (2.2) in the simpler case (3.4) reduces to

$$(3.5) \quad \alpha(\hat{x} - x) + [\hat{g}(\hat{x}) - g(x)] \nabla \hat{g}(\hat{x}) = 0,$$

with $\hat{x} = P(x)$.

By comparing (3.3) and (3.5) and assuming that $\hat{x} = P(x)$ and $\nabla \hat{g}(P(x))$ are relatively close to $\hat{x} = \bar{P}(x)$ and $\nabla \hat{g}(\bar{P}(x))$, respectively, we can gain some insight into the appropriate size for α :

$$\alpha \lambda(x) \approx \hat{g}(\hat{x}) - g(x).$$

In Figure 3.1 a simple model example is displayed. The fine and coarse models are $g(x) = x^2$ and $\hat{g}(\hat{x}) = (\hat{x} - 1)^2 + 1$, respectively. The upper left plot shows the fine and coarse models together with the surrogate $\hat{g} \circ \bar{P}$, where $\bar{P}$ is given by (3.1)–(3.2). In the upper right plot the new smooth surrogate is displayed in dashed lines. The third plot focuses on the behavior near the critical kink at $x = 1$. From the above relation between $\lambda(x)$ and $\hat{g}(\hat{x}) - g(x)$ and the structure of (3.4) we infer that the smaller α becomes, the closer the smooth and nonsmooth surrogates are.

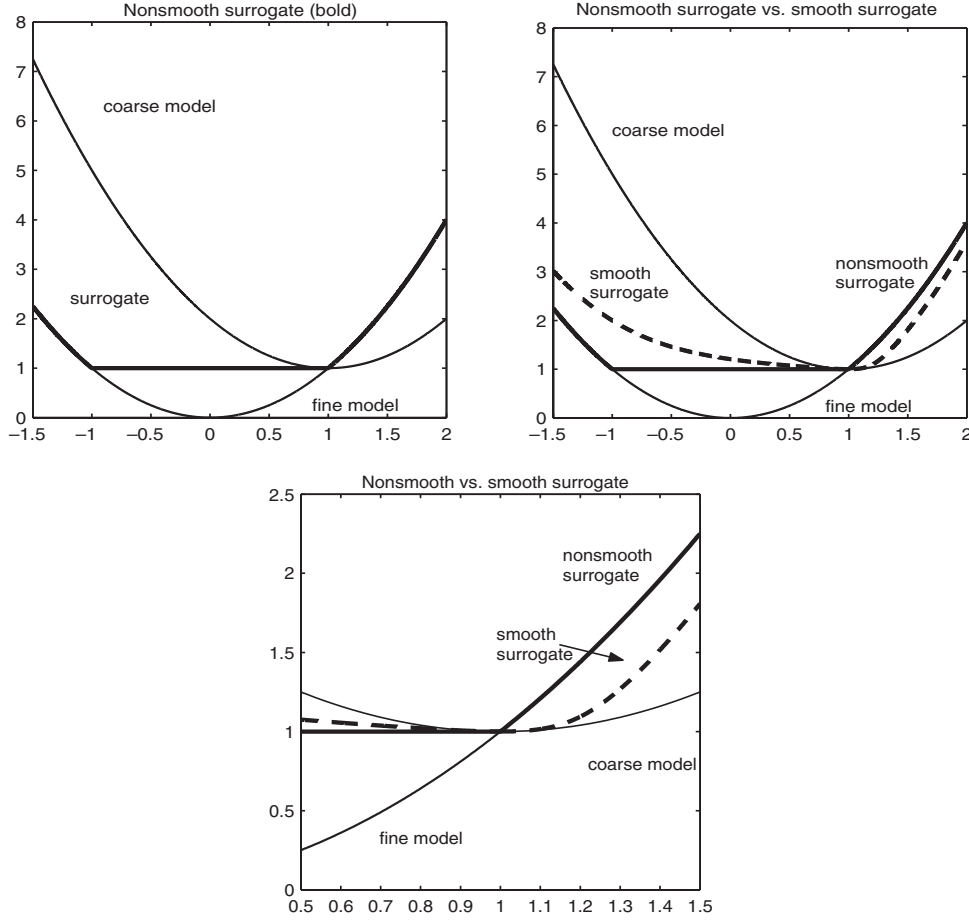


FIG. 3.1. Comparison between the nonsmooth surrogate [27] and the new smooth surrogate.

4. Aggressive space-mapping method. As we have shown in section 2, the computation of the sensitivities $D_P(x)$ requires first and second order derivative information of the coarse model and, more importantly, first order derivatives of the fine model. Requiring the gradient of the fine model can pose problems in many practical situations where the evaluation of the fine model is itself very expensive. To overcome this difficulty Bandler et al. [6] introduced a Broyden approach to space mapping, later globalized by Bakr et al. [2] with the help of the trust-region technique. This aggressive space-mapping method using trust regions is described next for the space-mapping definition (2.1).

The derivative $D_P(x)$ appears both in the formula for the gradient of g_P given by

$$(4.1) \quad \nabla g_P(x) = D_P(x)^\top \nabla \hat{g}(\hat{x}) \quad \text{with } \hat{x} = P(x),$$

and in the local linearization of P at x , along the increment Δx , of the form

$$(4.2) \quad P(x + \Delta x) \approx P(x) + D_P(x)\Delta x.$$

The Broyden updating formula provides a matrix B which can be used to replace $D_P(x)$ in both (4.1) and (4.2).

ALGORITHM 4.1. *Aggressive space-mapping method.*

Choose $x_0 \in \mathbb{R}^n$, $\Delta_0 > 0$, $B_0 \in \mathbb{R}^{\hat{n} \times n}$, and $\gamma_1, \eta_1 \in (0, 1)$.

0. Compute $P(x_0)$ by solving (2.1) with $x = x_0$.

For $k = 0, 1, 2, \dots$

1. Compute an approximated solution Δx_k for the trust-region subproblem

$$\text{minimize } \hat{g}(P(x_k) + B_k \Delta x) \quad \text{subject to } \|\Delta x\| \leq \Delta_k,$$

over $\Delta x \in \mathbb{R}^n$.

2. Compute $P(x_k + \Delta x_k)$ by solving (2.1) with $x = x_k + \Delta x_k$.

3. Compute the ratio between actual and predicted reductions:

$$\rho_k = \frac{\text{ared}(x_k, \Delta x_k)}{\text{pred}(x_k, \Delta x_k)} = \frac{\hat{g}(P(x_k)) - \hat{g}(P(x_k + \Delta x_k))}{\hat{g}(P(x_k)) - \hat{g}(P(x_k) + B_k \Delta x_k)}.$$

4. If $\rho_k \geq \eta_1$, then $x_{k+1} = x_k + \Delta x_k$ and Δ_{k+1} is chosen so that $\Delta_{k+1} \geq \Delta_k$. In this case, update B_{k+1} using Broyden's formula:

$$(4.3) \quad B_{k+1} = B_k + \frac{\Delta P_k - B_k \Delta x_k}{\|\Delta x_k\|_2^2} \Delta x_k^\top,$$

where $\Delta P_k = P(x_k + \Delta x_k) - P(x_k)$.

5. If $\rho_k < \eta_1$, then $x_{k+1} = x_k$ and $\Delta_{k+1} = \gamma_1 \Delta_k$. Keep $B_{k+1} = B_k$.

end

The initial value for B can be given by the classical choice $B_0 = I$ if $\hat{n} = n$. In section 6 we will introduce an appropriate choice for B_0 in a problem context where $\hat{n} \neq n$. The choice of Δ_0 is discussed in [11].

The norm used to define the trust region can be chosen according to practical considerations, but it is typically either the ℓ_2 or the ℓ_∞ norm. The mechanism given in steps 4–5 to update the trust radius is quite elementary, but it suffices to prove global convergence of trust-region algorithms. More sophisticated strategies can be found in [11].

The global convergence analysis is described in the next theorem, for which the classical theory of trust regions provides a proof (see [11, section 8.4] and the references therein). It is not our goal to investigate this subject further but only to list the ingredients necessary for global convergence.

THEOREM 4.1. *Let g_P be a continuously differentiable function with uniformly continuous gradient in X and bounded below on $L(x_0) = \{x \in X \mid g_P(x) \leq g_P(x_0)\}$. Consider a sequence $\{x_k\}$ generated by a trust-region method of the form of Algorithm 4.1, where the step Δx_k provides a fraction of the Cauchy decrease [11, section 6.3] and the Hessian used in the trust-region model $\hat{g}(P(x_k) + B_k \Delta x)$ is uniformly bounded. Finally, let B_k satisfy Carter's condition (see [10] and [11, section 8.4.1]) for all k :*

$$(4.4) \quad \frac{\|\nabla g_P(x_k) - B_k^\top \nabla \hat{g}(P(x_k))\|}{\|B_k^\top \nabla \hat{g}(P(x_k))\|} \leq \frac{\kappa_{mdc}(1 - \eta_1)}{2},$$

where $\kappa_{mdc} \in (0, 1)$ is the fraction of the Cauchy decrease achieved by step Δx_k . Then

$$\lim_{k \rightarrow +\infty} \|\nabla g_P(x_k)\| = 0.$$

We remark that the use of the exact sensitivities of P —in other words, the use of $B_k = D_P(x_k)$ —trivially satisfies (4.4).

Finally, for a practical implementation of Algorithm 4.1, a stopping rule has to be implemented. A possible choice is to terminate the method as soon as

$$\|\nabla g_P(x_k)\| \leq \epsilon_{\text{rel}} \|\nabla g_P(x_0)\| + \epsilon_{\text{abs}},$$

where $\epsilon_{\text{abs}} \leq \epsilon_{\text{rel}} \ll 1$ with ϵ_{abs} a suitable small positive constant. For further safeguards and additional numerical considerations in the context of trust-region methods, we refer to [11].

5. A new Broyden update for the aggressive space-mapping method.

The Broyden update (4.3) is the good Broyden update for solving systems of nonlinear equations [14]. However, the goal in space mapping is not to solve the system $P(x) = 0$ but rather to exploit the minimization of the surrogate $g_P = \hat{g} \circ P$. Note that the derivative $D_P(x)$ appears in the formula for the gradient of g_P given in (4.1). Our goal is to modify Broyden's formula to better reflect the use of $D_P(x)$ in the formula for $\nabla g_P(x)$.

The good Broyden formula is a rank one update B of B_k that satisfies the secant equation

$$(5.1) \quad B\Delta x_k = \Delta P_k.$$

The matrix B replaces the role of $D_P(x_k)$ in

$$P(x_k) + D_P(x_k)\Delta x_k \approx P(x_k + \Delta x_k).$$

Now, we also want to use B to approximate the role of $D_P(x_k)$ in (4.1), from iteration k to $k+1$:

$$\hat{g}(P(x_k)) + (D_P(x_k)^\top \nabla \hat{g}(P(x_k)))^\top \Delta x_k \approx \hat{g}(P(x_k + \Delta x_k)).$$

This motivation leads to the new secant condition

$$(5.2) \quad \nabla \hat{g}_k^\top B \Delta x_k = \Delta \hat{g}_k,$$

where $\nabla \hat{g}_k = \nabla \hat{g}(P(x_k))$ and $\Delta \hat{g}_k = \hat{g}(P(x_k + \Delta x_k)) - \hat{g}(P(x_k))$. The simultaneous satisfaction of (5.1) and (5.2) is possible only if

$$\nabla \hat{g}_k^\top \Delta P_k = \Delta \hat{g}_k,$$

a condition that would in turn reflect

$$(5.3) \quad \nabla \hat{g}(P(x_k))^\top (P(x_k + \Delta x_k) - P(x_k)) = \hat{g}(P(x_k + \Delta x_k)) - \hat{g}(P(x_k)).$$

It is unlikely that (5.3) is strictly satisfied and therefore unreasonable to compute B based on the simultaneous satisfaction of (5.1) and (5.2).

A way to circumvent this problem is to relax (5.1) by determining B as the optimal solution of

$$(5.4) \quad \text{minimize } \frac{1}{2} \|B\Delta x_k - \Delta P_k\|_2^2 \quad \text{subject to } \nabla \hat{g}_k^\top B \Delta x_k = \Delta \hat{g}_k$$

over $B \in \mathbb{R}^{\hat{n} \times n}$. The following proposition gives a characterization of the optimal solution of problem (5.4).

PROPOSITION 5.1. *Let Δx_k and $\nabla \hat{g}_k$ be nonzero vectors. The optimal solution B^* of (5.4) satisfies*

$$B^* \Delta x_k - \Delta P_k = \frac{\Delta \hat{g}_k - \nabla \hat{g}_k^\top \Delta P_k}{\|\nabla \hat{g}_k\|_2^2} \nabla \hat{g}_k.$$

Proof. Let us rewrite problem (5.4) as

$$(5.5) \quad \text{minimize } \frac{1}{2} \|V_k v - \Delta P_k\|_2^2 \quad \text{subject to } \nabla \hat{g}_k^\top V_k v = \Delta \hat{g}_k$$

over $v \in \mathbb{R}^{\hat{n}n}$, with the help of the change of variables $v_{(i-1)n+j} = B_{ij}$, $i = 1, \dots, \hat{n}$, $j = 1, \dots, n$. The rows of the $\hat{n} \times \hat{n}n$ matrix V_k are composed by the elements of Δx_k and by $(\hat{n} - 1)n$ zeros. The matrix V_k has full row rank because $\Delta x_k \neq 0$.

From the assumptions on Δx_k and $\nabla \hat{g}_k$ we know that $V_k^\top \nabla \hat{g}_k \neq 0$. The first order necessary conditions for (5.5) can then be stated by assuming the existence of a Lagrange multiplier λ_k such that

$$V_k^\top V_k v - V_k^\top \Delta P_k + \lambda_k V_k^\top \nabla \hat{g}_k = 0.$$

Since $V_k^\top$ has full column rank, we obtain

$$(5.6) \quad V_k v - \Delta P_k + \lambda_k \nabla \hat{g}_k = 0.$$

By multiplying this equation on the left by $\nabla \hat{g}_k^\top$ and using the problem's constraint in (5.5), we get

$$(5.7) \quad \Delta \hat{g}_k - \nabla \hat{g}_k^\top \Delta P_k + \lambda_k \|\nabla \hat{g}_k\|_2^2 = 0.$$

Thus, (5.6) and (5.7) together imply

$$V_k v - \Delta P_k = \frac{\Delta \hat{g}_k - \nabla \hat{g}_k^\top \Delta P_k}{\|\nabla \hat{g}_k\|_2^2} \nabla \hat{g}_k.$$

The proof is completed by returning to the formulation (5.4). $\square$

Proposition 5.1 suggests a perturbation for the right-hand side of the secant equation (5.1):

$$B \Delta x_k = \Delta P_k + \frac{\Delta \hat{g}_k - \nabla \hat{g}_k^\top \Delta P_k}{\|\nabla \hat{g}_k\|_2^2} \nabla \hat{g}_k.$$

For numerical purposes it might be advantageous to reduce the size of the new term that is added to ΔP_k :

$$(5.8) \quad \widetilde{\Delta P}_k = \Delta P_k + \sigma_k \frac{\Delta \hat{g}_k - \nabla \hat{g}_k^\top \Delta P_k}{\|\nabla \hat{g}_k\|_2^2} \nabla \hat{g}_k,$$

with $\sigma_k \in (0, 1]$, depending on the impact that $\hat{g}$ has in the definition of the space mapping P . The new Broyden update is therefore given by

$$B_{k+1} = B_k + \frac{\widetilde{\Delta P}_k - B_k \Delta x_k}{\|\Delta x_k\|_2^2} \Delta x_k^\top.$$

Notice that if we allow $\sigma_k = 0$ in (5.8), then the new Broyden update becomes the classical Broyden update as discussed, e.g., in [14]. In section 8 we will see that, for appropriate choices of $\sigma_k \in (0, 1)$, the new Broyden update leads to better numerical results than the classical one for an instance problem of optimal control of PDEs.

6. Application of the space-mapping method for optimal control of PDEs. In this section, we apply the space-mapping approach introduced in section 2 to the reduced problem (1.2). Let h and H with $H \geq h$ denote mesh sizes of discretizations of (1.2) yielding the fine model space $U_h = \mathbb{R}^{n_h}$ and the coarse model space $U_H = \mathbb{R}^{n_H}$. We have $n = n_h$, $X = U_h$, $\hat{n} = n_H$, and $\hat{X} = U_H$. For the ease of exposition we argue only for an L^2 -setting with standard inner product. Thus, by rescaling on the discrete level we essentially have to deal with ℓ_2 inner products only.

We introduce now discretized versions of the reduced problem (1.2). Let $y^h(u^h)$ denote the solution of the discretized PDE in (1.1b) with mesh size h . Moreover, let J^h be an appropriate discretization of the cost functional $\mathcal{J}$. Then

$$J_{\text{red}}^h(u^h) = J^h(y^h(u^h), u^h).$$

In an analogous way one obtains the coarse model J_{red}^H :

$$J_{\text{red}}^H(u^H) = J^H(y^H(u^H), u^H).$$

In order to simplify the notation and to make it similar to the one used in section 2, we will use J , J_{red} , y , and u for fine model quantities, and $\hat{J}$, $\hat{J}_{\text{red}}$, $\hat{y}$, and $\hat{u}$ for coarse model quantities.

Since $\dim U_h$ and $\dim U_H$ may differ, we define the linear restriction operator

$$I_H^h : U_h \longrightarrow U_H,$$

which maps a fine model quantity to a coarse model quantity. Typically, the definition of I_H^h depends on (infinite-dimensional) regularity properties of the control variable. Here we adopt restriction operators coming from multigrid methods; see [16, 23, 28].

The introduction of I_H^h enables us to define the space mapping $P : U_h \rightarrow U_H$ by

$$(6.1) \quad P(u) = \arg\min \left\{ \frac{\alpha_1}{2} \|\hat{y}(\hat{u}) - K_H^h y(u)\|_{\hat{M}_{\hat{y}}}^2 + \frac{\alpha_2}{2} \|\hat{u} - I_H^h u\|_{\hat{M}_{\hat{u}}}^2 + \frac{\alpha_3}{2} |\hat{J}_{\text{red}}(\hat{u}) - J_{\text{red}}(u)|^2 \mid \hat{u} \in U_H \right\},$$

with fixed $\alpha_1, \alpha_2, \alpha_3 \geq 0$, and $\alpha_1 + \alpha_2 + \alpha_3 > 0$. Above, $\hat{M}_{\hat{y}}$ represent a symmetric positive definite matrix resulting from discretizing a function space norm yielding $\|\hat{y}\|_{\hat{M}_{\hat{y}}}^2 = \hat{y}^T \hat{M}_{\hat{y}} \hat{y}$; this holds analogously for $\|\cdot\|_{\hat{M}_{\hat{u}}}^2$. Moreover, K_H^h denotes a restriction operator, possibly different from I_H^h . Throughout the rest of this paper we assume that $P(u)$ is single valued for every $u \in U_h$. Instead of $\hat{y}(\hat{u}) - K_H^h y(u)$ we could have used $\hat{R}\hat{y}(\hat{u}) - K_H^h R y(u)$, restricting the matching of the coarse and fine state variables to parts of its discretized domains.

The parallel to what has been introduced in section 2 is made by setting

$$x = u, \quad \hat{x} = \hat{u}, \quad p = n_H^{\hat{y}} + n_H, \quad \alpha = \alpha_1 = \alpha_2,$$

$$r(u) = \begin{pmatrix} K_H^h y(u) \\ I_H^h u \end{pmatrix}, \quad \hat{r}(\hat{u}) = \begin{pmatrix} \hat{y}(\hat{u}) \\ \hat{u} \end{pmatrix}, \quad \text{and}$$

$$M = \begin{pmatrix} M_{\hat{y}} & 0 \\ 0 & M_{\hat{u}} \end{pmatrix},$$

where $n_H^{\hat{y}}$ is the dimension of $\hat{y}(\hat{u})$.

Following the space-mapping philosophy presented in the previous sections, we now replace the problem of finding a solution to the fine model

$$(6.2) \quad \text{minimize } J_{\text{red}}(u) \quad \text{over } u \in U_h$$

with that of finding a solution of the problem involving the surrogate $J_{\text{red}}^P = \hat{J}_{\text{red}} \circ P$:

$$(6.3) \quad \text{minimize } J_{\text{red}}^P(u) = \hat{J}_{\text{red}}(P(u)) \quad \text{over } u \in U_h.$$

When solving (6.3) numerically, one has to evaluate J_{red}^P repeatedly, which, in turn, requires repeated evaluations of the fine model and repeated solutions of the minimization problem (6.1). As we have seen before, given a fixed fine model point u , the computational effort can be reduced by considering the following first order approximation of the space mapping

$$(6.4) \quad P(u + s) \approx P_\ell(u; s) = P(u) + D_P(u)s \in U_H,$$

with $D_P : U_h \mapsto \mathbb{R}^{n_H \times n_h}$ denoting the Jacobian of P . Consequently, J_{red}^P is approximated around u by

$$(6.5) \quad \hat{J}_{\text{red}}(P(u + s)) \approx \hat{J}_{\text{red}}(P_\ell(u; s)).$$

The evaluation of $\hat{J}_{\text{red}}(P_\ell(u; s))$ in (6.5) requires only the computation of the action of $D_P(u)$ on s .

The calculation of the gradient of $J_{\text{red}}^P(u)$ in (6.3) involves $D_P(u)$ in the following way:

$$(6.6) \quad \nabla J_{\text{red}}^P(u) = D_P(u)^\top \nabla \hat{J}_{\text{red}}(\hat{u}) \quad \text{with } \hat{u} = P(u).$$

If we use the approximation (6.4) for P centered at u as a way of computing a step s by minimizing $\hat{J}_{\text{red}}(P_\ell(u; s))$ in (6.5), we also need the evaluation of (6.6). In fact, $\nabla J_{\text{red}}^P(u)$ is the gradient of $\hat{J}_{\text{red}}(P_\ell(u; s))$ with respect to the increment s . The evaluation of $\nabla J_{\text{red}}^P(u)$ requires the computation of the action of $D_P(u)^\top$ on $\nabla \hat{J}_{\text{red}}(\hat{u})$.

6.1. Computation of the sensitivities of the space mapping. In order to characterize $D_P(u)$ or $D_P(u)s$ we need to consider the first order necessary conditions of (6.1), given by

$$(6.7) \quad \begin{aligned} & \alpha_1 D_{\hat{y}}(\hat{u})^\top \hat{M}_{\hat{y}} (\hat{y}(\hat{u}) - K_H^h y(u)) + \alpha_2 \hat{M}_{\hat{u}} (\hat{u} - I_H^h u) \\ & + \alpha_3 [\hat{J}_{\text{red}}(\hat{u}) - J_{\text{red}}(u)] \nabla \hat{J}_{\text{red}}(\hat{u}) = 0, \end{aligned}$$

with $\hat{u} = P(u)$. Above, $D_{\hat{y}}(\hat{u})$ denotes the Jacobian of $\hat{y}(\hat{u})$ with respect to $\hat{u}$. We obtain the characterizing equation for the sensitivities of the space mapping P by differentiation of (6.7) with respect to u . This results in

$$(6.8) \quad \begin{aligned} & \alpha_1 D_{\hat{y}}(\hat{u})^\top \hat{M}_{\hat{y}} (D_{\hat{y}}(\hat{u}) D_P(u) - K_H^h D_y(u)) \\ & + \alpha_1 H_{\hat{y}}(\hat{u}; \hat{M}_{\hat{y}}(\hat{y}(\hat{u}) - K_H^h y(u))) D_P(u) \\ & + \alpha_2 \hat{M}_{\hat{u}} (D_P(u) - I_H^h) + \alpha_3 \nabla \hat{J}_{\text{red}}(\hat{u}) \left(\nabla \hat{J}_{\text{red}}(\hat{u})^\top D_P(u) - \nabla J_{\text{red}}(u)^\top \right) \\ & + \alpha_3 [\hat{J}_{\text{red}}(\hat{u}) - J_{\text{red}}(u)] H_{\hat{J}_{\text{red}}}(\hat{u}) D_P(u) = 0, \end{aligned}$$

with $\hat{u} = P(u)$. In the above equation $H_{\hat{y}}(\hat{u}; z)$ denotes the derivative of $D_{\hat{y}}(\hat{u})z$ with respect to $\hat{u}$, $D_y(u)$ represents the Jacobian of $y(u)$ with respect to u , and $H_{\hat{J}_{\text{red}}}$ is the Hessian of $\hat{J}_{\text{red}}$. Let

$$\begin{aligned} G_{P(u)} &= \alpha_1 D_{\hat{y}}(\hat{u})^\top \hat{M}_{\hat{y}} D_{\hat{y}}(\hat{u}) + \alpha_1 H_{\hat{y}}(\hat{u}; \hat{M}_{\hat{y}}(\hat{y}(\hat{u}) - K_H^h y(u))) \\ &\quad + \alpha_2 \hat{M}_{\hat{u}} + \alpha_3 [\hat{J}_{\text{red}}(\hat{u}) - J_{\text{red}}(u)] H_{\hat{J}_{\text{red}}}(\hat{u}) + \alpha_3 \nabla \hat{J}_{\text{red}}(\hat{u}) \nabla \hat{J}_{\text{red}}(\hat{u})^\top, \end{aligned}$$

with $\hat{u} = P(u)$, and

$$\begin{aligned} r_{P(u)} &= \alpha_1 D_{\hat{y}}(\hat{u})^\top \hat{M}_{\hat{y}} K_H^h D_y(u) + \alpha_2 \hat{M}_{\hat{u}} I_H^h \\ &\quad + \alpha_3 \nabla \hat{J}_{\text{red}}(\hat{u}) \nabla J_{\text{red}}(u)^\top, \end{aligned}$$

also with $\hat{u} = P(u)$. This notation allows us to write (6.8) in a more compact way as

$$G_{P(u)} D_P(u) = r_{P(u)}.$$

For given $s \in U_h$ let us define

$$s_{P(u)} = D_P(u)s \quad \text{and} \quad r_{P(u)}^s = r_{P(u)}s.$$

Then the action of $D_P(u)$ on s , given by $s_{P(u)} = D_P(u)s$, satisfies

$$G_{P(u)} s_{P(u)} = r_{P(u)}^s \quad \text{in } U_H.$$

We point out that in the case where the PDE on the coarse level is linear, one has $H_{\hat{y}} = 0$ and the expression for $G_{P(u)}$ simplifies considerably.

6.2. A practical aggressive space-mapping method for optimal control of PDEs. Now we adapt the aggressive space-mapping method, introduced in [2, 6] and described in Algorithm 4.1, to optimal control of PDEs using the setting and notation chosen in this paper for these problems.

Before we describe the algorithm, we need to adapt some of the notation of sections 4 and 5 to the optimal control framework. In fact, let

$$\nabla \hat{J}_{\text{red}}^k = \nabla \hat{J}_{\text{red}}(P(u_k))$$

and

$$\Delta \hat{J}_{\text{red}}^k = \hat{J}_{\text{red}}(P(u_k + \Delta u_k)) - \hat{J}_{\text{red}}(P(u_k)).$$

As before, we have $\Delta P_k = P(u_k + \Delta u_k) - P(u_k)$, and we use the new Broyden update introduced in section 5 with $\sigma_k \in (0, 1]$:

$$(6.9) \quad \widetilde{\Delta P_k} = \Delta P_k + \sigma_k \frac{\Delta \hat{J}_{\text{red}}^k - (\nabla \hat{J}_{\text{red}}^k)^\top \Delta P_k}{\|\nabla \hat{J}_{\text{red}}^k\|_2^2} \nabla \hat{J}_{\text{red}}^k.$$

ALGORITHM 6.1. *Aggressive space-mapping method for optimal control of PDEs.* Choose $u_0 \in \mathbb{R}^n = \mathbb{R}^{n_h}$, $\Delta_0 > 0$, $B_0 \in \mathbb{R}^{\hat{n} \times n} = \mathbb{R}^{n_H \times n_h}$, and $\gamma_1, \eta_1 \in (0, 1)$.

0. Compute $P(u_0)$ by solving (6.1) with $u = u_0$.

For $k = 0, 1, 2, \dots$

1. Compute an approximated solution Δu_k for the trust-region subproblem

$$(6.10) \quad \text{minimize } \hat{g}(P(u_k) + B_k \Delta u) \quad \text{subject to } \|\Delta u\| \leq \Delta_k,$$

over $\Delta u \in \mathbb{R}^n = \mathbb{R}^{n_h}$.

2. Compute $P(u_k + \Delta u_k)$ by solving (6.1) with $u = u_k + \Delta u_k$.
3. Compute the ratio between actual and predicted reductions:

$$\rho_k = \frac{\text{ared}(u_k, \Delta u_k)}{\text{pred}(u_k, \Delta u_k)} = \frac{\hat{g}(P(u_k)) - \hat{g}(P(u_k + \Delta u_k))}{\hat{g}(P(u_k)) - \hat{g}(P(u_k) + B_k \Delta u_k)}.$$

4. If $\rho_k \geq \eta_1$, then $u_{k+1} = u_k + \Delta u_k$ and Δ_{k+1} is chosen so that $\Delta_{k+1} \geq \Delta_k$. In this case, update B_{k+1} using Broyden's formula:

$$(6.11) \quad B_{k+1} = B_k + \frac{\widetilde{\Delta P_k} - B_k \Delta u_k}{\|\Delta u_k\|_2^2} \Delta u_k^\top,$$

where $\widetilde{\Delta P_k}$ is given by (6.9) with $\Delta P_k = P(u_k + \Delta u_k) - P(u_k)$ and $\sigma_k \in (0, 1]$. Put $P(u_{k+1}) = P(u_k + \Delta u_k)$.

5. If $\rho_k < \eta_1$, then $u_{k+1} = u_k$ and $\Delta_{k+1} = \gamma_1 \Delta_k$. Keep $B_{k+1} = B_k$ and $P(u_{k+1}) = P(u_k)$.

end

The comments made about the norm used to shape the trust region and about the mechanisms to manage the size of the trust radius remain pertinent here. When $H = h$ the initial value for B can be given by the classical choice $B_0 = I_{n_h}$, with I_{n_h} the $n_h \times n_h$ identity matrix. When $H > h$ we can choose $B_0 = I_H^h$.

In analogy to the aggressive space-mapping method of section 4, one would expect that

$$(6.12) \quad \hat{g}(P(u_k)) = \hat{J}_{\text{red}}(P(u_k))$$

and

$$(6.13) \quad \hat{g}(P(u_k) + B_k \Delta u) = \hat{J}_{\text{red}}(P(u_k) + B_k \Delta u).$$

However, in the case where $H > h$, this last choice would result in an underdetermined problem in step 1 of the algorithm, in the sense that $\Delta u \in \mathbb{R}^{n_h}$ is a fine grid quantity, whereas $\hat{g}$ is defined in the coarse grid setting (yielding a singular Hessian in $\hat{J}_{\text{red}}(P(u_k) + B_k \Delta u)$). There exist two immediate remedies to this situation.

(i) One possibility is to use

$$(6.14) \quad \hat{g}(P(u_k) + B_k \Delta u) = \hat{J}_{\text{red}}(P(u_k) + B_k \Delta u) + \frac{\gamma}{2} \|u_k + \Delta u - u_d\|_{M_u}^2$$

and

$$(6.15) \quad \hat{g}(P(u_k)) = \hat{J}_{\text{red}}(P(u_k)) + \frac{\gamma}{2} \|u_k - u_d\|_{M_u}^2,$$

where $\gamma > 0$ and u_d denotes some reference value for the expected optimal control. For instance, u_d can be obtained by interpolating coarse grid solutions (easy to obtain) to the fine grid. The parameter γ plays the role of a regularization parameter which penalizes deviations of $u_k + \Delta u$ from u_d . In our numerical tests, γ is chosen according to the mesh sizes H and h in the following way:

$$\gamma = c_\gamma (1 - h/H) \quad \text{with} \quad 0 < c_\gamma \ll 1.$$

Note that when $H = h$ we have $\gamma = 0$ and no regularization takes place (and the coarse model in step 1 is likely not underdetermined in the sense discussed above).

(ii) An alternative remedy, using the original choices (6.12)–(6.13), is given by solving an approximate problem of the type

$$(6.16) \quad \text{minimize } \hat{g}(P(u_k) + B_k I_h^H \Delta \hat{u}) \quad \text{subject to } \|\Delta \hat{u}\| \leq \Delta_k$$

instead of problem (6.10) in step 1. By using a prolongation operator $I_h^H : U_H \rightarrow U_h$, the independent variable $\Delta \hat{u}$ is mapped to a fine grid quantity and the new problem is usually well-determined. Again, whenever $h = H$ we may choose $I_h^H = I_{n_h}$, and problem (6.16) becomes the original problem (6.10).

Both remedies have additional costs. The first one requires the computation of the reference value u_d and the second one the application of the prolongation operator I_h^H . However, in the latter case only an n_H -dimensional problem has to be solved.

7. Computation of coarse and fine model derivatives.

7.1. Adjoint calculation of the coarse model gradient and Hessian. The computation of the gradient $\nabla \hat{J}_{\text{red}}(\hat{u})$ can be carried out by the so-called adjoint technique. In what follows we briefly explain some of the details.

Let $\hat{E}(\hat{y}, \hat{u}) = 0$ denote the discretized PDE on the coarse grid. Further let $\hat{E}_{\hat{y}}$, $\hat{E}_{\hat{u}}$ denote the partial Jacobians of $\hat{E}$ with respect to $\hat{y}$ and $\hat{u}$, respectively. From the assumption that the state equation admits a unique solution $y(u)$ for $u \in U$, we infer that there exists a unique $\hat{y}(\hat{u})$ such that $\hat{E}(\hat{y}(\hat{u}), \hat{u}) = 0$ and that $\hat{E}_{\hat{y}}(\hat{y}(\hat{u}), \hat{u})$ is invertible (at least for sufficiently small H). Differentiation of $\hat{E}(\hat{y}(\hat{u}), \hat{u}) = 0$ with respect to $\hat{u}$ yields

$$(7.1) \quad \hat{E}_{\hat{y}}(\hat{y}(\hat{u}), \hat{u}) D_{\hat{y}}(\hat{u}) + \hat{E}_{\hat{u}}(\hat{y}(\hat{u}), \hat{u}) = 0.$$

Hence, we obtain from (7.1)

$$(7.2) \quad D_{\hat{y}}(\hat{u}) = -\hat{E}_{\hat{y}}(\hat{y}(\hat{u}), \hat{u})^{-1} \hat{E}_{\hat{u}}(\hat{y}(\hat{u}), \hat{u}).$$

From the definition of $\hat{J}_{\text{red}}$ we deduce that

$$(7.3) \quad \nabla \hat{J}_{\text{red}}(\hat{u}) = \nabla_{\hat{u}} \hat{J}(\hat{y}(\hat{u}), \hat{u}) + D_{\hat{y}}(\hat{u})^\top \nabla_{\hat{y}} \hat{J}(\hat{y}(\hat{u}), \hat{u}),$$

where $\nabla_{\hat{y}} \hat{J}$, $\nabla_{\hat{u}} \hat{J}$ represent the partial derivatives of $\hat{J}$ with respect to the first and second argument, evaluated in (7.3) at $(\hat{y}(\hat{u}), \hat{u})$. Utilizing (7.2) in (7.3) yields

$$(7.4) \quad \nabla \hat{J}_{\text{red}}(\hat{u}) = \nabla_{\hat{u}} \hat{J}(\hat{y}(\hat{u}), \hat{u}) + \hat{E}_{\hat{u}}(\hat{y}(\hat{u}), \hat{u})^\top \hat{p}(\hat{u})$$

with

$$(7.5) \quad \hat{E}_{\hat{y}}(\hat{y}(\hat{u}), \hat{u})^\top \hat{p}(\hat{u}) = -\nabla_{\hat{y}} \hat{J}(\hat{y}(\hat{u}), \hat{u}).$$

Equation (7.5) is the so-called (discrete) adjoint equation. For computing $\nabla \hat{J}_{\text{red}}(\hat{u})$ one can proceed as follows: Given $\hat{u}$, solve the state equation for $\hat{y}(\hat{u})$, then solve the adjoint equation (7.5) for $\hat{p}(\hat{u})$, and finally compute the gradient according to (7.4).

By using the definition

$$\hat{W}(\hat{y}(\hat{u}), \hat{u}) = \begin{pmatrix} D_{\hat{y}}(\hat{u}) \\ I_{n_H} \end{pmatrix},$$

it is possible to rewrite (7.3) as

$$\nabla \hat{J}_{\text{red}}(\hat{u}) = \hat{W}(\hat{y}(\hat{u}), \hat{u})^\top \nabla \hat{J}(\hat{y}(\hat{u}), \hat{u}).$$

Also, it is possible to show (see, e.g., [17]) that the Hessian of the coarse model $\hat{J}_{\text{red}}(\hat{u})$ is given by

$$H_{\hat{J}_{\text{red}}}(\hat{u}) = H_{\hat{y}}(\hat{u}; \nabla_{\hat{y}} \hat{J}(\hat{y}(\hat{u}), \hat{u})) + \hat{W}(\hat{y}(\hat{u}), \hat{u})^\top H_{\hat{J}}(\hat{y}(\hat{u}), \hat{u}) \hat{W}(\hat{y}(\hat{u}), \hat{u}).$$

When the state equation in (1.1b) is linear, i.e., when $\hat{E}(\hat{y}, \hat{u}) = \hat{L}\hat{y} + \hat{M}\hat{u} - \hat{f}$, where $\hat{M}$ and $\hat{L}$ are suitable matrices with $\hat{L}$ invertible and $\hat{f}$ a coarse model vector, and when the cross derivatives $\hat{J}_{\hat{y}\hat{u}}(\cdot)$ and $\hat{J}_{\hat{u}\hat{y}}(\cdot)$ are zero, one can simplify considerably the expression for the Hessian of the coarse model $\hat{J}_{\text{red}}(\hat{u})$. The assumption $\hat{J}_{\hat{y}\hat{u}}(\cdot) = \hat{J}_{\hat{u}\hat{y}}(\cdot) = 0$ is satisfied for the commonly used objective functional of tracking type, i.e., for

$$\mathcal{J}(y, u) = \frac{1}{2} \|y - y_d\|_{L^2(\Omega)}^2 + \frac{\delta}{2} \|u\|_{L^2(\Omega)}^2,$$

with $y_d \in L^2(\Omega)$ and $\delta > 0$ fixed. Under the simplified assumptions of this paragraph, the model Hessian becomes

$$H_{\hat{J}_{\text{red}}}(\hat{u}) = \hat{L}^{-\top} \hat{J}_{\hat{y}\hat{y}}(\hat{L}^{-1}(\hat{f} - \hat{M}\hat{u}), \hat{u}) \hat{L}^{-1} + \hat{J}_{\hat{u}\hat{u}}(\hat{L}^{-1}(\hat{f} - \hat{M}\hat{u}), \hat{u}).$$

7.2. Approximation of the fine model gradient. The gradient ∇J_{red} of the fine model can be computed also by using the adjoint technique of section 7.1. Per each gradient evaluation, this technique requires one solve for the (possibly nonlinear) state equation and one for the (linear) adjoint equation.

Since we are working on the fine model these evaluations might be extremely costly. One way to reduce or avoid fine model solves is based on restriction operators I_H^h and their analogues, prolongation operators I_h^H .

Alternatively, each fine model gradient evaluation can be calculated by a hybrid approach based on a fine model adjoint solve and a coarse model solve of the state equation.

In what follows we describe these techniques for computing ∇J_{red} depending on whether $H > h$ (and both fine and coarse models are nonlinear) or $H = h$ (and the coarse model is linear). We present this material because of its relevance in the context of this paper despite the fact that we do not make use of any approximation to the gradient of the fine model in our numerical testing.

7.2.1. The case $H > h$ (fine and coarse models are nonlinear). Let I_h^H denote the (linear) prolongation operator from U_H to U_h . Analogously, one also introduces K_h^H . In the case where the fine and the coarse models of the PDE are nonlinear, a suitable approximation of the gradient is given by

$$(7.6) \quad \nabla J_{\text{red}}^{\text{app}}(u) = I_h^H \nabla \hat{J}_{\text{red}}(I_H^h u);$$

i.e., we restrict the fine model point $u \in U_h$ to the coarse setting U_H by using I_H^h , evaluate the gradient on the coarse level by means of the adjoint technique, and then we transfer the coarse model gradient back to the fine model setting with the help of I_h^H .

Alternatively, one can use a hybrid approach which combines coarse and fine model solves and which is still numerically less expensive than the full fine model approach. The hybrid technique is particularly useful when the fine model involves

nonlinearities. In fact, we can compute

$$(7.7) \quad \nabla J_{\text{red}}^{\text{app}}(u) = \nabla_u J(K_h^H \hat{y}(I_H^h u), u) + E_u(K_h^H \hat{y}(I_H^h u), u)^\top p(u)$$

with

$$(7.8) \quad E_y(K_h^H \hat{y}(I_H^h u), u)^\top p(u) = -\nabla_y J(K_h^H \hat{y}(I_H^h u), u).$$

The advantage of this strategy is related to the fact that the nonlinear state equation must be solved only on the coarse grid for a given $I_H^h u$. On the fine grid, one has to solve the (linear) adjoint equation. Typically, solving linear equations is significantly less expensive than computing solutions to nonlinear ones. Thus, the hybrid approach is less expensive than computing ∇J_{red} by the adjoint technique on the fine grid.

Using (7.6), or the hybrid approach in (7.7) and (7.8), yields the approximate sensitivity $D_P^{\text{app}}(u)$ and the approximate action $s_{P(u)}^{\text{app}}$. We remark that the accuracies of these approximations can be controlled by tuning the mesh size H . In fact, in the extreme case $H = h$ with $I_H^h = I_h^H = I_{n_h}$ (I_{n_h} the $n_h \times n_h$ identity matrix) only exact quantities are computed.

7.2.2. The case $H = h$ (coarse model is linear). In this case (7.7) would require the full adjoint technique on the fine grid. In order to reduce the computational burden, one may consider as the coarse model a linear approximation of the discretized PDE. If the linear equation can be solved efficiently (e.g., by fast Fourier transformation techniques), then a suitable approximate gradient is given by

$$\nabla J_{\text{red}}^{\text{app}}(u) = \nabla_u J(\hat{y}_L(u), u) + E_u(\hat{y}_L(u), u)^\top p(u)$$

with

$$E_y(\hat{y}_L(u), u)^\top p(u) = -\nabla_y J(\hat{y}_L(u), u)$$

and $\hat{y}_L(u)$ denoting the solution of the linear coarse model $\hat{E}(\hat{y}, u) = 0$. Here we assume $\hat{u} = u$. Clearly, the approximation properties depend now on the error between the linear coarse model and the nonlinear fine model.

8. Numerical experiments. Let us now report some numerical results attained by the aggressive space-mapping method for the optimal control of PDEs. Our test examples are of the following type:

$$(8.1a) \quad \text{minimize } \frac{1}{2} \|y - y_d\|_{L^2(\Omega)}^2 + \frac{\delta}{2} \|u\|_{L^2(\Omega)}^2 \quad \text{over } (y, u) \in H_0^1(\Omega) \times L^2(\Omega)$$

$$(8.1b) \quad \text{subject to } -\nu \Delta y + f(y) = u \quad \text{in } \Omega = (0, 1)^2,$$

with $y_d \in L^2(\Omega)$ and $\nu, \delta > 0$. Here f denotes some nonlinear mapping in y . Note that the parameter $\nu > 0$ allows us to emphasize the nonlinear term $f(y)$ by considering $0 < \nu \ll 1$.

We use a standard five-point stencil for discretizing the Laplacian with homogeneous Dirichlet boundary conditions. The prolongation operators I_h^H, K_h^H and restriction operators I_H^h, K_H^h are chosen as follows: Motivated by an a posteriori analysis (in function spaces) of a solution (y, u) to our control problem (see, e.g., [1]), we choose $K_h^H = I_h^H$ and $K_H^h = I_H^h$. The interpolation from the coarse to the fine grid, i.e., I_h^H ,

is achieved by a nine-point prolongation. Its stencil is symbolized by

$$\begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}.$$

The restriction I_H^h is the adjoint of the nine-point prolongation with symbol

$$\frac{1}{16} \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}.$$

For more details on prolongation and restriction operators of the above kind, we refer the reader to, e.g., [16].

For the numerical solution of the discretized counterpart of the nonlinear PDE involved in (8.1b) we use the Newton conjugate gradient method [20]. The discrete linearized PDE as well as the discrete adjoint equation are solved by means of the conjugate gradient method.

In the aggressive space-mapping method for optimal control, i.e., Algorithm 6.1, we use the following adjustment strategy for the trust radius Δ_k : Let $0 < \eta_1 \leq \eta_2 < 1$, $\gamma_1 \in (0, 1)$, and $\xi_1 > 1$ be given. If $\rho_k \geq \eta_2$, then accept the current step and enlarge the trust radius by $\Delta_{k+1} = \xi_1 \Delta_k$. If $\eta_1 \leq \rho_k < \eta_2$, then the current step is accepted and the trust radius is kept, i.e., $\Delta_{k+1} = \Delta_k$. Finally, whenever $\rho_k < \eta_1$, then the trust radius is reduced by $\Delta_{k+1} = \gamma_1 \Delta_k$ without accepting the current step. In the examples reported below we used $\eta_1 = 10^{-5}$, $\eta_2 = 10^{-1}$, $\gamma_1 = 0.25$, and $\xi_1 = 2$. We initialize the trust radius as $\Delta_0 = 50$.

The Broyden update procedure of Algorithm 6.1 is based on a full limited memory version of Broyden's method [9]. In fact, since n_h is typically very large, and B_k tends to be a dense matrix, storing B_k is infeasible in the context of optimal control problems for PDEs. Rather we store the vectors $\{\Delta u_i\}_{i=0}^k$ and $\{\Delta P_i\}_{i=0}^k$ and perform the product $B_k v$ of the Broyden's matrix B_k by a vector $v \in \mathbb{R}^{n_h}$ using vector×vector-multiplications only. We initialize the Broyden matrix as $B_0 = I_H^h$.

In the examples below the fine model consists of the fully nonlinear PDE discretized uniformly on the fine grid with mesh size h , resulting in n_h unknowns in the reduced fine model problem (6.2) with $U_h = \mathbb{R}^{n_h}$. This PDE has to be solved once per iteration of Algorithm 6.1. In order to reduce the computational cost we use an inexact iterative solution technique; i.e., the stopping rule for the iterative solver for the fine model nonlinear PDE becomes increasingly stringent as the iterates of Algorithm 6.1 approach the solution. The coarse model is given by the discretization of the linearized PDE on the coarse grid with mesh size H . The linearization is performed with respect to $\hat{y}_k^\dagger$, which results in

$$\nu \hat{A} \hat{y} + D_{\hat{f}}(\hat{y}_k^\dagger) \hat{y} = \hat{u} - \hat{f}(\hat{y}_k^\dagger) + D_{\hat{f}}(\hat{y}_k^\dagger) \hat{y}_k^\dagger,$$

with $\hat{u}, \hat{y}, \hat{y}_k^\dagger \in \mathbb{R}^{n_H}$. Above, the $n_H \times n_H$ -matrix $\hat{A}$ represents the discretization of the operator $-\Delta$ with homogeneous Dirichlet boundary conditions on the (uniform) coarse grid with mesh size H . In all test runs reported below, we chose $u_0 = 0$, $\hat{y}_0^\dagger = 0$, and $\hat{y}_k^\dagger = I_H^h y_k$, where y_k solves the discretized nonlinear PDE for $u = u_k$ on the fine grid.

We use (6.14)–(6.15) for $\hat{g}$ in the trust-region subproblem in Algorithm 6.1. Unless otherwise specified, the reference value u_d is chosen as $u_d = 0$ in all iterations. The

corresponding regularization parameter γ is reported in the examples below. The norm used to define the trust region is the ℓ_∞ one.

Algorithm 6.1 was stopped when

$$\max\{|pred(u_k, \Delta u_k)|, H\|\nabla \hat{J}_{\text{red}}(P(u_k))\|_2, h\|\Delta u_k\|_2\} \leq \text{tol},$$

where $\text{tol} = \epsilon_1 H\|\nabla \hat{J}_{\text{red}}(P(u_0))\|_2 + \epsilon_2$, with $0 < \epsilon_2 \ll \epsilon_1$. Unless otherwise specified, we chose $\epsilon_1 = 10^{-5}$ and $\epsilon_2 = 10^{-14}$.

We used a globalized semismooth Newton method [15] for the solution of the minimization problem subject to bounds on the variables, namely, problem (6.10), which is a quadratic programming problem with simple bounds. The unconstrained problem (6.1) is solved by an inexact Newton method. We made a modification in the Hessian of the objective in (6.1) that consisted of neglecting the term involving the Hessian of $\hat{J}_{\text{red}}(\hat{u})$. As a result we obtained a positive definite approximation of the Hessian in all iterations.

In the examples below we also report on results obtained by a genuine nonlinear multigrid method applied to the first order optimality system for the discrete analogue of (8.1):

$$(8.2a) \quad \nu Ay + f_h(y) - u = 0,$$

$$(8.2b) \quad y + \delta f'_h(y)u + \nu \delta Au = y_d,$$

where A denotes the discrete Laplacian on the fine grid, and $f_h(y)$, y , u , $f'_h(y)u$, and y_d are vectors in $\mathbb{R}^{n_h}$. We implemented the FAS scheme [8, 16] with a balanced nonlinear Gauss–Seidel smoother which approximates the solution of the scalar nonlinear equations (one per grid point) by applying one Newton step. Below we use the shortcut `NMG_OC` for our nonlinear multigrid solver.

8.1. Example 1. The first example is related to a simplified Ginzburg–Landau model for superconductivity [18, 19]. The data are as follows:

$$y_d = \frac{1}{6} \sin(2\pi x_1) \sin(2\pi x_2) \exp(2x_1), \quad f(y) = y^3 + y,$$

and $\delta = \nu = 10^{-3}$. Figure 8.1 shows the optimal control and the optimal state of (8.1), with data as specified before, computed on a 255×255 grid.

In Table 8.1 we report the results obtained from our space-mapping algorithm, Algorithm 6.1, with $\alpha_1 = \alpha_2 = 100$, $\alpha_3 = 10^{-5}$, and $\gamma = 10^{-3}(1 - h/H)$. The parameter σ_k (see (6.9)) was set to $\sigma_k = 0.1$ for all k . By *level* we denote the number of grid coarsenings, i.e., $H = 2^{\text{level}}h$. Furthermore, *#it* denotes the number of iterations until successful termination, and *CPU-ratio* represents the ratio between the CPU-time required by the space-mapping method versus the CPU-time elapsed by `NMG_OC` when applied to the fine model problem and stopped as soon as it reaches the norm for the residual in (8.2) of the space-mapping solution. Finally, *res-ratio* is the ratio between the residual in (8.2) of the space-mapping solution and the one of the coarse grid solution (computed by `NMG_OC`) after prolongation.

From Table 8.1 we can see that the new space-mapping method produces more accurate approximations in a significantly smaller amount of CPU-time than `NMG_OC` if $n_H \ll n_h$. We further point out that *res-ratio* = 0.185 if we compare the space-mapping solution for $n_H = 15^2$, $n_h = 255^2$ with the prolonged solution on a 127×127 grid. This shows that the space mapping P contains a substantial amount of fine grid information. As one would expect, if the level of coarsening is decreased, the accuracy

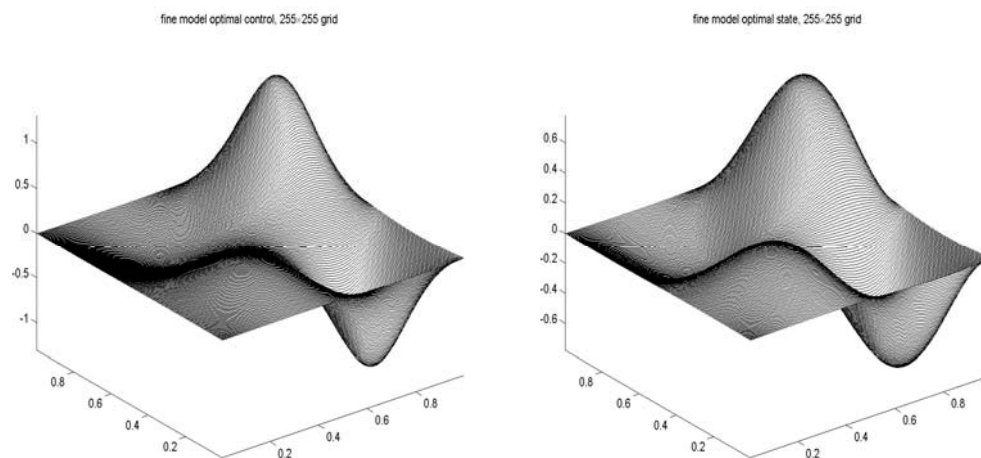


FIG. 8.1. Optimal control (left) and optimal state (right) for the simplified Ginzburg–Landau model on a 255×255 grid.

TABLE 8.1

Results of space mapping versus fine model solution and interpolated coarse model solutions for Example 1.

n_h	level	n_H	# it	CPU-ratio	res-ratio
255^2	4	15^2	4	0.089	0.0443
255^2	3	31^2	4	0.105	0.0269
255^2	2	63^2	4	0.307	0.0110
255^2	1	127^2	4	1.040	0.0206

as well as the computation time for the space-mapping method are increased. In our tests it turns out that there is a trade-off coarse mesh size H at which $CPU\text{-ratio} \approx 1$, but still we have $res\text{-ratio} \ll 1$. However, let us reemphasize that in this paper, space mapping is not designed to be a new fast fine grid solver. Rather it is a tool which, in contrast to classical multigrid techniques, allows us to combine different models on different levels for a fast computation of approximate solutions.

In Figure 8.2 we display the controls obtained by the space-mapping method for $n_H = 15^2$, $n_h = 127^2$ (top left graph) and $n_h = 255^2$ (top right graph), respectively. Furthermore, in Figure 8.2, we plot the difference in absolute value between the optimal controls obtained from the fine model and from our space-mapping technique. The graphs in the second row of Figure 8.2 show that the error between the space-mapping solution and the true solution of the fine model behaves rather stably with respect to the coarse level. Indeed, the coarse level for both results is $H = 1/16$, while the fine levels are $h = 1/128$ and $h = 1/256$, respectively. In Figure 8.3 we further investigate the dependence of the error on the levels of coarsening. Now we use $h = 1/128$ and $H = 1/32$, which corresponds to $level = 2$ (compared to $level = 3$ previously). From the graphs in Figure 8.3 we conclude that the error is significantly reduced. Also, the graph of the space-mapping solution appears to be smoother compared to the ones in Figure 8.2. This is related to the fact that the restriction and prolongation operators approach the unit matrix as the level of coarsening decreases.

In the following we briefly comment on the effect of the new Broyden update (6.11). In Table 8.2 we compare the new Broyden update with $\sigma_k = 0.1$ to the

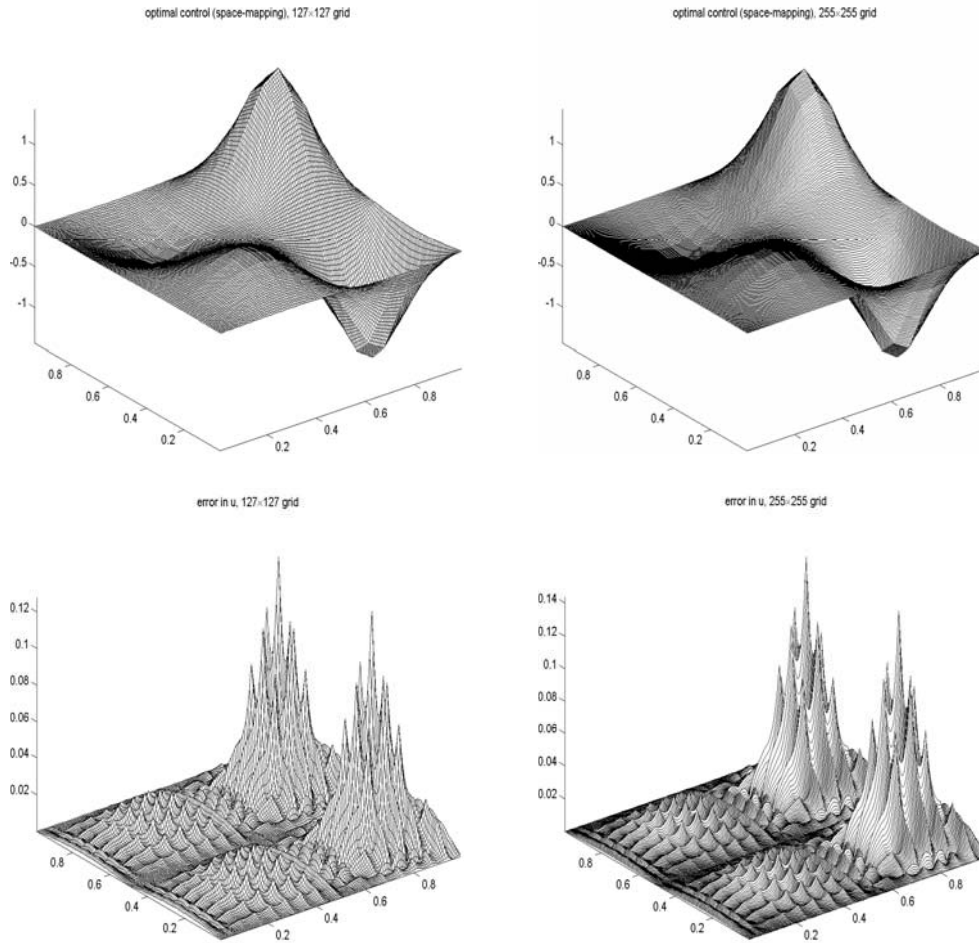


FIG. 8.2. Optimal controls obtained by Algorithm 6.1 (upper plots) and differences to the fine model solutions (lower plots) for $n_h = 127^2$ (left column), $n_h = 255^2$ (right column), and $n_H = 15^2$, respectively.

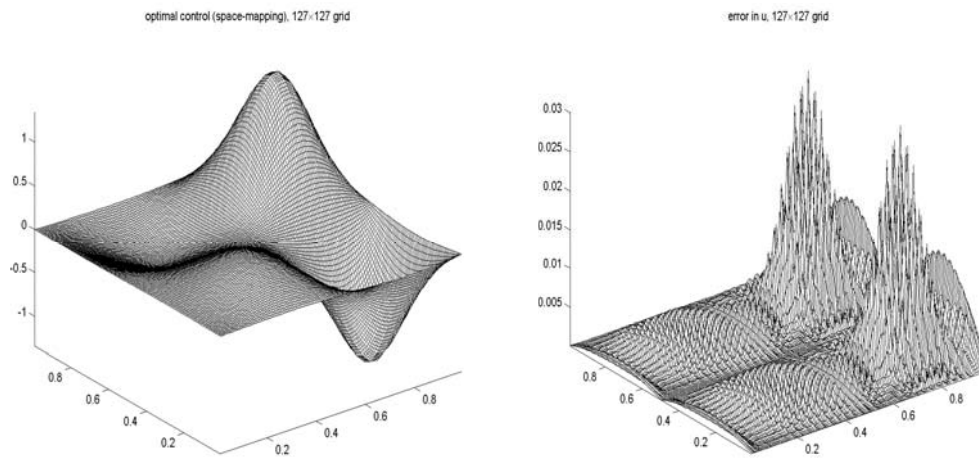


FIG. 8.3. Optimal control obtained by Algorithm 6.1 for $h = 1/128$ (left plot) and the difference to the fine model solution (right plot) for level = 2.

TABLE 8.2
Comparison between the new and the classical Broyden update for Example 1.

σ_k	n_h	level	n_H	# it	CPU-ratio
0.1	255^2	4	15^2	4	0.072
0.0	255^2	4	15^2	4	0.089

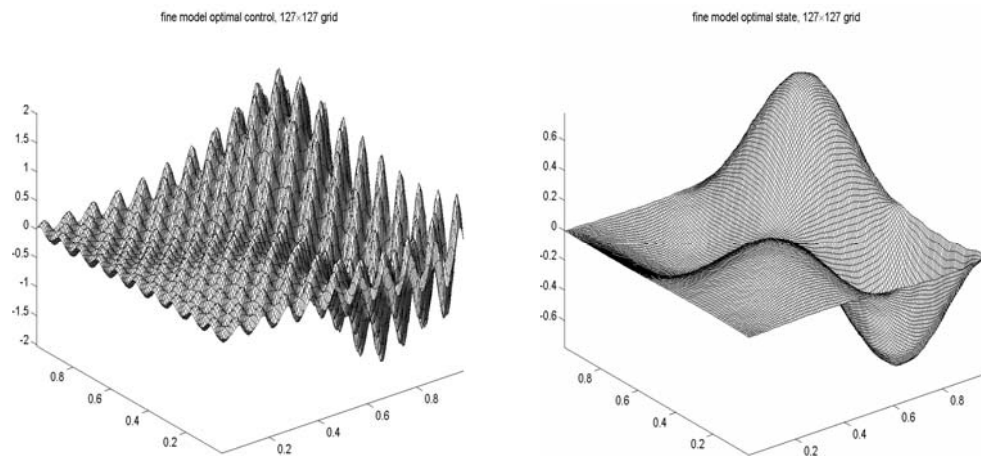


FIG. 8.4. Optimal control (left) and optimal state (right) for Example 2 on a 127×127 grid.

classical Broyden update, i.e., $\sigma_k = 0$ for all k . The results in Table 8.2 indicate that the new Broyden update reduces the computation time and, according to our numerical experience, sometimes also the number of iterations of the new space-mapping algorithm. In general, we found that the behavior of the new method depends on the choice of σ_k . In our test runs for Example 1, the choices $\sigma_k \in [0.1, 0.001)$ yielded results comparable to $\sigma_k = 0.1$ for all k . For $\sigma_k < 0.001$ there was no significant difference between the new and the classical Broyden update. The choice $\sigma_k > 0.1$ typically degraded the performance of the method when compared to runs with $\sigma_k = 0.1$.

8.2. Example 2. The following example shows that the space-mapping method benefits from eventual evaluations of the fine model and the possibility of specifying u_d in step 1. In fact, when considering steps 2 and 4 of Algorithm 6.1 we find that every computation of the space mapping P is based on the matching of fine model solves. This fact is highlighted in our definition (6.1) of P , where u and $y(u)$ are fine model quantities. As a consequence, we expect that the space-mapping solution yields a better approximation to the fine model solution than, e.g., interpolated coarse model solutions. This is also true when the coarse mesh does not capture oscillations which exist on the fine mesh.

The data for Example 2 are the same as those for Example 1 except for f . Now we have

$$f(y) = y^3 + y + f_0, \text{ with } f_0(x) = \frac{1}{6} \sin(20\pi x_1) \sin(20\pi x_2) \exp(2x_1).$$

The zero order term f_0 induces oscillations to the optimal control, as can be seen from Figure 8.4, which displays the optimal control and the corresponding optimal state for the fine model problem on a 127×127 grid. We ran Algorithm 6.1 with $\alpha_1 = 75.0$,

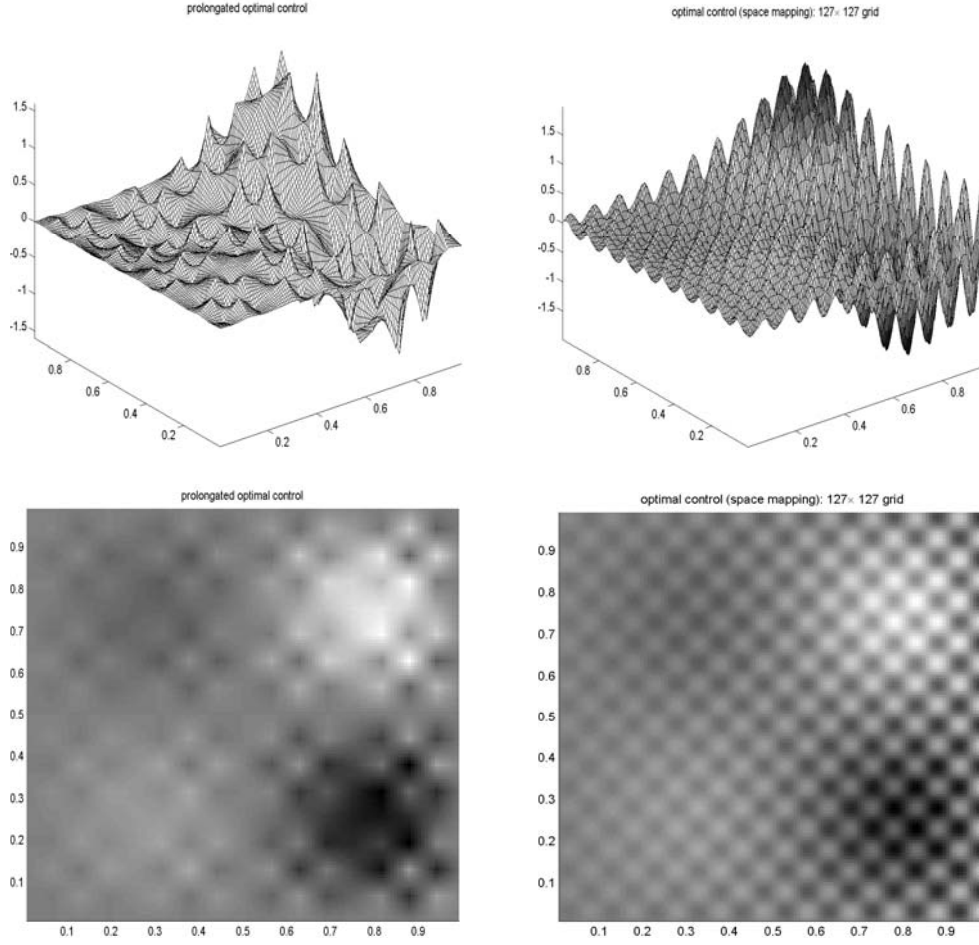


FIG. 8.5. *Interpolated optimal control (left column) and space-mapping solution (right column) for Example 2.*

$\alpha_2 = 0.5$, $\alpha_3 = 10^{-5}$, $\gamma = 2.25 \cdot 10^{-1}(1 - h/H)$, $\sigma_k = 10^{-3}$ for all k , and $level = 3$ for $h = 1/128$. Further, we set $u_d = f_0$ in all iterations.

Figure 8.5 shows (in the upper left part) the prolonged coarse grid optimal control, i.e., $u_{pro} = K_h^H \bar{u}$, where $\bar{u}$ is the optimal solution of the nonlinear optimal control problem on the coarse grid, as well as the space-mapping solution (in the upper right part). The figures in the lower part show the top view of both solutions (left for prolonged and right for space mapping). First of all, we point out that there is a significant difference in the scale of both solutions, as can be seen from the graphs in the first row of Figure 8.5. The space-mapping scaling is significantly closer to the fine model one. The lower plots show that the space-mapping solution identified more of the fine model resolution. This is a clear indication of what we mentioned before, in the sense that fine model information has a beneficial impact on the quality of the solution obtained by space mapping.

Finally we mention that we also tested the following variant of Algorithm 6.1: Similar to the nested iteration concept, we start at a coarse-fine grid pair with coarse mesh size $H = 1/2^i$, a fine mesh size $h = 1/2^{i+1}$, and $i = 2$. Algorithm 6.1 was run

until its stopping rule was satisfied. Then the solution was interpolated to the next finer grid, and i was increased by 1, and this interpolated solution was used as the starting point for the run of Algorithm 6.1 on the finer grid. As soon as the coarse grid mesh size reached the specified value, only the fine mesh was further refined until its mesh size reached its specified value. For each grid pair Algorithm 6.1 was run until convergence and then the solution was interpolated to the next finer level. This procedure typically decreased slightly both the overall runtime and *res-ratio*.

9. Conclusions and future work. In this paper we have investigated the use of the space-mapping technique in the numerical solution of optimal control problems governed by PDEs. We have identified a space-mapping framework for this purpose that allows the integration of different coarse models, arising from linearizing and/or coarsening the fine model. The new definition for the space mapping that we introduced uses the concept of Tikhonov-type regularization as a way of finding the coarse (control and state) variables closest to some corresponding fine model values. We have also suggested a new Broyden update to approximate the derivatives of the space mapping, with broad applicability to most of the existent space-mapping approaches.

A number of issues require further investigation. In this paper we have not considered, for instance, optimal control problems with constraints on the control variables like simple bounds. Adapting our approach to cover this case is relatively straightforward, but it would add another layer of complexity in the numerical computations.

A topic for future research is the use of more than one coarse model in the space-mapping approach. The existence of, say, two coarse models with increasing level of accuracy and cost of evaluation is an appealing idea in some application problems. Another aspect that has not been considered in this paper is the appropriate use of different optimization algorithms for coarse and fine models along the spirit of multigrid methods.

Acknowledgment. Luís Vicente would like to thank the IBM T.J. Watson Research Center and the Institute for Mathematics and Its Applications for their local support.

REFERENCES

- [1] N. ARADA, E. CASAS, AND F. TRÖLTZSCH, *Error estimates for the numerical approximation of a semilinear elliptic control problem*, Comp. Optim. Appl., 23 (2002), pp. 201–229.
- [2] M. H. BAKR, J. W. BANDLER, R. M. BIERNACKI, S. H. CHEN, AND K. MADSEN, *A trust region aggressive space mapping algorithm for EM optimization*, IEEE Trans. Microwave Theory Tech., 46 (1998), pp. 2412–2425.
- [3] M. H. BAKR, J. W. BANDLER, K. MADSEN, AND J. SØNDERGAARD, *Review of the space mapping approach to engineering optimization and modeling*, Optim. Engrg., 1 (2000), pp. 241–276.
- [4] M. H. BAKR, J. W. BANDLER, K. MADSEN, AND J. SØNDERGAARD, *An introduction to the space mapping technique*, Optim. Engrg., 2 (2002), pp. 369–384.
- [5] J. W. BANDLER, R. M. BIERNACKI, S. H. CHEN, P. A. GROBELNY, AND R. H. HEMMERS, *Space mapping technique for electromagnetic optimization*, IEEE Trans. Microwave Theory Tech., 42 (1994), pp. 2536–2544.
- [6] J. W. BANDLER, R. M. BIERNACKI, S. H. CHEN, R. H. HEMMERS, AND K. MADSEN, *Electromagnetic optimization exploiting aggressive space mapping*, IEEE Trans. Microwave Theory Tech., 43 (1995), pp. 2874–2882.
- [7] J. W. BANDLER AND K. MADSEN, *Editorial—Surrogate modelling and space mapping for engineering optimization*, Optim. Engrg., 2 (2002), pp. 367–368.
- [8] A. BRANDT, *Multi-level adaptive solutions to boundary-value problems*, Math. Comp., 31 (1977), pp. 333–390.

- [9] R. BYRD, J. NOCEDAL, AND R. SCHNABEL, *Representations of quasi-Newton matrices and their use in limited-memory methods*, Math. Programming A, 63 (1994), pp. 129–156.
- [10] R. G. CARTER, *On the global convergence of trust region algorithms using inexact gradient information*, SIAM J. Numer. Anal., 28 (1991), pp. 251–265.
- [11] A. R. CONN, N. I. M. GOULD, AND PH. L. TOINT, *Trust-Region Methods*, MPS/SIAM Ser. Optim. 1, SIAM, Philadelphia, 2000.
- [12] R. DAUTRAY AND J.-L. LIONS, *Analyse Mathématique et Calcul Numérique* 3, Masson, Paris, 1987.
- [13] J. E. DENNIS, *Surrogate Modelling and Space Mapping for Engineering Optimization. A Summary of the Danish Technical University November 2000 Workshop*, Tech. Report TR00–35, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 2000.
- [14] J. E. DENNIS AND R. B. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice–Hall, Englewood Cliffs, NJ, 1983. Reprinted, Classics Appl. Math. 16, SIAM, Philadelphia, 1996.
- [15] F. FACCHINEI AND J.-S. PANG, *Finite-dimensional variational inequalities and complementarity problems, Vol. II*, Springer Ser. Oper. Res., Springer-Verlag, New York, 2003.
- [16] W. HACKBUSCH, *Multigrid Methods and Applications*, Springer Ser. Comput. Math. 4, Springer-Verlag, Berlin, 1985.
- [17] M. HEINKENSCHLOSS, *Projected sequential quadratic programming methods*, SIAM J. Optim., 6 (1996), pp. 373–417.
- [18] M. HINTERMÜLLER, *On a globalized augmented Lagrangian-SQP algorithm for nonlinear optimal control problems with box constraints*, in Fast Solution Methods for Discretized Optimization Problems, K. H. Hoffmann, R. H. W. Hoppe, and V. Schulz, eds., Birkhäuser-Verlag, Basel, 2001, pp. 139–153.
- [19] K. ITO AND K. KUNISCH, *Augmented Lagrangian-SQP methods for nonlinear optimal control problems of tracking type*, SIAM J. Control Optim., 34 (1996), pp. 874–891.
- [20] C. T. KELLEY, *Iterative Methods for Linear and Nonlinear Equations*, SIAM, Philadelphia, 1995.
- [21] R. M. LEWIS AND S. G. NASH, *A Multigrid Approach to the Optimization of Systems Governed by Differential Equations*, Tech. Report AIAA-2000-4890, American Institute of Aeronautics and Astronautics, NASA Langley Research Center, 2000.
- [22] R. M. LEWIS AND S. G. NASH, *Model problems for the multigrid optimization of systems governed by differential equations*, submitted for publication.
- [23] S. F. MCCORMICK, ED., *Multigrid Methods*, Frontiers Appl. Math. 3, SIAM, Philadelphia, 1987.
- [24] S. G. NASH, *A multigrid approach to discretized optimization problems*, Optim. Methods Softw., 14 (2000), pp. 99–116.
- [25] R. T. ROCKAFELLAR, *Directional differentiability of the optimal value function in a nonlinear programming problem*, Math. Programming Stud., 21 (1984), pp. 213–226.
- [26] J. SØNDERGAARD, *Optimization Using Surrogate Models—by the Space Mapping Technique*, Ph.D. thesis, Department of Mathematical Modelling, Technical University of Denmark, 2003.
- [27] L. N. VICENTE, *Space mapping: Models, sensitivities, and trust-regions methods*, Optim. Engng., 4 (2003), pp. 159–175.
- [28] P. WESSELING, *An Introduction to Multigrid Methods*, Wiley Interscience, Chichester, UK, 1992.

PERTURBATION ANALYSIS OF ERROR BOUNDS FOR SYSTEMS OF CONIC LINEAR INEQUALITIES IN BANACH SPACES*

XI YIN ZHENG[†] AND KUNG FU NG[‡]

Abstract. For a general (constrained or unconstrained) system of conic inequalities, we provide sufficient conditions to ensure the existence of a common error bound for each of the systems obtained from the original one when the given data undergoes a small perturbation. Moreover, we also give some dual characterizations on the existence of error bounds for systems of conic inequalities.

Key words. abstract conic linear inequality, error bound, stability, least error bound, normal cone

AMS subject classifications. 90C31, 90C05, 49K40, 49J52

DOI. 10.1137/03060179X

1. Introduction. This paper is devoted to studying the effect on error bounds for a system of conic linear inequalities when the given data undergoes perturbations. The system to be considered is one of the following types (listed in order of generality):

(a) A constrained system of a conic linear inequality:

$$(1.1) \quad Tx \leq_C b, \quad x \in \Omega.$$

(b) A system of conic linear inequalities:

$$(1.2) \quad Tx \leq_C b, \quad x \leq_{C_X} a.$$

(c) An unconstrained system:

$$(1.3) \quad Tx \leq_C b.$$

Here and throughout, $T \in L(X, Y)$, namely, T is a bounded linear operator from a Banach space X into another Banach space Y with (partial) orders $\leq_{C_X}$ and $\leq_C$, respectively, defined by closed convex cones $C_X \subset X$ and $C \subset Y$; Ω is a closed convex subset of X , $a \in X$, and $b \in Y$. More details on notation/definitions are given in the next section.

Since the pioneering work of Hoffman [9], many publications have addressed the error bound conditions in a wide range of applications (see [6, 14, 15, 16, 19] and references therein). As far as we know, Luo and Tseng [17] (who consider the classical case $X = R^n$, $Y = R^m$, and $C = R_+^m$) were the first to thoroughly study the effect on error bounds of (1.3) when both T and b are perturbed; they proved the following interesting result.

*Received by the editors November 11, 2003; accepted for publication (in revised form) December 20, 2004; published electronically June 30, 2005. This research was supported by a postdoctoral fellowship scheme (CUHK) and an Earmarked grant from the Research Grant Council of Hong Kong.

<http://www.siam.org/journals/siopt/15-4/60179.html>

[†]Department of Mathematics, Yunnan University, Kunming 650091, P. R. China (xyzheng@ynu.edu.cn). Research of this author was supported by the National Natural Science Foundation of P. R. China (grant 10361008) and the Natural Science Foundation of Yunnan Province, China (grant 2003A0002M).

[‡]Department of Mathematics, The Chinese University of Hong Kong, Shatin, New Territory, Hong Kong (kfng@math.cuhk.edu.hk).

THEOREM LT. Let $X = R^n$, $Y = R^m$, $C = R_+^m$. Then the following conditions are equivalent:

(a) Either $Tx <_C 0$ (i.e., $-Tx \in \text{int}(C)$, the interior of C) is solvable, or $Tx <_C b$ is solvable and the solution set $S(T, b)$ of (1.3) is bounded.

(b) There exist $\tau, \delta > 0$ such that

$$\text{dist}(x, S(\tilde{T}, \tilde{b})) \leq \tau \text{dist}(\tilde{T}x, \tilde{b} - C) \quad \text{for all } x \in X$$

whenever $\tilde{T} \in L(X, Y)$ and $\tilde{b} \in Y$ satisfy $\|\tilde{T} - T\| < \delta$ and $\|\tilde{b} - b\| < \delta$.

Li and Singer [15] considered an infinite-dimensional setting with a constraint and proved that if the solution set $S(T, b)$ of (1.1) is bounded, and if the following constraint qualification of Robinson (cf. [3]) is assumed,

$$(1.4) \quad b \in \text{int}(T(\Omega) + C) \quad (\text{RCQ}),$$

then there exist $\tau, \delta > 0$ such that

$$(1.5) \quad \text{dist}(x, S(\tilde{T}, \tilde{b})) \leq \tau \text{dist}(\tilde{T}x, \tilde{b} - C) \quad \text{for all } x \in \Omega$$

whenever $\tilde{T} \in L(X, Y)$ and $\tilde{b} \in Y$ satisfy $\|\tilde{T} - T\| < \delta$ and $\|\tilde{b} - b\| < \delta$. (Note that, in the case when the ordering cone C is a general closed convex cone, the Robinson constraint qualification (RCQ) is weaker than the assumption that $Tx <_C b$ is solvable.)

Theorem 3.1 shows that (under the assumptions of Li and Singer) the conclusion (1.5) can be strengthened into

$$(1.6) \quad \text{dist}(x, S(\tilde{T}, \tilde{b})) \leq \tau [\text{dist}(x, \Omega) + \text{dist}(\tilde{T}x, \tilde{b} - C)] \quad \text{for all } x \in X$$

whenever $\tilde{T} \in L(X, Y)$ and $\tilde{b} \in Y$ satisfy $\|\tilde{T} - T\| < \delta$ and $\|\tilde{b} - b\| < \delta$. The other sets of sufficient conditions ensuring (1.6) holds are provided in Theorems 3.2 and 4.1; the implication (a) $\implies$ (b) is thereby extended to the case when X and Y are not necessarily finite-dimensional. In particular, in the case when X is finite-dimensional (while Y is infinite-dimensional with a general closed convex cone C), we show that the following conditions are equivalent:

(a*) Either $0 \in \text{int}(T(X) + C)$, or $b \in \text{int}(T(X) + C)$ and the solution set $S(T, b)$ of (1.3) is bounded.

(b) There exist $\tau, \delta > 0$ such that

$$\text{dist}(x, S(\tilde{T}, \tilde{b})) \leq \tau \text{dist}(\tilde{T}x, \tilde{b} - C) \quad \text{for all } x \in X$$

whenever $\tilde{T} \in L(X, Y)$ and $\tilde{b} \in Y$ satisfy $\|\tilde{T} - T\| < \delta$ and $\|\tilde{b} - b\| < \delta$.

Clearly, condition (a*) is in general weaker than condition (a) of Theorem LT, but the two conditions coincide when $\text{int}(C)$ is assumed nonempty. A key step for establishing (1.6) is a result presented in Proposition 2.1, which provides a useful sufficient condition ensuring the existence of error bounds for system (1.1). Using this later result and, in terms of normal cones, we provide in sections 4 and 5 some dual representations of error bounds for systems (1.1) and (1.3).

Our work here was inspired by [12, 13], where RCQ was employed to study the perturbed systems, but only regarding the issue of solvability (rather than that of error bounds) and under the restriction that Ω is a closed convex cone.

2. Preliminary results and error bounds. Throughout this paper, X, Y denote Banach spaces and $L(X, Y)$ the Banach space of all bounded linear operators from X into Y . Let X^* denote the dual of X and let B_X denote the closed unit ball of X . Similar notation will be used for other normed vector spaces. Moreover, for any x in a normed vector space, we use $B(x, r)$ to denote the open ball with center x and radius r . Let $C \subset Y$ be a closed convex cone and let $\leq_C$ denote the partial preorder defined by C : for $y_1, y_2 \in Y$ one defines $y_1 \leq_C y_2$ (resp., $y_1 <_C y_2$) to mean that $y_2 - y_1 \in C$ (resp., $y_2 - y_1 \in \text{int}(C)$). Similarly, when a closed convex cone $C_X \subset X$ is specified, $\leq_{C_X}$ denotes the partial preorder defined by C_X . For subsets $A_1, A_2 \subset X$, let $H(A_1, A_2)$ denote the Hausdorff distance between A_1 and A_2 :

$$H(A_1, A_2) := \max\{\sup\{\text{dist}(a_1, A_2) : a_1 \in A_1\}, \sup\{\text{dist}(a_2, A_1) : a_2 \in A_2\}\}.$$

We will use the following consequence of the RCQ defined in section 1.

LEMMA 2.1. (see [14, Theorem 19]). *Let the solution set $S(T, b)$ of (1.1) be nonempty and bounded. Assume that the RCQ is satisfied. Then there exist $\delta_0, \lambda_0 > 0$ such that*

$$H(S(\tilde{T}, \tilde{b}), S(T, b)) \leq \lambda_0(\|\tilde{T} - T\| + \|\tilde{b} - b\|)$$

whenever $\tilde{T} \in L(X, Y)$ and $\tilde{b} \in Y$ satisfy $\|\tilde{T} - T\| < \delta_0$ and $\|\tilde{b} - b\| < \delta_0$.

The following two lemmas are from functional analysis; the second is known as the Radstrom cancellation lemma (cf. [20]), while the first is an immediate consequence of Jameson's result on CS-closed and CS-compact sets (for the definition of these notions see [10, p. 183]. In addition, the reader might wish to refer to [23, Cor. 1.3.4]).

LEMMA 2.2. *Let A be a bounded closed convex subset of X and K be a closed convex subset of Y . Let $T \in L(X, Y)$, $b \in Y$, and $r > 0$. Then*

$$b + rB_Y \subset \overline{T(A) + K} \implies b + \text{rint}(B_Y) \subset T(A) + K,$$

where B_Y denotes the closed unit ball of Y .

Proof. By completeness of X , it is easy to verify that A is CS-compact, and so must be its image $T(A)$ under T . Since K , being a closed convex subset of a Banach space, is CS-closed, it follows that $T(A) + K$ is CS-closed. The result follows from [10, Thms. A.1 and A.2, pp. 183–184]. $\square$

LEMMA 2.3. *Let E and K be subsets of Y and suppose that K is convex. Let $r > 0$ be that*

$$(2.1) \quad E + rB_Y \subset K + rB_Y.$$

Then $E \subset \overline{K}$.

Let $\phi : X \rightarrow R \cup \{+\infty\}$ be a proper lower semicontinuous function and ϕ° denote the generalized directional derivative of ϕ , that is,

$$\phi^\circ(x; v) := \lim_{\varepsilon \downarrow 0} \limsup_{\substack{y \downarrow_\phi x \\ t \downarrow 0}} \inf_{w \in v + \varepsilon B_X} \frac{\phi(y + tw) - \phi(y)}{t}$$

for $x \in \text{dom}(\phi) := \{x \in X : \phi(x) < +\infty\}$ and $v \in X$, where $y \downarrow_\phi x$ means that $y \rightarrow x$ and $\phi(y) \rightarrow \phi(x)$. Recall that the Clarke subdifferential of ϕ at $x \in \text{dom}(\phi)$ can be defined by

$$\partial\phi(x) := \{x^* \in X^* : \langle x^*, v \rangle \leq \phi^\circ(x; v) \text{ for all } v \in X\}.$$

For a closed subset A of X , let $N(A, x)$ denote the normal cone of A at $x \in X$, that is, $N(A, x) := \partial\delta_A(x)$ if $x \in A$ and $N(A, x) := \emptyset$ otherwise, where δ_A is the indicator function of A . It is well known that if ϕ (resp., A) is convex, then $\partial\phi(x)$ (resp., $N(A, x)$) is the subdifferential (resp., normal cone) in the sense of convex analysis (cf. [7]). The next result is on the existence of a local error bound. One of the referees kindly informed us that a similar result was obtained by Azé and Corvellec [2]. Some authors proved corresponding results on global error bounds (cf. [1, 11] and references therein).

LEMMA 2.4. *Let $\phi : X \rightarrow R \cup \{+\infty\}$ be a proper lower semicontinuous function and $S := \{x \in X : \phi(x) \leq 0\}$. Let $\tau, \varepsilon \in (0, +\infty)$ and $x_0 \in S$ such that*

$$(2.2) \quad \text{dist}(0, \partial\phi(x)) \geq \frac{1}{\tau} \quad \text{for all } x \in B(x_0, 2\varepsilon) \setminus S.$$

Then

$$\text{dist}(x, S) \leq \tau[\phi(x)]_+ \quad \text{for all } x \in B(x_0, \varepsilon),$$

where $[\phi(x)]_+ = \max\{0, \phi(x)\}$.

Proof. Suppose to the contrary that there exist $z_0 \in B(x_0, \varepsilon) \setminus S$ and $\sigma \in (0, 1)$ such that $\text{dist}(z_0, S) > (\tau + \sigma)\phi(z_0)$. Let $\psi(x) := [\phi(x)]_+$ for all $x \in X$. Then $\inf\{\psi(x) : x \in X\} = 0$ and $\psi(z_0) = \phi(z_0)$. By the Ekeland variational principle there exists $z' \in X$ such that

$$(2.3) \quad \|z' - z_0\| < \text{dist}(z_0, S) \leq \|z_0 - x_0\|$$

and

$$\psi(z') \leq \psi(z) + \frac{1}{\tau + \sigma} \|z - z'\| \quad \text{for all } z \in X.$$

It follows that $0 \in \partial\psi(z') + \frac{1}{\tau + \sigma} B_{X^*}$. This shows that

$$(2.4) \quad \text{dist}(0, \partial\psi(z')) < \frac{1}{\tau}.$$

On the other hand, by (2.3) one has that $z' \notin S$. Hence $X \setminus S$ is an open neighborhood of z' because ϕ is lower semicontinuous. Since, by definition, ψ and ϕ are identical on $X \setminus S$, it follows that $\partial\phi(z') = \partial\psi(z')$. Thus (2.4) is rewritten as $\text{dist}(0, \partial\phi(z')) < \frac{1}{\tau}$, contradicting (2.2) (because $\|z' - x_0\| \leq \|z' - z_0\| + \|z_0 - x_0\| \leq 2\|z_0 - x_0\| < 2\varepsilon$). $\square$

The following lemma is a useful tool for the proofs of our main results.

LEMMA 2.5. *Consider (1.1) and let $x \in S(T, b)$ ($= \Omega \cap T^{-1}(b - C)$). Suppose there exists $r > 0$ such that*

$$(2.5) \quad b + rB_Y \subset T((x + B_X) \cap \Omega) + C.$$

Then, for any $\varepsilon \in (0, r/4(\|T\| + 1)^2)$,

$$\text{dist}(z, S(T, b)) \leq \tau_\varepsilon(\text{dist}(z, \Omega) + \text{dist}(Tz, b - C)) \quad \text{for all } z \in B(x, \varepsilon),$$

where $\tau_\varepsilon := \max\{2, \frac{1}{r-2(1+\|T\|)\varepsilon}, \frac{2\|T\|}{r-4(\|T\|^2+\|T\|)\varepsilon}\}$.

If, in addition, $\Omega = X$, then

$$\text{dist}(z, S(T, b)) \leq \frac{1}{r - 2(1 + \|T\|)\varepsilon} \text{dist}(Tz, b - C) \quad \text{for all } z \in B(x, \varepsilon).$$

Proof. Let $\phi(z) := \text{dist}(z, \Omega) + \text{dist}(Tz, b - C)$ for all $z \in X$. Then ϕ is a continuous convex function on X . By Lemma 2.4 we need only show that

$$(2.6) \quad \text{dist}(0, \partial\phi(u)) \geq \frac{1}{\tau_\varepsilon} \quad \text{for all } u \in B(x, 2\varepsilon) \setminus S(T, b)$$

and that

$$(2.7) \quad \text{dist}(0, \partial\phi(u)) \geq r - 2(1 + \|T\|)\varepsilon \quad \text{for all } u \in B(x, 2\varepsilon) \setminus T^{-1}(b - C).$$

Let $u \in B(x, 2\varepsilon)$ and $u^* \in \partial\phi(u)$. By the Moreau–Rockafellar theorem, there exist $x^* \in \partial\text{dist}(\cdot, \Omega)(u)$ and $x_1^* \in \partial\text{dist}(T(\cdot), b - C)(u)$ such that $u^* = x^* + x_1^*$. Noting that

$$\partial\text{dist}(T(\cdot), b - C)(u) = T^*(\partial\text{dist}(\cdot, b - C)(Tu))$$

(cf. [7, Thm. 2.3.10]), there exists $y^* \in \partial\text{dist}(\cdot, b - C)(Tu)$ such that $u^* = x^* + T^*y^*$. Recalling (cf. [5, Thm. 1]) that

$$(2.8) \quad \partial\text{dist}(\cdot, \Omega)(v) = \begin{cases} \text{bd}(B_{X^*}) \cap N(\Omega + r_v B_X, v), & v \notin \Omega, \\ B_{X^*} \cap N(\Omega, v), & v \in \Omega \end{cases}$$

and

$$(2.9) \quad \partial\text{dist}(\cdot, b - C)(Tv) = \begin{cases} \text{bd}(B_{Y^*}) \cap N(b - C + r_{Tv} B_Y, Tv), & v \notin T^{-1}(b - C), \\ B_{Y^*} \cap N(b - C, Tv), & v \in T^{-1}(b - C), \end{cases}$$

where $r_v := \text{dist}(v, \Omega)$ and $r_{Tv} := \text{dist}(Tv, b - C)$, one has

$$(2.10) \quad \langle x^*, u - \omega \rangle \geq 0 \quad \text{for all } \omega \in \Omega \quad \text{and} \quad \langle y^*, Tu - b + c \rangle \geq 0 \quad \text{for all } c \in C.$$

Take a sequence $\{y_n\}$ in B_Y satisfying

$$(2.11) \quad \langle y^*, y_n \rangle \geq \left(1 - \frac{1}{n}\right) \|y^*\|.$$

By (2.5) there exist $\omega_n \in (x + B_X) \cap \Omega$ and $c_n \in C$ such that $T\omega_n + c_n = b - ry_n$. Therefore,

$$\begin{aligned} \|u^*\| &= \|x^* + T^*y^*\| \geq \langle x^* + T^*y^*, x - \omega_n \rangle \\ &= \langle x^*, x - \omega_n \rangle + \langle y^*, Tx - T\omega_n \rangle \\ &= \langle x^*, x - \omega_n \rangle + \langle y^*, Tx - b + c_n + ry_n \rangle. \end{aligned}$$

It follows from (2.10) and (2.11) that

$$\begin{aligned} \|u^*\| &\geq \langle x^*, x - u \rangle + \langle y^*, T(x - u) \rangle + r \left(1 - \frac{1}{n}\right) \|y^*\| \\ &\geq -\|x^*\| \|x - u\| - \|y^*\| \|T\| \|x - u\| + r \left(1 - \frac{1}{n}\right) \|y^*\| \\ &\geq -2(1 + \|T\|)\varepsilon + r \left(1 - \frac{1}{n}\right) \|y^*\|. \end{aligned}$$

Letting $n \rightarrow \infty$, one has

$$(2.12) \quad \|u^*\| \geq -2(1 + \|T\|)\varepsilon + r\|y^*\|.$$

If $u \notin T^{-1}(b - C)$, then $\|y^*\| = 1$, and this implies that (2.7) holds. To verify (2.6) it remains to consider the case when $u \in T^{-1}(b - C) \setminus \Omega$. To do this, note first that if $\|y^*\| \geq \frac{1}{2\|T\|}$, then (2.12) implies that

$$\|u^*\| \geq \frac{r}{2\|T\|} - 2(1 + \|T\|)\varepsilon \geq \frac{1}{\tau_\varepsilon}.$$

Next if $\|y^*\| < \frac{1}{2\|T\|}$, then

$$\|u^*\| = \|x^* + T^*y^*\| \geq \|x^*\| - \|T\|\|y^*\| \geq 1 - \frac{1}{2} = \frac{1}{2},$$

and so (2.6) holds in any case. $\square$

We end the section with the following proposition that gives a useful sufficient condition ensuring the existence of global error bounds for (1.1).

PROPOSITION 2.1. *Let $r > 0$ and suppose that*

$$b + rB_Y \subset T((u + B_X) \cap \Omega) + C \quad \text{for all } u \in \text{bd}(S(T, b)).$$

Then

$$\text{dist}(x, S(T, b)) \leq \max \left\{ 2, \frac{1}{r}, \frac{2\|T\|}{r} \right\} (\text{dist}(x, \Omega) + \text{dist}(Tx, b - C)) \quad \text{for all } x \in X.$$

If, in addition, $\Omega = X$, then

$$\text{dist}(x, S(T, b)) \leq \frac{1}{r} \text{dist}(Tx, b - C) \quad \text{for all } x \in X.$$

Proof. Let $\varepsilon \in (0, r/4(\|T\| + 1)^2)$, and let τ_ε be defined by

$$\tau_\varepsilon = \begin{cases} \max \left\{ 2, \frac{1}{r-2(1+\|T\|)\varepsilon}, \frac{2\|T\|}{r-4(\|T\|^2+\|T\|)\varepsilon} \right\}, & \Omega \neq X, \\ \frac{1}{r-2(1+\|T\|)\varepsilon}, & \Omega = X. \end{cases}$$

Let $x \in X \setminus S(T, b)$ with $\text{dist}(x, S(T, b)) < \varepsilon$. Then there exists $u \in \text{bd}(S(T, b))$ such that $x \in B(u, \varepsilon)$. By the given assumption and Lemma 2.5, one has that

$$\text{dist}(x, S(T, b)) \leq \tau_\varepsilon (\text{dist}(x, \Omega) + \text{dist}(Tx, b - C)).$$

This and [15, Thm. 4] imply that

$$\text{dist}(x, S(T, b)) \leq \tau_\varepsilon (\text{dist}(x, \Omega) + \text{dist}(Tx, b - C)) \quad \text{for all } x \in X.$$

Letting $\varepsilon \rightarrow 0^+$, we prove the result. $\square$

3. Stable error bounds. The notion of stable error bounds is defined in section 1. Recall also that the RCQ is defined by (1.4). The first two theorems in this section extend (a) $\implies$ (b) of Theorem LT to the situation when X, Y are infinite-dimensional with a constraint set Ω that is a general closed convex subset of X . The first result below deals with the case when the solution set of (1.1) is bounded.

THEOREM 3.1. *Assume that the solution set $S(T, b)$ of system (1.1) is bounded. Then system (1.1) satisfies the RCQ if and only if there exist $\delta, \tau \in (0, +\infty)$ such that (1.6) holds whenever $\tilde{T} \in L(X, Y)$ and $\tilde{b} \in Y$ satisfy $\|\tilde{T} - T\| < \delta$ and $\|\tilde{b} - b\| < \delta$.*

Proof. It is straightforward to verify the sufficiency. Conversely, suppose that $b \in \text{int}(T(\Omega) + C)$. Then, by Lemma 2.1 there exist $\delta_0, \lambda_0 > 0$ such that

$$(3.1) \quad H(S(\tilde{T}, \tilde{b}), S(T, b)) \leq \lambda_0(\|\tilde{T} - T\| + \|\tilde{b} - b\|)$$

whenever $\tilde{T} \in L(X, Y)$ and $\tilde{b} \in Y$ satisfy $\|\tilde{T} - T\| < \delta_0$ and $\|\tilde{b} - b\| < \delta_0$. Let $Fx := Tx + C$ if $x \in \Omega$ and $Fx = \emptyset$ otherwise. It is clear that F is a closed convex set-valued mapping from X to Y with $F(X) = T(\Omega) + C$ and $S(T, b) = F^{-1}b$. Let $x_0 \in S(T, b)$. Making use of the assumption $b \in \text{int}(T(\Omega) + C)$, we see that the Robinson–Ursescu theorem (cf. [21, Thm. 1] and [15, Thm. 1]) implies that there exists $r > 0$ such that $b + 4rB_Y \subset F(x_0 + \frac{1}{2}B_X)$, that is,

$$(3.2) \quad b + 4rB_Y \subset T\left(\left(x_0 + \frac{1}{2}B_X\right) \cap \Omega\right) + C.$$

On the other hand, by (3.1) there exists $\tilde{x}_0 \in S(\tilde{T}, \tilde{b})$ such that

$$(3.3) \quad \|\tilde{x}_0 - x_0\| \leq 2\lambda_0(\|\tilde{T} - T\| + \|\tilde{b} - b\|)$$

whenever $\tilde{T} \in L(X, Y)$ and $\tilde{b} \in Y$ satisfy $\|\tilde{T} - T\| < \delta_0$ and $\|\tilde{b} - b\| < \delta_0$. Let $M := \sup\{\|x\| : x \in S(T, b)\}$, and take $\delta \in (0, \delta_0)$ such that

$$(3.4) \quad \delta(M + 1) + 4\lambda_0\delta < \min\left\{r, \frac{1}{2}\right\}.$$

Let $\tilde{T} \in L(X, Y)$ and $\tilde{b} \in Y$ satisfy $\|\tilde{T} - T\| < \delta$ and $\|\tilde{b} - b\| < \delta$. We claim that

$$(3.5) \quad \tilde{b} + rB_Y \subset \tilde{T}((\tilde{x}_0 + B_X) \cap \Omega) + C.$$

To see this, note that

$$\begin{aligned} T\left(\left(x_0 + \frac{1}{2}B_X\right) \cap \Omega\right) &\subset \tilde{T}\left(\left(x_0 + \frac{1}{2}B_X\right) \cap \Omega\right) + \|T - \tilde{T}\|\left(\|x_0\| + \frac{1}{2}\right)B_Y \\ &\subset \tilde{T}\left(\left(x_0 + \frac{1}{2}B_X\right) \cap \Omega\right) + \delta(M + 1)B_Y. \end{aligned}$$

Hence, by (3.2) one has

$$\begin{aligned} b + 4rB_Y &\subset \tilde{T}\left(\left(x_0 + \frac{1}{2}B_X\right) \cap \Omega\right) + C + \delta(M + 1)B_Y \\ &\subset \tilde{T}\left(\left(\tilde{x}_0 + B_X\right) \cap \Omega\right) + C + \delta(M + 1)B_Y, \end{aligned}$$

where the last inclusion holds because $x_0 + \frac{1}{2}B_X \subset \tilde{x}_0 + B_X$ holds by (3.3) and (3.4). Consequently, by (3.4) and $\|\tilde{b} - b\| < \delta$, one has

$$\tilde{b} + 3rB_Y \subset \tilde{T}((\tilde{x}_0 + B_X) \cap \Omega) + C + rB_Y.$$

It follows from Lemmas 2.2 and 2.3 that (3.5) holds. Now let u be an arbitrary point in $\text{bd}(S(\tilde{T}, \tilde{b}))$. By (3.1) and the definition of M , one has

$$\sup\{\|x\| : x \in S(\tilde{T}, \tilde{b})\} \leq M + 2\lambda_0\delta.$$

Hence

$$\tilde{x}_0 + B_X \subset u + (4\lambda_0\delta + 2M + 1)B_X.$$

Letting $\alpha := \frac{1}{4\lambda_0\delta + 2M + 1}$ and noting that $u \in \Omega$, one has

$$(1 - \alpha)u + \alpha[(\tilde{x}_0 + B_X) \cap \Omega] \subset (u + B_X) \cap \Omega.$$

Since C is a convex cone and $\tilde{b} \in \tilde{T}(u) + C$, it follows from (3.5) and linearity of $\tilde{T}$ that

$$\tilde{b} + \alpha rB_Y \subset \tilde{T}((u + B_X) \cap \Omega) + C.$$

Thus, noting that $\|\tilde{T}\| < \|T\| + \delta$, Proposition 2.1 entails that

$$\begin{aligned} \text{dist}(z, S(\tilde{T}, \tilde{b})) &\leq \max\left\{2, \frac{1}{\alpha r}, \frac{2\|\tilde{T}\|}{\alpha r}\right\} (\text{dist}(z, \Omega) + \text{dist}(\tilde{T}(z), \tilde{b} - C)) \quad \text{for all } z \in X \\ &\leq \tau(\text{dist}(z, \Omega) + \text{dist}(\tilde{T}(z), \tilde{b} - C)) \quad \text{for all } z \in X, \end{aligned}$$

where $\tau := \max\{2, \frac{1}{\alpha r}, \frac{2(\|T\| + \delta)}{\alpha r}\}$. Note that τ is independent on $\tilde{T}$ and $\tilde{b}$. This completes the proof of the necessity. $\square$

The following example shows that Theorem 3.1 is not necessarily true if the solution set is unbounded.

Example 3.1. Let $X = R$, $Y = R^2$ (with the Euclidean norm), $\Omega = X$, $C = R \times R_+$, $b = (0, 1)$, and let $T : X \rightarrow Y$ be defined by $Tx = 0$ for all $x \in X$. It is clear that $b \in \text{int}(Tx + C)$ for each $x \in X$, and hence the RCQ is satisfied. For every natural number n , define $T_n : X \rightarrow Y$ by $T_n x = (0, \frac{x}{n})$ for each $x \in X$. Then $\|T_n - T\| = \frac{1}{n} \rightarrow 0$ and $S(T_n, b) = (-\infty, n]$. Therefore,

$$\text{dist}(x, S(T_n, b)) = x - n = n \text{dist}(b, T_n x + C) \quad \text{for all } x > n.$$

This shows that the system $(Tx \leq_C b, x \in \Omega)$ has no stable error bounds.

Our next theorem supplements Theorem 3.1 by considering the case when the solution set of (1.1) is not necessarily bounded. Let $\text{Rec}(\Omega)$ denote the recession cone of Ω , that is,

$$\text{Rec}(\Omega) := \{h \in X : \exists \omega_n \in \Omega \text{ and } \exists t_n > 0 \text{ such that } t_n \rightarrow 0 \text{ and } t_n \omega_n \rightarrow h\}.$$

It is known (cf. [22]) that $\Omega = \Omega + \text{Rec}(\Omega)$ and

$$\begin{aligned} \text{Rec}(\Omega) &= \{h \in X : x + R_+ h \subset \Omega \text{ for some } x \in \Omega\} \\ &= \{h \in X : x + R_+ h \subset \Omega \text{ for all } x \in \Omega\}. \end{aligned}$$

In the case when X is finite-dimensional, Ω is bounded if and only if $\text{Rec}(\Omega) = \{0\}$. The next result is stronger than the result that system (1.1) has a stable error bound for each $b \in Y$ because (3.6) refers to any $\tilde{b}$, and also because τ, δ do not depend on b .

THEOREM 3.2. *Suppose that $0 \in \text{int}(T(\text{Rec}(\Omega)) + C)$ (equivalently, $T(\text{Rec}(\Omega)) + C = Y$). Then there exist $\tau, \delta \in (0, +\infty)$ depending only on $T, \text{Rec}(\Omega)$, and C such that*

$$(3.6) \quad \text{dist}(x, S(\tilde{T}, \tilde{b})) \leq \tau[\text{dist}(x, \Omega) + \text{dist}(\tilde{T}x, \tilde{b} - C)] \quad \text{for all } x \in X$$

whenever $\tilde{T} \in L(X, Y)$ with $\|\tilde{T} - T\| < \delta$ and $\tilde{b} \in Y$.

Proof. By the given assumption and the Robinson–Ursescu theorem,

$$0 \in \text{int}(T(B_X \cap \text{Rec}(\Omega)) + C).$$

Let

$$\delta := \sup\{r > 0 : \text{rint}(B_Y) \subset T(B_X \cap \text{Rec}(\Omega)) + C\}.$$

Thus, $\delta > 0$ and $\delta \text{int}(B_Y) \subset T(B_X \cap \text{Rec}(\Omega)) + C$. Let $\tilde{T} \in L(X, Y)$ with $\|\tilde{T} - T\| < \delta$. Then

$$\delta \text{int}(B_Y) \subset \tilde{T}(B_X \cap \text{Rec}(\Omega)) + \|\tilde{T} - T\|B_Y + C.$$

It follows from Lemmas 2.2 and 2.3 that

$$(3.7) \quad (\delta - \|\tilde{T} - T\|)\text{int}(B_Y) \subset \tilde{T}(B_X \cap \text{Rec}(\Omega)) + C.$$

Let $\tilde{b} \in Y$ and $u \in \text{bd}(S(\tilde{T}, \tilde{b}))$. Then

$$u + B_X \cap \text{Rec}(\Omega) = (u + B_X) \cap (u + \text{Rec}(\Omega)) \subset (u + B_X) \cap \Omega.$$

By (3.7) and $\tilde{b} \in \tilde{T}u + C$, it follows that

$$\tilde{b} + (\delta - \|\tilde{T} - T\|)\text{int}(B_Y) \subset \tilde{T}((u + B_X) \cap \Omega) + C.$$

Letting

$$\tilde{\tau} := \max \left\{ 2, \frac{1}{\delta - \|\tilde{T} - T\|}, \frac{2\|\tilde{T}\|}{\delta - \|\tilde{T} - T\|} \right\},$$

it follows from Proposition 2.1 that

$$\text{dist}(z, S(\tilde{T}, \tilde{b})) \leq \tilde{\tau}(\text{dist}(z, \Omega) + \text{dist}(\tilde{T}z, \tilde{b} - C)) \quad \text{for all } z \in X.$$

Let $\tau := \max\{2, \frac{2}{\delta}, \frac{4(\|\tilde{T}\| + \delta)}{\delta}\}$. Then, for any $\tilde{T} \in L(X, Y)$ with $\|\tilde{T} - T\| < \frac{\delta}{2}$, we have $\tilde{\tau} \leq \tau$, and (3.6) holds. The proof is completed. $\square$

The remainder of this section is devoted to studying system (1.2) for any $(a, b) \in X \times Y$. The solution set of (1.2) is denoted by $S(T, a, b)$; thus $S(T, a, b) = (a - C_X) \cap T^{-1}(b - C)$.

THEOREM 3.3. *For any bounded linear operator $T \in L(X, Y)$ and any closed convex cones $C_X \subset X$ and $C \subset Y$, the following conditions are equivalent:*

- (a) $0 \in \text{int}(T(-C_X) + C)$ (equivalently, $Y = T(-C_X) + C$).

(b) *There exist $\tau, \delta \in (0, +\infty)$ depending only on T, C_X , and C such that*

$$\text{dist}(x, S(\tilde{T}, \tilde{a}, \tilde{b})) \leq \tau[\text{dist}(x, \tilde{a} - C_X) + \text{dist}(\tilde{T}x, \tilde{b} - C)] \quad \text{for all } x \in X$$

whenever $(\tilde{a}, \tilde{b}) \in X \times Y$ and $\tilde{T} \in L(X, Y)$ with $\|\tilde{T} - T\| < \delta$.

Proof. (a) $\implies$ (b) Let $\tilde{a} \in X$ and $\Omega := \tilde{a} - C_X$. Then $\text{Rec}(\Omega) = -C_X$, which is independent on $\tilde{a}$. Thus, (a) means that $0 \in \text{int}(T(\text{Rec}(\Omega)) + C)$. Therefore, (b) follows from Theorem 3.2.

(b) $\implies$ (a) Let b be an arbitrary point of Y . Then

$$\text{dist}(x, S(T, 0, b)) \leq \tau[\text{dist}(x, -C_X) + \text{dist}(Tx, b - C)] \quad \text{for all } x \in X.$$

Therefore, $S(T, 0, b) \neq \emptyset$; that is, there exists $u \in -C_X$ such that $Tu \in b - C$. This shows that $b \in T(-C_X) + C$. $\square$

In the special case when $X = R^m$, $Y = R^n$, and $C = R_+^n$, Theorem 3.3 is due to Luo and Tseng (cf. [17, Thm. 3.1]).

4. Unconstrained systems of conic linear inequalities. In this section, we consider error bounds of (1.3), an unconstrained abstract linear inequality. We use $S(T, b)$ to denote the solution set of (1.3), namely, $S(T, b) = T^{-1}(b - C)$. We say that (1.3) satisfies the Slater condition if

$$(4.1) \quad 0 \in \text{int}(T(X) + C) \quad (\text{equivalently, } T(X) + C = Y).$$

Note that condition (4.1) holds if $Tx <_C 0$ is solvable; moreover, in the case when $\text{int}(C) \neq \emptyset$, (4.1) holds if and only if $Tx <_C 0$ is solvable. Let

$$\Delta := \{\delta > 0 : B_Y \subset T(\delta B_X) + C\}.$$

In view of the Robinson–Ursescu theorem, Δ is nonempty if and only if (4.1) holds. For convenience, let

$$\tau(T, C) := \inf\{\tau > 0 : \text{dist}(x, T^{-1}(b - C)) \leq \tau \text{dist}(Tx, b - C) \quad \text{for all } (x, b) \in X \times Y\}.$$

THEOREM 4.1. *Suppose that (1.3) satisfies the Slater condition (4.1). Then $\tau(T, C) = \inf\{\delta : \delta \in \Delta\} < \infty$. In particular, (1.3) has a common error bound for each $b \in Y$.*

Proof. Let $\delta(T, C) := \inf\{\delta : \delta \in \Delta\}$. By assumption, Δ is nonempty, and so $\delta(T, C)$ is finite. Let $\delta > \delta(T, C)$. Then $B_Y \subset T(\delta B_X) + C$, that is,

$$\frac{1}{\delta} B_Y \subset T(B_X) + C.$$

Let $b \in Y$. Since C is a convex cone, it follows that

$$b + \frac{1}{\delta} B_Y \subset T(u + B_X) + C \quad \text{for all } u \in \text{bd}(S(T, b)).$$

This and Proposition 2.1 imply that

$$\text{dist}(x, S(T, b)) \leq \delta \text{dist}(Tx, b - C) \quad \text{for all } x \in X.$$

This shows that $\tau(T, C) \leq \delta(T, C)$. On the other hand, let $y \in B_Y$. By the definition of $\tau(T, C)$,

$$\text{dist}(0, S(T, y)) \leq \tau(T, C) \text{dist}(T0, y - C) \leq \tau(T, C).$$

Thus, for any $\varepsilon > 0$ there exists $a \in S(T, y) = T^{-1}(y - C)$ such that $\|a\| < \tau(T, C) + \varepsilon$. This shows that

$$B_Y \subset T((\tau(T, C) + \varepsilon)B_X) + C.$$

Since $\varepsilon > 0$ is arbitrary, it follows from the definition of $\delta(T, C)$ that $\delta(T, C) \leq \tau(T, C)$. Thus, $\tau(T, C) = \delta(T, C)$. $\square$

The following theorem provides other representations of $\tau(T, C)$.

THEOREM 4.2. *Suppose that (1.3) satisfies the Slater condition (4.1). Then*

$$(4.2) \quad \tau(T, C) = (\inf\{\|T^*y^*\| : y^* \in C^+ \text{ with } \|y^*\| = 1\})^{-1}$$

$$(4.3) \quad = \sup\{\|y^*\| : y^* \in C^+ \text{ with } \|T^*y^*\| = 1\}.$$

Proof. It is straightforward to verify that the numbers represented in the right-hand sides of (4.2) and (4.3) are equal. They are henceforth denoted by $\tilde{\tau}$. Let $\delta(T, C)$ be defined as in the proof of Theorem 4.1 and let $\delta > \delta(T, C)$. Then one has that $B_Y \subset T(\delta B_X) + C$. Thus, for each $y^* \in C^+$ with $\|y^*\| = 1$,

$$\begin{aligned} 1 &\leq \sup\{\langle -y^*, y \rangle : y \in T(\delta B_X) + C\} \\ &= \sup\{\langle -y^*, y \rangle : y \in T(\delta B_X)\} \\ &= \delta \sup\{\langle -T^*y^*, x \rangle : x \in B_X\} = \delta \|T^*y^*\|, \end{aligned}$$

showing $\tilde{\tau} \leq \delta$. This and the arbitrariness of δ in $(\delta(T, C), +\infty)$ imply that $\tilde{\tau} \leq \delta(T, C)$. To prove $\tilde{\tau} = \delta(T, C)$, suppose to the contrary that $\tilde{\tau} < \delta(T, C)$. Let $\tau' \in (\tilde{\tau}, \delta(T, C))$. Then, by the definition of $\delta(T, C)$, B_Y is not a subset of $T(\tau' B_X) + C$; pick $b \in B_Y$ such that

$$b \notin T(\tau' B_X) + C.$$

By (4.1), one has that $\text{int}(T(\tau' B_X) + C) \neq \emptyset$. Hence, one can apply the separation theorem to obtain $y^* \in Y^*$ with $\|y^*\| = 1$ such that

$$1 \geq \langle y^*, b \rangle \geq \sup\{\langle y^*, y \rangle : y \in T(\tau' B_X) + C\}.$$

Since C is a cone, it follows that $-y^* \in C^+$, and hence

$$1 \geq \sup\{\langle y^*, y \rangle : y \in T(\tau' B_X)\} = \tau' \|T^*(-y^*)\|,$$

contradicting the definition of $\tilde{\tau}$ and the choice of τ' . Therefore we must have $\tilde{\tau} = \delta(T, C)$. The proof is now completed by Theorem 4.1. $\square$

Note. Theorem 4.2 can be proved via Borwein's duality theorem on a closed convex process (cf. [4, Thm. 2.1]).

COROLLARY 4.1. *Let $\sigma(T, C) := \{y^* \in C^+ : \|T^*y^*\| \leq 1\}$ and suppose that (1.3) satisfies the Slater condition (4.1). Then*

$$(4.4) \quad \tau(T, C) = \sup\{\|y^*\| : y^* \text{ is an extreme point of } \sigma(T, C)\}.$$

Proof. Since T^* is weak*-weak* continuous, $\sigma(T, C) = C^+ \cap (T^*)^{-1}(B_{X^*})$ is a weak* closed convex subset of Y^* . Moreover, $\sigma(T, C)$ is bounded by Theorems 4.1 and 4.2. Therefore, $\sigma(T, C)$ is a weak* compact convex set. By the Krein–Milman theorem, $\sigma(T, C)$ is the weak* closed convex hull of the set of its extreme points. Since the dual norm of Y^* is convex and weak* lower semicontinuous, it follows that

$$\sup\{\|y^*\| : y^* \text{ is an extreme point of } \sigma(T, C)\} = \sup\{\|y^*\| : y^* \in \sigma(T, C)\}.$$

Thus (4.4) holds by (4.3). $\square$

Güler, Hoffman, and Rothblum [8] proved that

$$\tau(T, C) \leq \sup\{\|y^*\| : y^* \text{ is an extreme point of } \sigma(T, C)\}$$

in the case when $X = R^m$, $Y = R^n$, and $C = R_+^n$, but without the Slater condition assumption. Representations of $\tau(T, C)$ in the case when $Y = R^n$ and $C = R_+^n$ also have been given recently by Zălinescu [23].

Using Theorem 4.1, we now prove the following result on the existence of stable error bounds of (1.3).

THEOREM 4.3. *Suppose that (1.3) satisfies the Slater condition (4.1). Then*

$$(4.5) \quad \text{dist}(x, S(\tilde{T}, b)) \leq \frac{1}{\frac{1}{\tau(T, C)} - \|\tilde{T} - T\|} \text{dist}(\tilde{T}x, b - C) \quad \text{for all } (x, b) \in X \times Y$$

whenever $\tilde{T} \in L(X, Y)$ with $\|\tilde{T} - T\| < \frac{1}{\tau(T, C)}$.

Proof. Let $\tilde{T} \in L(X, Y)$ with $\|\tilde{T} - T\| < \frac{1}{\tau(T, C)}$ and let γ be an arbitrary number in $(\|\tilde{T} - T\|, \frac{1}{\tau(T, C)})$. It follows from Theorem 4.1 that

$$\gamma B_Y \subset T(B_X) + C \subset \tilde{T}(B_X) + C + \|\tilde{T} - T\| B_Y,$$

where the second inclusion holds because $T(B_X) \subset \tilde{T}(B_X) + \|\tilde{T} - T\| B_Y$. By Lemmas 2.2 and 2.3, one then has

$$(\gamma - \|\tilde{T} - T\|) \text{int}(B_Y) \subset \tilde{T}(B_X) + C,$$

that is,

$$\text{int}(B_Y) \subset \tilde{T} \left(\frac{1}{\gamma - \|\tilde{T} - T\|} B_X \right) + C.$$

It is now easy to verify from Theorem 4.1 that

$$\text{dist}(x, S(\tilde{T}, b)) \leq \frac{1}{\gamma - \|\tilde{T} - T\|} \text{dist}(\tilde{T}(x), b - C) \quad \text{for all } (x, b) \in X \times Y.$$

Letting $\gamma \rightarrow \tau(T, C)^{-1}$, one has that (4.7) holds. $\square$

PROPOSITION 4.1. *Suppose that (1.3) does not satisfy the Slater condition (4.1) and that there exist $\tau, \delta > 0$ such that*

$$(4.6) \quad \text{dist}(x, S(\tilde{T}, \tilde{b})) \leq \tau \text{dist}(\tilde{T}x, \tilde{b} - C) \quad \text{for all } x \in X$$

whenever $\tilde{T} \in L(X, Y)$ and $\tilde{b} \in Y$ satisfy $\|\tilde{T} - T\| < \delta$ and $\|\tilde{b} - b\| < \delta$. Then $\text{Rec}(S(T, b)) = \{0\}$.

Proof. Suppose to the contrary that there exists $z \in \text{Rec}(S(T, b))$ with $\|z\| = 1$. Take $z^* \in X^*$ such that

$$(4.7) \quad \langle z^*, z \rangle = \|z^*\| = \|z\| = 1.$$

Take $x_0 \in S(T, b)$. Then $T(x_0 + rz) \leq_C b$ for all $r > 0$. It follows that $Tz \leq_C 0$. By the given assumption, one has that $0 \notin \text{int}(T(X) + C)$ and $b \in \text{int}(T(X) + C)$. This

and the separation theorem imply that there exists $c^* \in Y^*$ with $\|c^*\| = 1$ such that $\inf\{\langle c^*, y \rangle : y \in T(X) + C\} \geq 0$ and $\langle c^*, b \rangle > 0$. It follows that

$$(4.8) \quad c^* \in C^+ \quad \text{and} \quad T^*c^* = 0.$$

Let $e := b - T(x_0)$, and for any $\varepsilon > 0$, define

$$T_\varepsilon x := Tx + \varepsilon \langle z^*, x \rangle e \quad \text{for all } x \in X.$$

Then, by (4.8), $\langle c^*, e \rangle = \langle c^*, b \rangle > 0$. Let $x_\varepsilon := x_0 + r_\varepsilon z$, where $r_\varepsilon = \frac{1}{\varepsilon} - \langle z^*, x_0 \rangle$. Thus,

$$T_\varepsilon(x_\varepsilon) = T(x_\varepsilon) + \varepsilon \langle z^*, x_\varepsilon \rangle e = Tx_0 + r_\varepsilon Tz + \varepsilon(\langle z^*, x_0 \rangle + r_\varepsilon)e = Tx_0 + r_\varepsilon Tz + e,$$

that is,

$$(4.9) \quad T_\varepsilon(x_\varepsilon) = b + r_\varepsilon Tz.$$

It follows from $Tz \leq_C 0$ that $x_\varepsilon \in S(T_\varepsilon, b)$ for all $\varepsilon > 0$ sufficiently small. Since

$$\langle T_\varepsilon^*(c^*), x_\varepsilon \rangle = \langle c^*, T_\varepsilon(x_\varepsilon) \rangle = \langle c^*, b + r_\varepsilon Tz \rangle = \langle c^*, b \rangle$$

(the second and third equalities hold by (4.9) and (4.8), respectively) and

$$\langle T_\varepsilon^*(c^*), x \rangle = \langle c^*, T_\varepsilon x \rangle \leq \langle c^*, b \rangle \quad \text{for all } x \in S(T_\varepsilon, b),$$

one has that $T_\varepsilon^*(c^*) \in N(S(T_\varepsilon, b), x_\varepsilon)$. By (4.8) and the definition of T_ε , one has that

$$T_\varepsilon^*(c^*) = T^*(c^*) + \varepsilon \langle c^*, e \rangle z^* = \varepsilon \langle c^*, e \rangle z^*.$$

It follows from (4.7) and [5, Thm. 1] that $z^* \in \partial \text{dist}(\cdot, S(T_\varepsilon, b))(x_\varepsilon)$, and so

$$r_\varepsilon = \langle z^*, (x_\varepsilon + r_\varepsilon z) - x_\varepsilon \rangle \leq \text{dist}(x_\varepsilon + r_\varepsilon z, S(T_\varepsilon, b)).$$

This and assumption (4.6) imply that

$$(4.10) \quad r_\varepsilon \leq \tau \text{dist}(T_\varepsilon(x_\varepsilon + r_\varepsilon z), b - C).$$

By (4.9) one has that

$$\begin{aligned} T_\varepsilon(x_\varepsilon + r_\varepsilon z) &= T_\varepsilon(x_\varepsilon) + r_\varepsilon T_\varepsilon(z) = b + r_\varepsilon Tz + r_\varepsilon T_\varepsilon(z) \\ &= b + 2r_\varepsilon Tz + r_\varepsilon \varepsilon \langle z^*, z \rangle e = b + 2r_\varepsilon Tz + r_\varepsilon \varepsilon e. \end{aligned}$$

This shows that

$$\text{dist}(T_\varepsilon(x_\varepsilon + r_\varepsilon z), b - C) = \text{dist}(r_\varepsilon \varepsilon e, -C - 2r_\varepsilon Tz) \leq r_\varepsilon \varepsilon \|e\|$$

(the last inequality holds because $Tz \leq_C 0$). This and (4.10) imply that $1 \leq \tau \varepsilon \|e\|$, which is absurd if one takes ε sufficiently small. The proof is completed. $\square$

Remark. From the first half of the proof of Proposition 4.1, one can see that $\text{Rec}(S(T, b)) \subset T^{-1}(-C)$. It is clear that $T^{-1}(-C) \subset \text{Rec}(S(T, b))$ if $S(T, b) \neq \emptyset$. Therefore, $T^{-1}(-C) = \text{Rec}(S(T, b))$ if $S(T, b) \neq \emptyset$. Thus, by virtue of Theorem 4.3 and Proposition 4.1, one has the following result.

Suppose that $T^{-1}(C) \neq \{0\}$. Then (1.3) satisfies the Slater condition if and only if there exist $\tau, \delta > 0$ such that

$$\text{dist}(x, S(\tilde{T}, \tilde{b})) \leq \tau \text{dist}(\tilde{T}x, \tilde{b} - C) \quad \text{for all } x \in X$$

whenever $\tilde{T} \in L(X, Y)$ and $\tilde{b} \in Y$ satisfy $\|\tilde{T} - T\| < \delta$ and $\|\tilde{b} - b\| < \delta$.

Note. The condition $T^{-1}(C) \neq \{0\}$ is satisfied either if T is not injective or if $T(X) \cap C \neq \{0\}$.

Recall that in the case when X is finite-dimensional, $S(T, b)$ is bounded if and only if $\text{Rec}(S(T, b)) = \{0\}$. The following corollary improves Luo and Tseng's result (Theorem LT in section 1); here Y is not restricted to being finite-dimensional and the ordering cone C is a general closed convex cone.

COROLLARY 4.2. *Let X be finite-dimensional. The following conditions are equivalent:*

(a) *Either $0 \in \text{int}(T(X) + C)$, or $b \in \text{int}(T(X) + C)$ and the solution set $S(T, b)$ of (1.3) is bounded.*

(b) *There exist $\tau, \delta > 0$ such that*

$$\text{dist}(x, S(\tilde{T}, \tilde{b})) \leq \tau \text{dist}(\tilde{T}x, \tilde{b} - C) \quad \text{for all } x \in X$$

whenever $\tilde{T} \in L(X, Y)$ and $\tilde{b} \in Y$ satisfy $\|\tilde{T} - T\| + \|\tilde{b} - b\| < \delta$.

Proof. The implication (a) $\implies$ (b) follows from Theorems 3.1 and 3.2. Conversely, suppose that (b) holds. Then it is easy to verify that $b \in \text{int}(T(X) + C)$. By Proposition 4.1, we see now that (a) holds. $\square$

5. Characterizations of error bounds. In this section, we will give dual characterizations of error bounds of system (1.1). For convenience, throughout this section let S denote the solution set of (1.1), namely, $S = \Omega \cap T^{-1}(b - C)$. We always assume that $S \neq \emptyset$.

THEOREM 5.1. *Let $\tau \in (0, \infty)$. Then*

$$(5.1) \quad \text{dist}(x, S) \leq \tau [\text{dist}(x, \Omega) + \text{dist}(Tx, b - C)] \quad \text{for all } x \in X$$

if and only if

$$(5.2) \quad B_{X^*} \cap N(S, z) \subset \tau [B_{X^*} \cap N(\Omega, z) - T^*(B_{Y^*} \cap N(C, b - Tz))] \quad \text{for all } z \in \text{bd}(S).$$

Proof. Suppose that (5.2) holds. Let $x \in X \setminus S$ and $\gamma \in (0, 1)$. By [18, Prop. 1.3], there exist $z \in \text{bd}(S)$ and $x^* \in B_{X^*} \cap N(S, z)$ such that

$$(5.3) \quad \gamma \|x - z\| \leq \langle x^*, x - z \rangle.$$

According to (5.2), there exist $u^* \in B_{X^*} \cap N(\Omega, z)$ and $v^* \in B_{Y^*} \cap N(C, b - Tz)$ such that $x^* = \tau(u^* - T^*(v^*))$. This and (5.3) imply that

$$(5.4) \quad \gamma \|x - z\| \leq \tau (\langle u^*, x - z \rangle - \langle v^*, Tx - Tz \rangle).$$

Noting that

$$u^* \in \partial \text{dist}(\cdot, \Omega)(z) \quad \text{and} \quad v^* \in \partial \text{dist}(\cdot, C)(b - Tz)$$

(by (2.8)), one has that

$$\langle u^*, x - z \rangle \leq \text{dist}(x, \Omega) - \text{dist}(z, \Omega) = \text{dist}(x, \Omega)$$

and

$$-\langle v^*, Tx - Tz \rangle = \langle v^*, -Tx + b - (b - Tz) \rangle \leq \text{dist}(-Tx + b, C) = \text{dist}(Tx, b - C).$$

It follows from (5.4) that

$$\gamma \|x - z\| \leq \tau(\text{dist}(x, \Omega) + \text{dist}(Tx, b - C)).$$

Noting that $\text{dist}(x, S) \leq \|x - z\|$ and letting $\gamma \rightarrow 1$, we see that (5.1) holds.

Conversely, suppose that (5.1) holds. Letting $f(x) := \text{dist}(x, \Omega) + \text{dist}(Tx, b - C)$ for all $x \in X$, we have that $\inf\{f(x) : x \in X\} = 0$ and $S = \text{argmin} f$. It follows from (5.1) and [24, Thm. 3.10.1] that $B_{X^*} \cap N(S, z) \subset \tau \partial f(z)$ for all $z \in \text{bd}(S)$. On the other hand, by [7, Thm. 2.3.10], (2.8), and (2.9), one has that for any $z \in \text{bd}(S)$,

$$\begin{aligned} \partial f(z) &= \partial \text{dist}(\cdot, \Omega)(z) + \partial \text{dist}(T(\cdot), b - C)(z) \\ &= B_{X^*} \cap N(\Omega, z) + T^*(B_{Y^*} \cap N(b - C, Tz)) \\ &= B_{X^*} \cap N(\Omega, z) - T^*(B_{Y^*} \cap N(C, b - Tz)). \end{aligned}$$

It follows that (5.2) holds. $\square$

Next we give dual characterization of error bounds for system (1.2) of conic linear inequalities.

THEOREM 5.2. *Let $C_X \subset X$ be a closed convex cone and let $(a, b) \in X \times Y$ satisfy $Ta = b$. Let $S(T, a, b)$ denote the solution set of (1.2). Let $\tau \in (0, +\infty)$. Then*

$$(5.5) \quad \text{dist}(x, S(T, a, b)) \leq \tau[\text{dist}(x, a - C_X) + \text{dist}(Tx, b - C)] \quad \text{for all } x \in X$$

if and only if

$$(5.6) \quad B_{X^*} \cap N(S(T, a, b), a) \subset -\tau[B_{X^*} \cap N(C_X, 0) + T^*(B_{Y^*} \cap N(C, 0))].$$

Proof. Let $\Phi : X \rightarrow X \times Y$ be a bounded linear operator defined by $\Phi(x) = (x, Tx)$ for all $x \in X$ and equip $X \times Y$ with the norm $\|(x, y)\| = \|x\| + \|y\|$ for any $(x, y) \in X \times Y$. Let

$$f(x) := \text{dist}(x, \Phi^{-1}(-(C_X \times C))) \quad \text{and} \quad g(x) := \tau \text{dist}(\Phi(x), -(C_X \times C))$$

for all $x \in X$.

By $Ta = b$, it is clear that (5.5) is equivalent to $f(x) \leq g(x)$ for all $x \in X$. Since C_X and C are closed convex cones, f and g are positively homogeneous, subadditive, and continuous functions. Thus,

$$f(x) = \sup\{\langle x^*, x \rangle : x^* \in \partial f(0)\} \quad \text{and} \quad g(x) = \sup\{\langle x^*, x \rangle : x^* \in \partial g(0)\}$$

for all $x \in X$.

This and [7, Prop. 2.1.4] imply that $f(x) \leq g(x)$ for all $x \in X$ if and only if $\partial f(0) \subset \partial g(0)$. By (2.8) and (2.9) one has that

$$\partial f(0) = B_{X^*} \cap N(\Phi^{-1}(-(C_X \times C)), 0)$$

and

$$\partial g(0) = \tau \Phi^*[B_{(X \times Y)^*} \cap N(-(C_X \times C), \Phi(0))].$$

Since $\Phi^*(x^*, y^*) = x^* + T^*(y^*)$ for all $(x^*, y^*) \in X^* \times Y^*$,

$$\Phi^{-1}(-(C_X \times C)) = S(T, a, b) - a \quad \text{and} \quad N(-(C_X \times C), \Phi(0)) = N(-C_X, 0) \times N(-C, 0),$$

it follows that $\partial f(0) \subset \partial g(0)$ is equivalent to (5.6). This completes the proof. $\square$

Acknowledgments. The authors thank the two referees for their helpful comments, which improved the proofs of some results.

REFERENCES

- [1] D. AZÉ AND J.-N. CORVELLEC, *On the sensitivity analysis of Hoffman constants for systems of linear inequalities*, SIAM J. Optim., 12 (2002), pp. 913–927.
- [2] D. AZÉ AND J.-N. CORVELLEC, *Characterizations of error bounds for lower semicontinuous functions on metric spaces*, ESAIM Control Optim. Calc. Var., 10 (2004), pp. 409–425.
- [3] J. F. BONNANS AND A. SHAPIRO, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, New York, 2000.
- [4] J. M. BORWEIN, *Norm duality for convex process and applications*, J. Optim. Theory Appl., 48 (1986), pp. 53–64.
- [5] J. V. BURKE AND M. C. FERRIS, *On the Clarke subdifferential of the distance function of a closed set*, J. Math. Anal. Appl., 166 (1992), pp. 512–534.
- [6] J. V. BURKE AND P. TSENG, *A unified analysis of Hoffman’s bound via Fenchel duality*, SIAM J. Optim., 6 (1996), pp. 265–282.
- [7] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, CRM, Montreal, Canada, 1989.
- [8] O. GÜLER, A. J. HOFFMAN, AND U. G. ROTHBLUM, *Approximations to solutions to systems of linear inequalities*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 688–696.
- [9] A. J. HOFFMAN, *On approximate solutions of systems of linear inequalities*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 263–265.
- [10] G. JAMESON, *Ordered Linear Spaces*, Springer-Verlag, Berlin, 1970.
- [11] A. JOURANI, *Hoffman’s error bound, local controllability, and sensitivity analysis*, SIAM J. Control Optim., 38 (2000), pp. 947–970.
- [12] A. S. LEWIS, *Ill-conditioned convex processes and conic linear systems*, Math. Oper. Res., 24 (1999), pp. 829–834.
- [13] A. S. LEWIS, *Ill-conditioned inclusions*, Set-Valued Anal., 9 (2001), pp. 375–381.
- [14] A. LEWIS AND J. S. PANG, *Error bounds for convex inequality systems*, in Generalized Convexity, Generalized Monotonicity: Recent Results, Proceedings of the Fifth Symposium on Generalized Convexity, Luminy, June 1996, J.-P. Crouzeix, J.-E. Martinez-Legaz, and M. Volle, eds., Kluwer Academic Publishers, Dordrecht, 1998, pp. 75–100.
- [15] W. LI AND I. SINGER, *Global error bounds for convex multifunctions and applications*, Math. Oper. Res., 23 (1998), pp. 443–462.
- [16] X.-D. LUO AND Z.-Q. LUO, *Extension of Hoffman’s error bound to polynomial systems*, SIAM J. Optim., 4 (1994), pp. 383–392.
- [17] Z.-Q. LUO AND P. TSENG, *Perturbation analysis of a condition number for linear systems*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 636–660.
- [18] K. F. NG AND W. H. YANG, *Error bounds for abstract linear inequality systems*, SIAM J. Optim., 13 (2002), pp. 24–43.
- [19] J. S. PANG, *Error bounds in mathematical programming*, Math. Prog., 79 (1997), pp. 299–332.
- [20] H. RÅDSTRÖM, *An embedding theorem for spaces of convex sets*, Proc. Amer. Math. Soc., 3 (1952), pp. 165–169.
- [21] S. M. ROBINSON, *Regularity and stability for convex multivalued functions*, Math. Oper. Res., 1 (1976), pp. 130–143.
- [22] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [23] C. ZĂLINESCU, *Sharp estimates for Hoffman’s constant for systems of linear inequalities and equalities*, SIAM J. Optim., 14 (2003), pp. 517–533.
- [24] C. ZĂLINESCU, *Convex Analysis in General Vector Spaces*, World Scientific, Singapore, 2002.

ERROR BOUNDS FOR SOME CONVEX FUNCTIONS AND DISTANCE COMPOSITE FUNCTIONS*

KUNG FU NG[†] AND WEI HONG YANG[‡]

Abstract. In this paper we study error bounds for function $\varphi(x) = \text{dist}(Ax + b, K)$, where A is a continuous linear operator from a Banach space X into another Banach space Y , $b \in Y$, and $K \subset Y$ is a closed set. Assuming that A is of closed range $\text{Im}(A)$, φ is shown to have a Lipschitz error bound (Υ -error bound) if and only if $\{\text{Im}(A) + b, K\}$ is linearly regular (regular). Next we identify exactly when the convex function $g = \sqrt{\zeta} - l$ has error bounds (of various types), where ζ is a quadratic function satisfying $\zeta \geq 0$ and l is an affine function on $\mathbb{R}^n$. As an application, we study the existence of various error bounds for the function $\varphi : x \mapsto \text{dist}(Ax + b, K)$ to supplement the results reported in an earlier paper [K. F. Ng and W. H. Yang, *SIAM J. Optim.*, 13 (2002), pp. 24–43] for the standard second-order cone K in $\mathbb{R}^n$.

Key words. abstract linear inequality system, error bound, regular, linearly regular, Robinson–Ursescu theorem, second-order cone

AMS subject classifications. 90C31, 90C25, 49J52

DOI. 10.1137/040603346

1. Introduction and notation. In generalizing the well-known fundamental result of Hoffman [7, 17], Burke and Tseng studied in [3] the convex set inclusion problem

$$(1.1) \quad Ax + b \in C$$

and derived general conditions on which (1.1) has a Lipschitz error bound in the sense that there exists a constant $\tau > 0$ such that

$$(1.2) \quad \text{dist}(x, S) \leq \tau \text{dist}(Ax + b, C) \quad \forall x \in X.$$

Here A is a continuous linear operator from a Banach space X (or a normed vector space) into another Banach space Y , b is an element of Y , C is a closed convex cone (or, more generally, a closed convex set) in Y , $\text{dist}(y, C)$ denotes the distance from y to C , while $\text{dist}(x, S)$ has a similar meaning, and S denotes the solution set of (1.1):

$$(1.3) \quad S = \{x \in X : Ax + b \in C\}.$$

The results and the methods of [3] are mainly in terms of the (Fenchel) duality. In section 3 of this paper, we will give conditions under which (1.1) has a Lipschitz error bound for any closed set K instead of C . Our conditions are in terms of b and the image $\text{Im}(A)$ of A . In fact, we deal with not only the Lipschitz case but also what we call the Υ -error bounds (such as the Hölder error bounds, that is, error bounds with fractional exponents).

*Received by the editors January 19, 2004; accepted for publication (in revised form) November 2, 2004; published electronically June 30, 2005. This research was supported by an Earmarked Grant from the Research Grant Council of Hong Kong.

<http://www.siam.org/journals/siopt/15-4/60334.html>

[†]Department of Mathematics, The Chinese University of Hong Kong, Shatin, New Territory, Hong Kong (kfng@math.cuhk.edu.hk).

[‡]Department of Mathematics, Fudan University, Shanghai, P.R. China (whyang@fudan.edu.cn).

When $Y = \mathbb{R}^n$, in addition to the most important cone $\mathbb{R}_+^n$ there is another interesting and important cone $\mathcal{K}$, consisting of all vectors $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ such that

$$(1.4) \quad \sqrt{\sum_{i=1}^{n-1} x_i^2} \leq x_n.$$

This cone is the simplest second-order cone in $\mathbb{R}^n$, and it can be used to represent several common convex constraints in mathematical programming. There are several names for $\mathcal{K}$: the standard second-order cone, ice-cream cone, and Lorentz cone. See [5, 10] for further information. Many researchers have studied error bounds for second-order cone programming; see [11, 12, 15]. In [15], the authors only gave conditions on which (1.1) with $C = \mathcal{K}$ has a Lipschitz error bound. Luo and Sturm consider a Cartesian product of second-order cones in [12]. However, their study in this connection was focused on the issue of requiring (1.2) to hold for x in a bounded set. In section 5, we consider the global issue and clarify exactly when (1.1) with $C = \mathcal{K}$ has a Lipschitz error bound, or error bounds of other types. Our analysis is based on section 4, where we study error bounds for function g of the form

$$(1.5) \quad g(x) = \sqrt{\zeta(x)} - l(x) \quad \forall x \in \mathbb{R}^n,$$

where l is an affine function and ζ is a quadratic function on $\mathbb{R}^n$ such that

$$(1.6) \quad \zeta(y) \geq 0 \quad \forall y \in \mathbb{R}^n.$$

2. Definitions and preliminary results. We use $X, Y, Z \dots$ to denote Banach, normed, and metric spaces. For a set K , $\text{int}(K)$ and $\text{bd}(K)$ denote the (topological) interior and the boundary of K , respectively. Let $B(x, r)$ denote the closed ball with center x and radius r . For a map T from X into Y , let $\text{Ker}(T)$ and $\text{Im}(T)$ denote the kernel and the image of T , respectively. A linear operator from $\mathbb{R}^m$ to $\mathbb{R}^n$ is identified with a matrix of order $n \times m$. Thus the transpose Q^T of an $n \times m$ matrix Q is regarded as an adjoint operator from $\mathbb{R}^n$ to $\mathbb{R}^m$. We use $x, y, z \dots$ for vectors, $m, n, k \dots$ for natural numbers, and $\alpha, \beta, \gamma \dots$ for scalars. An element y in $\mathbb{R}^n$ is represented as $y = (y_1, \dots, y_n)$ and is also identified with the “column” vector when it involves matrix calculations. For a scalar-valued function $f : X \rightarrow \mathbb{R}$, the solution set of f is denoted by $S_f = \{x \in X : f(x) \leq 0\}$. Let $\underline{d}^+ f(x)(h)$ denote the lower Dini-directional derivative of f at x along the direction h :

$$(2.1) \quad \underline{d}^+ f(x)(h) = \liminf_{\alpha \rightarrow 0+} \frac{f(x + \alpha h) - f(x)}{\alpha}.$$

For a real number α , let α_+ denote the number $\max\{\alpha, 0\}$. Let $\varrho : \mathbb{R}^n \rightarrow \mathbb{R}$ be defined by

$$(2.2) \quad \varrho(y) = \sqrt{\sum_{i=1}^{n-1} y_i^2} - y_n \quad \forall y \in \mathbb{R}^n.$$

Thus $\mathcal{K} = S_{\varrho}$. The following lemma will be useful to us. Part (iii) was proved by Lewis and Pang [9], while other parts can be proved with ease: (i) holds because $t \mapsto t^2$ is strictly convex, (iv) follows from (i), and (ii) can be proved by the fact that $\{\alpha s : \alpha \geq 0\}$ is the normal cone of $\mathcal{K}$ at point u . We omit the details.

LEMMA 2.1. *Let $\mathcal{K}$ be defined by (1.4). Then we have the following:*

- (i) Let $s, t \in \mathcal{K} \setminus \{0\}$. Suppose that s is not a multiple of t . Then the open line-segment (t, s) is contained in $\text{int}(\mathcal{K})$.
- (ii) Let $u = (u_1, \dots, u_{n-1}, u_n) \in \text{bd}(\mathcal{K}) \setminus \{0\}$, and let $s = (u_1, \dots, u_{n-1}, -u_n)$. Then, for any $h \in \mathbb{R}^n$, $\langle s, h \rangle < 0$ if and only if $u + \alpha h \in \text{int}(\mathcal{K})$ for some $\alpha > 0$.
- (iii) $\frac{[\varrho(y)]_+}{\sqrt{2}} \leq \text{dist}(y, \mathcal{K}) \leq [\varrho(y)]_+ \forall y \in \mathbb{R}^n$.
- (iv) Suppose that L is a linear variety (a translate of a vector subspace) in $\mathbb{R}^n$. If $0 \notin L$ and $L \cap \text{int}(\mathcal{K}) = \emptyset$, then $L \cap \mathcal{K}$ is either empty or a singleton.

A function $\Upsilon : (0, +\infty) \rightarrow (0, +\infty)$ is called an admissible function if Υ is nondecreasing and $\lim_{t \rightarrow 0} \Upsilon(t) = 0$. The set of all admissible functions is denoted by Γ .

DEFINITION 2.1. Let Z be a metric space, and let $\Upsilon \in \Gamma$. A lower semicontinuous function $f : Z \rightarrow (-\infty, +\infty]$ is said to have

(a) a Υ -error bound if

$$(2.3) \quad \text{dist}(x, S_f) \leq \Upsilon([f(x)]_+) \quad \forall x \in Z.$$

(b) a Lipschitz error bound if there exists $\tau > 0$ such that

$$(2.4) \quad \text{dist}(x, S_f) \leq \tau [f(x)]_+ \quad \forall x \in Z.$$

(c) a Hölder error bound with exponent $\gamma \in (0, 1)$ if there exists $\tau > 0$ such that

$$(2.5) \quad \text{dist}(x, S_f) \leq \tau [f(x)]_+^\gamma \quad \forall x \in Z.$$

(d) a Robinson–Ursescu type error bound if there exist $\alpha, \beta > 0$, $\gamma \in (0, 1]$, and $\bar{x} \in S_f$ such that

$$(2.6) \quad \text{dist}(x, S_f) \leq (\alpha \text{dist}(x, \bar{x}) + \beta) [f(x)]_+^\gamma \quad \forall x \in Z.$$

Remark 2.1. Let X, Y be normed spaces. If $\hat{g} : X \times Y \rightarrow \mathbb{R}$ is defined by $\hat{g}(x, y) := g(x)$, then g satisfies one of the conditions (a), (b), (c), (d) if and only if $\hat{g}$ satisfies the corresponding one (when $X \times Y$ is endowed with a usual norm of type p , for example, $\|(x, y)\| = (\|x\|^p + \|y\|^p)^{\frac{1}{p}}$ with $p \geq 1$, or the box norm). Similarly, if $A : X \rightarrow Y$ is a homeomorphic linear operator and $h = g \circ A$, then g satisfies one of the conditions (a), (b), (c), (d) if and only if h satisfies the corresponding one.

Remark 2.2. Definition (a) is related to the notion of well-conditioned functions (see [19, sect. 3.4] and the references therein); more precisely, f has a Υ -error bound for some $\Upsilon \in \Gamma$ if and only if $f_+ := \max\{f, 0\}$ is well-conditioned. Definition (d) is inspired by the Robinson–Ursescu theorem (see [18, Thm. 9.48]). For a convex function f , if there exists x_0 such that $f(x_0) < 0$, then f has a Robinson–Ursescu type error bound for $\gamma = 1$. Indeed, define the set-valued function $F : \mathbb{R}^n \rightarrow 2^{\mathbb{R}}$ by $F(x) = f(x) + [0, +\infty)$. Then 0 belongs to the interior of the range of F and $x_0 \in F^{-1}(0) = S_f$. By [18, Thm. 9.48], there exist $\alpha, \beta > 0$ such that

$$(2.7) \quad \text{dist}(x, S_f) \leq (\alpha \|x - x_0\| + \beta) [f(x)]_+ \quad \forall x \in \mathbb{R}^n.$$

The following proposition provides a concrete example of function f , which has a Robinson–Ursescu type error bound but does not have a Υ -error bound for any $\Upsilon \in \Gamma$; this function will be useful for our study in section 4.

PROPOSITION 2.1. Let m be a natural number such that $m \geq 2$. Let $\eta, \nu \geq 0$, and let $f : \mathbb{R}^m \rightarrow \mathbb{R}$ be defined by

$$(2.8) \quad f(u_1, \dots, u_m) = \sqrt{\sum_{i=1}^m u_i^2} + \eta - u_1 - \nu.$$

Suppose that S_f is nonempty. Then, there exists a sequence $\{z^k\} \subset \mathbb{R}^m$ such that

$$(2.9) \quad \lim_{k \rightarrow \infty} f(z^k) = 0 \quad \text{and} \quad \inf_k \text{dist}(z^k, S_f) > 0.$$

Consequently, for any $\Upsilon \in \Gamma$, f does not have a Υ -error bound (a fortiori, f has neither a Lipschitz error bound nor a Hölder error bound).

Proof. Suppose that $\nu = 0$. Since $S_f \neq \emptyset$, one has that $\eta = 0$ and $S_f = \mathbb{R}_+ \times \{0\}$. For each $k \in \mathbb{N}$, let z^k be the vector in $\mathbb{R}^m$, whose first two coordinates are $k, 1$, respectively, and whose remaining coordinates are zero. Then $\{z^k\}$ has the property stated in (2.9) because $\lim_{k \rightarrow \infty} f(z^k) = \lim_{k \rightarrow \infty} \sqrt{k^2 + 1} - k = 0$, while $\text{dist}(z^k, S_f) = 1$ for each $k \in \mathbb{N}$.

Next we consider the case when $\nu > 0$. For k large enough such that $2k\nu + \nu^2 - \eta > 0$, let $\xi_k := \sqrt{2k\nu + \nu^2 - \eta}$. Then $\xi_k \rightarrow \infty$ as $k \rightarrow +\infty$. Define sequences $\{y^k\}$ and $\{z^k\}$ in $\mathbb{R}^m$ by $y^k = (k, \xi_k, 0, \dots, 0)$ and $z^k = (k, \xi_k + 1, 0, \dots, 0)$. Then $\{y^k\} \subset \text{bd}(S_f)$, and $\{z^k\} \subset \mathbb{R}^m \setminus S_f$ has the desired property (2.9). In fact, it is easy to verify that $f(y^k) = 0$ and

$$\begin{aligned} f(z^k) &= \sqrt{k^2 + (2k\nu + \nu^2 - \eta + 2\xi_k + 1) + \eta} - k - \nu \\ &= \sqrt{(k + \nu)^2 + 2\xi_k + 1} - (k + \nu) \\ &= \frac{2\xi_k + 1}{\sqrt{(k + \nu)^2 + 2\xi_k + 1} + (k + \nu)} \\ &\leq \frac{2\xi_k + 1}{k + \nu}, \end{aligned}$$

which implies that

$$0 \leq \lim_{k \rightarrow \infty} f(z^k) \leq \lim_{k \rightarrow \infty} \frac{2\sqrt{2k\nu + \nu^2 - \eta} + 1}{k + \nu} = 0.$$

Observe that f is differentiable on $\mathbb{R}^m$ if $\eta > 0$ and is differentiable on $\mathbb{R}^m \setminus \{0\}$ if $\eta = 0$. Moreover, $\nabla f(y^k) = (-\frac{\nu}{k+\nu}, \frac{\xi_k}{k+\nu}, 0, \dots, 0)$. Let L_k denote the tangent hyperplane of S_f at y^k . Then z^k and S_f lie on the opposite sides of L_k as $\langle z^k - y^k, \nabla f(y^k) \rangle = \frac{\xi_k}{k+\nu} > 0$. Hence

$$\begin{aligned} \text{dist}(z^k, S_f) &\geq \text{dist}(z^k, L_k) = \left\langle \frac{\nabla f(y^k)}{\|\nabla f(y^k)\|}, z^k - y^k \right\rangle \\ &= \frac{\xi_k}{\sqrt{\xi_k^2 + \nu^2}} \rightarrow 1 \quad \text{as } k \rightarrow \infty. \end{aligned}$$

Thus the second relation of (2.9) holds. To prove the last assertion, we assume for contradiction that there exists $\Upsilon \in \Gamma$ such that (2.3) holds. Then

$$0 < \inf_k \text{dist}(z^k, S_f) \leq \lim_{k \rightarrow \infty} \Upsilon(f(z^k)) = 0,$$

which is a contradiction. This completes the proof. $\square$

THEOREM 2.1. *Let $f : \mathbb{R}^m \rightarrow \mathbb{R}$ be defined by (2.8). Suppose that $\eta, \nu \geq 0$ and $S_f \neq \emptyset$. Then there exist $\alpha, \beta > 0$, $\gamma \in (0, 1]$, and $\bar{x} \in S_f$ such that*

$$(2.10) \quad \text{dist}(z, S_f) \leq (\alpha\|z - \bar{x}\| + \beta)[f(z)]_+^\gamma \quad \forall z \in \mathbb{R}^m.$$

Proof. If $\nu > 0$, then there exists $\bar{x}$ such that $f(\bar{x}) < 0$. It follows from (2.7) that (2.10) holds with $\gamma = 1$.

If $\nu = 0$, then $\eta = 0$ and $S_f = \mathbb{R}_+ \times \{0\}$, as noted before. Next we prove that

$$(2.11) \quad \text{dist}(z, S_f) \leq (\|z\| + 1)[f(z)]_+^{1/2} \quad \forall z \in \mathbb{R}^m.$$

Let $z = (z_1, z_2, \dots, z_m) \in \mathbb{R}^m$. If $z_1 \leq 0$, then $\text{dist}(z, S_f) = \|z\|$ and $f(z) \geq \|z\|$. Thus,

$$\text{dist}(z, S_f) \leq 2\|z\| \leq (\|z\| + 1)\|z\|^{1/2} \leq (\|z\| + 1)[f(z)]_+^{1/2}.$$

If $z_1 > 0$, then $\text{dist}(z, S_f) = \sqrt{\sum_{i=2}^m z_i^2}$. Thus,

$$f(z) = \sqrt{\sum_{i=1}^m z_i^2} - z_1 = \frac{\sum_{i=2}^m z_i^2}{\sqrt{\sum_{i=1}^m z_i^2} + z_1} > \frac{d^2(z, S_f)}{2\|z\|} \geq \frac{d^2(z, S_f)}{(1 + \|z\|)^2}.$$

Therefore (2.11) holds. $\square$

We shall also need the following elementary fact.

LEMMA 2.2. For $\eta > 0$, let $\omega : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be defined by $\omega(z) = \frac{z}{\sqrt{\|z\|^2 + \eta}}$. Then ω is homeomorphic from $\mathbb{R}^n$ onto the open unit ball.

DEFINITION 2.2 (cf. [1, 2, 16]). A pair $\{C_1, C_2\}$ of closed sets in a metric space Z is said to be

(a) *regular* if, for any sequence $\{x^k\}$ in Z , it holds that

$$(2.12) \quad \max\{\text{dist}(x^k, C_1), \text{dist}(x^k, C_2)\} \rightarrow 0 \Rightarrow \text{dist}(x^k, C_1 \cap C_2) \rightarrow 0;$$

(b) *linearly regular* if there exists $\kappa > 0$ such that

$$(2.13) \quad \text{dist}(x, C_1 \cap C_2) \leq \kappa \max\{\text{dist}(x, C_1), \text{dist}(x, C_2)\} \quad \forall x \in Z.$$

3. Regularity and error bounds. Throughout this section, let A be a continuous linear operator from a Banach space X into a Banach space Y . For $b \in Y$, let $T : X \rightarrow Y$ be defined by $Tx = Ax + b$ for each $x \in X$. Let us begin with the following elementary result.

LEMMA 3.1. Assume that $\text{Im}(A)$ is closed. Let C be a nonempty subset of X such that

$$(3.1) \quad C + \text{Ker}(A) = C.$$

Then there exists $\alpha > 0$ such that

$$(3.2) \quad \alpha \text{dist}(x, C) \leq \text{dist}(Tx, TC) \leq \|A\| \text{dist}(x, C) \quad \forall x \in X.$$

Proof. The second inequality of (3.2) is obvious. Let $\hat{x}$ and $\hat{C}$ be the images of x and C by the natural quotient map of X onto $X/\text{Ker}(A)$, which is endowed with the quotient norm. Let $A^{-1} : \text{Im}(A) \rightarrow X/\text{Ker}(A)$ denote the natural homeomorphism defined by $A^{-1}(Ax) = \hat{x}$. By (3.1), it follows that

$$(3.3) \quad \begin{aligned} \text{dist}(x, C) &= \text{dist}(\hat{x}, \hat{C}) = \text{dist}(A^{-1}(Ax), A^{-1}(AC)) \\ &\leq \|A^{-1}\| \text{dist}(Ax, AC) = \|A^{-1}\| \text{dist}(Tx, TC) \quad \forall x \in X. \end{aligned}$$

This proves the first inequality of (3.2) with $\alpha := \frac{1}{\|A^{-1}\|}$. $\square$

Let K be a closed set in Y . Let $\psi : \text{Im}(T) \rightarrow \mathbb{R}$ be defined by

$$(3.4) \quad \psi(y) = \text{dist}(y, K) \quad \forall y \in \text{Im}(T),$$

and let $\varphi : X \rightarrow \mathbb{R}$ be defined by

$$(3.5) \quad \varphi = \psi \circ T, \quad \text{i.e.,} \quad \varphi(x) = \psi(T(x)) \quad \forall x \in X.$$

Then $S_\psi = \text{Im}(T) \cap K = T(S_\varphi)$.

The following blanket assumptions will be assumed in what follows:

$$(3.6) \quad \begin{array}{l} \text{Im}(A) \text{ is closed (hence } \text{Im}(T) = \text{Im}(A) + b \text{ is also closed)} \\ \text{and } \text{Im}(T) \cap K \neq \emptyset. \end{array}$$

THEOREM 3.1. *Let ψ, φ be defined by (3.4) and (3.5), and suppose that (3.6) holds. Then the following assertions are valid:*

(a) *There exists $\alpha > 0$ such that*

$$(3.7) \quad \alpha \text{dist}(x, S_\varphi) \leq \text{dist}(Tx, S_\psi) \leq \|A\| \text{dist}(x, S_\varphi) \quad \forall x \in X.$$

Consequently, φ has an error bound of any type ((a), (b), (c), or (d) from Definition 2.1) if and only if ψ has the corresponding one.

(b) *The following statements are equivalent:*

- (i) *φ has a Lipschitz error bound.*
- (ii) *ψ has a Lipschitz error bound.*
- (iii) *There exists $k > 0$ such that*

$$(3.8) \quad \text{dist}(y, \text{Im}(T) \cap K) \leq k \text{dist}(y, K) \quad \forall y \in \text{Im}(T).$$

(iv) *$\{\text{Im}(T), K\}$ is linearly regular.*

(c) *The following statements are equivalent:*

- (i) *φ has a Υ -error bound for some $\Upsilon \in \Gamma$.*
- (ii) *ψ has a Υ -error bound for some $\Upsilon \in \Gamma$.*
- (iii) *For any sequence $\{y^k\} \subset \text{Im}(T)$,*

$$(3.9) \quad \lim_{k \rightarrow \infty} \psi(y^k) = 0 \Rightarrow \lim_{k \rightarrow \infty} \text{dist}(y^k, S_\psi) = 0.$$

(iv) *$\{\text{Im}(T), K\}$ is regular.*

Proof. It is routine to verify that (3.1) is satisfied by $C := S_\varphi$. Since $S_\psi = TC$, by Lemma 3.1 it follows that (3.7) holds with $\alpha = \frac{1}{\|A^{-1}\|}$. Next suppose that φ has a Robinson–Ursescu type error bound: there exists $\alpha, \beta > 0$, $\gamma \in (0, 1]$, and $\bar{x} \in S_\varphi$ such that

$$(3.10) \quad \text{dist}(x, S_\varphi) \leq (\alpha \|x - \bar{x}\| + \beta) [\varphi(x)]_+^\gamma \quad \forall x \in X.$$

For any $y \in \text{Im}(T)$, there exist $x, x^k \in X$ such that $Tx = Tx^k = y$ and

$$(3.11) \quad \|\bar{x} - x^k\| \leq \text{dist}(\bar{x}, x + \text{Ker}(A)) + \frac{1}{k}.$$

By (3.7), (3.10), and (3.11), one has that

$$\begin{aligned}
 \text{dist}(y, S_\psi) &= \text{dist}(Tx^k, S_\psi) \leq \|A\| \text{dist}(x^k, S_\varphi) \\
 &\leq \|A\|(\alpha\|x^k - \bar{x}\| + \beta)[\varphi(x^k)]_+^\gamma \\
 &\leq \|A\| \left(\alpha \cdot \text{dist}(\bar{x}, x + \text{Ker}(A)) + \frac{\alpha}{k} + \beta \right) [\psi(y)]_+^\gamma \\
 (3.12) \quad &\leq \|A\| \left(\alpha\|A^{-1}\| \text{dist}(T\bar{x}, y) + \frac{\alpha}{k} + \beta \right) [\psi(y)]_+^\gamma,
 \end{aligned}$$

where (3.12) holds by Lemma 3.1. Note that $T\bar{x} \in S_\psi$. Letting $k \rightarrow \infty$, we prove that ψ has a Robinson–Ursescu type error bound. The converse of this implication can be proved directly as well as for other parts of (a) of Theorem 3.1 regarding the types (a), (b), and (c) of Definition 2.1.

In particular, (i) $\Leftrightarrow$ (ii) holds in (b) and in (c). By $S_\psi = \text{Im}(T) \cap K$, (iii) $\Leftrightarrow$ (iv) holds in (c) thanks to [16, Thm. 3.1(a)]. Similarly, [16, Thm. 3.1(b)] implies that (iii) $\Leftrightarrow$ (iv) holds in (b). Since (iii) is simply a restatement of (ii), (b) is proved. Finally, by [16, Cor. 3.1] (applied to the function ψ defined on the metric space $\text{Im}(T)$), one has (ii) $\Leftrightarrow$ (iii) in (c); thus (c) is also proved. $\square$

4. Error bounds for convex function $g = \sqrt{\zeta} - l$. This section is devoted to error bound issues for a function g defined by

$$(4.1) \quad g(x) = \sqrt{\zeta(x)} - l(x) \quad \forall x \in \mathbb{R}^n,$$

where l is an affine function and ζ is a quadratic function on $\mathbb{R}^n$ such that

$$(4.2) \quad \zeta(\cdot) \geq 0 \quad \text{on} \quad \mathbb{R}^n.$$

Note in particular that ζ must be convex. By an appropriate coordinate transform (i.e., by an affine map from $\mathbb{R}^n$ onto itself) if necessary, we assume without loss of generality that ζ takes the form

$$(4.3) \quad \zeta(x) = \sum_{i=1}^m x_i^2 + \eta \quad \forall x \in \mathbb{R}^n,$$

where $\eta \geq 0$ (because of (4.2)) and $m \leq n$. Note that $\sqrt{\zeta}$ is a convex function. Let $a \in \mathbb{R}^n$, and let $\nu \in \mathbb{R}$ be such that

$$(4.4) \quad l(x) = \langle a, x \rangle + \nu \quad \forall x \in \mathbb{R}^n.$$

Throughout this section we make the following blanket assumptions:

$$(4.5) \quad S_g \text{ is nonempty, } \eta \geq 0, \text{ and (4.1)–(4.4) hold.}$$

THEOREM 4.1. *Suppose there exists $i_0 > m$ such that $a_{i_0} \neq 0$. Then g has a Lipschitz error bound.*

Proof. Let $h \in \mathbb{R}^n$ with $\frac{a_{i_0}}{|a_{i_0}|}$ at the i_0 th coordinate and zero at the other coordinates. Then, for each $x \in \mathbb{R}^n$, one has that

$$\underline{d}^+ g(x)(h) = -\langle a, h \rangle = -|a_{i_0}| < 0.$$

Hence g has a Lipschitz error bound by [13, Thm. 2.5]. $\square$

In the rest of this section (except in Corollary 4.2), we further assume that

$$(4.6) \quad a_i = 0 \quad \text{for each } i > m.$$

The following result implies that the consideration for error bounds for g can be reduced to that for error bounds for a simpler function.

THEOREM 4.2. *Under assumptions (4.5) and (4.6), there exists a continuous linear operator $Q : \mathbb{R}^m \rightarrow \mathbb{R}^n$ such that*

$$(4.7) \quad gQ(u) = \sqrt{\sum_{i=1}^m u_i^2 + \eta - \|a\|u_1} - \nu \quad \forall u \in \mathbb{R}^m,$$

where gQ is interpreted as the composite map.

Proof. Note that $g(x) = g(y)$ if $x, y \in \mathbb{R}^n$ and $x_i = y_i$ for each $i \leq m$. Let M be an orthonormal matrix of order $m \times m$ (with m rows and m columns) such that if $a \neq 0$, the first column is composed of the first m coordinates of $\frac{a}{\|a\|}$, namely,

$$(4.8) \quad m_{i1} = \frac{a_i}{\|a\|} \quad \forall i = 1, \dots, m;$$

if $a = 0$, we take $M = I_{m \times m}$, the identity matrix. Let Q denote the matrix of order $n \times m$:

$$(4.9) \quad Q = \begin{pmatrix} M \\ O_{(n-m) \times m} \end{pmatrix};$$

that is, the first m rows of Q agree with that of M while the other rows are zero. Since $(a_1, \dots, a_m)$ is the $\|a\|$ -multiple of the first column of M and is perpendicular to the other columns, one has that

$$(a_1, \dots, a_m)Mu = \|a\|u_1 \quad \forall u \in \mathbb{R}^m.$$

Moreover, by (4.3), one has that

$$\zeta(Qu) = \|Mu\|^2 + \eta = (Mu)^T(Mu) + \eta = (u)^T u + \eta \quad \forall u \in \mathbb{R}^m.$$

It follows that

$$\begin{aligned} g(Qu) &= \sqrt{\zeta(Qu) + \eta} - \langle a, Qu \rangle - \nu \\ &= \sqrt{\sum_{i=1}^m u_i^2 + \eta - \|a\|u_1} - \nu \quad \forall u \in \mathbb{R}^m, \end{aligned}$$

which verifies (4.7). $\square$

COROLLARY 4.1. *Write $\tilde{g}$ for gQ . Then $\tilde{g}$ has an error bound of any type ((a), (b), (c), or (d) from Definition 2.1) if and only if g has the corresponding one.*

Proof. By assumption (4.6) and identifying $\mathbb{R}^n$ with $\mathbb{R}^m \times \mathbb{R}^{n-m}$, we can define $g_1 : \mathbb{R}^m \rightarrow \mathbb{R}$ by $g_1(u) = g(x)$, where $x = (u, v) \in \mathbb{R}^n$. Write $\tilde{g}$ for gQ . By (4.9) and the definition of g_1 , one has that $\tilde{g}(u) = g(Qu) = g_1(Mu) = g_1 \circ M$. Hence $\tilde{g}$ has an error bound of any type ((a), (b), (c), or (d) from Definition 2.1) if and only if g_1 , and so g , has the corresponding one, where g_1 can be replaced by g thanks to Remark 2.1. $\square$

By (4.7), one has that

$$(4.10) \quad \tilde{g}(u) = \sqrt{\|u\|^2 + \eta} - \|a\|u_1 - \nu \quad \forall u \in \mathbb{R}^m.$$

Moreover, $\tilde{g}$ is convex, differentiable on $\mathbb{R}^m$ if $\eta > 0$, and differentiable on $\mathbb{R}^m \setminus \{0\}$ if $\eta = 0$. For any $u \in \mathbb{R}^m$ (for $u \neq 0$ if $\eta = 0$), one has that

$$(4.11) \quad \nabla \tilde{g}(u) = \left(\frac{u_1}{\sqrt{\|u\|^2 + \eta}} - \|a\|, \frac{u_2}{\sqrt{\|u\|^2 + \eta}}, \dots, \frac{u_m}{\sqrt{\|u\|^2 + \eta}} \right)$$

$$(4.12) \quad = \frac{u}{\sqrt{\|u\|^2 + \eta}} - (\|a\|, 0, \dots, 0).$$

Since S_g is assumed to be nonempty, $S_{\tilde{g}} \neq \emptyset$. The following lemma follows from (4.12) and the convexity of $\tilde{g}$.

LEMMA 4.1. *Assume that $\|a\| < 1$ and $\eta > 0$. Then*

$$\bar{u} = \left\{ \frac{\|a\|\sqrt{\eta}}{\sqrt{1 - \|a\|^2}}, 0, \dots, 0 \right\}$$

is the unique point in $\mathbb{R}^m$ such that $\nabla \tilde{g}(\bar{u}) = 0$. Moreover, $\bar{u}$ is the unique minimum point of $\tilde{g}$, and

$$(4.13) \quad \tilde{g}(\bar{u}) = 0 \Leftrightarrow \sqrt{(1 - \|a\|^2)\eta} = \nu \Leftrightarrow S_{\tilde{g}} = \{\bar{u}\}.$$

THEOREM 4.3. *Suppose that assumptions (4.5) and (4.6) are satisfied. In each of the following cases, $\tilde{g}$ (and thus g) has a Lipschitz error bound:*

- (1) $\|a\| > 1$;
- (2) $\|a\| = 1$ and $m = 1$;
- (3) $\|a\| < 1$ and $\eta = 0$;
- (4) $\|a\| < 1$, $\eta > 0$, and $\tilde{g}(\bar{u}) < 0$,

where $\bar{u}$ is as in the preceding lemma.

Proof. (1) Suppose that $\|a\| > 1$. Let $e^1 = (1, 0, \dots, 0)$. By (4.12), one has that

$$\underline{d}^+ \tilde{g}(x)(e^1) \leq 1 - \|a\| < 0 \quad \forall x \in \mathbb{R}^m.$$

Then, by [13, Thm. 2.5], $\tilde{g}$ has a Lipschitz error bound.

(2) Suppose that $\|a\| = 1$ and $m = 1$. Then $\tilde{g}$ is defined on $\mathbb{R}$. If $\eta = 0$, then $\tilde{g}(\alpha)$ equals $-\nu$ for $\alpha \geq 0$ and $-2\alpha - \nu$ for $\alpha < 0$. It is trivial to see that $\tilde{g}$ has a Lipschitz error bound. Suppose that $\eta > 0$. Since $S_{\tilde{g}}$ is nonempty, it follows that $\nu > 0$. Let $\omega := \frac{\eta - \nu^2}{2\nu}$. Then $S_{\tilde{g}} = \{\alpha \in \mathbb{R} : \alpha \geq \omega\}$. Since $\tilde{g}'(\alpha) = \frac{\alpha}{\sqrt{\alpha^2 + \eta}} - 1$ and $\frac{\alpha}{\sqrt{\alpha^2 + \eta}}$ is a strictly increasing function, $\tilde{g}'(\alpha) < \frac{\omega}{\sqrt{\omega^2 + \eta}} - 1$ for each $\alpha \notin S_{\tilde{g}}$. Then by [13, Thm. 2.5] again, $\tilde{g}$ has a Lipschitz error bound when $\eta > 0$.

(3) Thus we have $\nu \geq 0$ (as $S_{\tilde{g}}$ is nonempty) and $0 \in S_{\tilde{g}}$. Hence $u \neq 0$ for each $u \notin S_{\tilde{g}}$. For each $u \neq 0$, by (4.12), one has $\|\nabla \tilde{g}(u)\|^2 = \|\frac{u}{\|u\|} - \|a\|e^1\|^2 \geq \text{dist}(\|a\|e^1, S(0, 1))^2 = 1 - \|a\|^2$, where $e^1 = (1, 0, \dots, 0)$ and $S(0, 1)$ denotes the unit sphere (consisting of all unit vectors). Then $\inf\{\|\nabla \tilde{g}(x)\| : x \in \mathbb{R}^m \setminus S_{\tilde{g}}\} \geq 1 - \|a\| > 0$. By [13, Thm. 2.5], $\tilde{g}$ has a Lipschitz error bound.

(4) By assumption, there exists a neighborhood V of $\bar{u}$ such that $V \subset S_{\tilde{g}}$. By (4.11) and Lemma 2.2, $\nabla \tilde{g}(\cdot)$ is a homeomorphism; since $\nabla \tilde{g}(\bar{u}) = 0$, it follows that

the image of V under $\nabla \tilde{g}(\cdot)$ must contain a ball $B(0, \epsilon)$ in $\mathbb{R}^m$ with some $\epsilon > 0$, and hence that the image of $\mathbb{R}^m \setminus V$ is contained in $\mathbb{R}^m \setminus B(0, \epsilon)$. Therefore $\|\nabla \tilde{g}(u)\| > \epsilon$ for each $u \in \mathbb{R}^m \setminus V$. Since $V \subset S_{\tilde{g}}$, one has that $\inf\{\|\nabla \tilde{g}(u)\| : u \in \mathbb{R}^m \setminus S_{\tilde{g}}\} \geq \epsilon$. Invoking [13, Thm. 2.5], we see that $\tilde{g}$ has a Lipschitz error bound. $\square$

Next, we treat the case when g has a Υ -error bound but has neither a Lipschitz error bound nor a Hölder error bound.

THEOREM 4.4. *Suppose that assumptions (4.5) and (4.6) are satisfied. If $\|a\| < 1$, $\eta > 0$, and $\tilde{g}(\bar{u}) = 0$, where $\bar{u}$ is as in Lemma 4.1, then g has neither Lipschitz error bound nor Hölder error bound, but there exists $\tau > 0$ such that*

$$(4.14) \quad \text{dist}(x, S_g) \leq \tau([g(x)]_+ + [g(x)]_+^{1/2}) \quad \forall x \in \mathbb{R}^n.$$

Proof. To prove (4.14), we assume without loss of generality that $g = \tilde{g}$. Then the Hessian matrix $H(\bar{u})$ of $\tilde{g}$ at $\bar{u}$ is positive definite. Indeed, $H(\bar{u})$ is a diagonal matrix (of zero elements off the diagonal) with the first diagonal element $\frac{(1-\|a\|^2)^{3/2}}{\sqrt{\eta}}$ and other diagonal elements $\sqrt{\frac{1-\|a\|^2}{\eta}}$. Take $\delta > 0$ such that 2δ is smaller than the eigenvalues of $H(\bar{u})$. Since $\nabla \tilde{g}(\bar{u}) = 0$ and $\tilde{g}(\bar{u}) = 0$, one has from the Taylor theorem that

$$(4.15) \quad \tilde{g}(u) > \delta \|u - \bar{u}\|^2 \quad \forall u \in W$$

for some neighborhood W of $\bar{u}$. This and (4.13) imply that

$$(4.16) \quad \text{dist}(u, S_{\tilde{g}}) = \|u - \bar{u}\| \leq \frac{1}{\sqrt{\delta}} [\tilde{g}(u)]^{1/2} \quad \forall u \in W.$$

On the other hand, take $r > 0$ such that

$$(4.17) \quad B(\bar{u}, 2r) \subseteq W.$$

Let $V := B(\bar{u}, r)$. Since $\|a\| < 1$ and $\tilde{g}$ is coercive, there exists $\xi > 0$ such that $\xi = \inf_{x \in \mathbb{R}^m \setminus V} \tilde{g}(x)$. Let $f(u) := \tilde{g}(u) - \frac{\xi}{2}$ for each $u \in \mathbb{R}^m$. Then $S_f \subset V$. By (4.12) and Lemma 2.2, $\nabla \tilde{g}(\cdot)$ is a homeomorphism; since $\nabla \tilde{g}(\bar{u}) = 0$, it follows that the image of V under $\nabla \tilde{g}(\cdot)$ must contain a ball $B(0, \epsilon)$ in $\mathbb{R}^m$ with some $\epsilon > 0$, and hence that the image of $\mathbb{R}^m \setminus V$ is contained in $\mathbb{R}^m \setminus B(0, \epsilon)$. Thus,

$$\|\nabla f(u)\| = \|\nabla \tilde{g}(u)\| > \epsilon \quad \forall u \in \mathbb{R}^m \setminus V.$$

By [13, Thm. 2.5], we have

$$(4.18) \quad \text{dist}(u, V) \leq \text{dist}(u, S_f) \leq \frac{1}{\epsilon} [f(u)]_+ \leq \frac{1}{\epsilon} \tilde{g}(u) \quad \forall u \in \mathbb{R}^m \setminus V.$$

Now consider those u such that $u \in \mathbb{R}^m \setminus W$. This and (4.17) as well as the definition of V imply that $\|u - \bar{u}\| \leq \text{dist}(u, V) + r \leq 2\text{dist}(u, V)$. It follows from (4.18) that

$$\text{dist}(u, S_{\tilde{g}}) = \|u - \bar{u}\| \leq \frac{2}{\epsilon} \tilde{g}(u) \quad \forall u \in \mathbb{R}^m \setminus W.$$

Letting $\tau = \max\{\frac{1}{\sqrt{\delta}}, \frac{2}{\epsilon}\}$ and combining the above with (4.16), we have

$$(4.19) \quad \text{dist}(u, S_{\tilde{g}}) \leq \tau([\tilde{g}(u)]_+ + [\tilde{g}(u)]_+^{1/2}) \quad \forall u \in \mathbb{R}^m$$

(recall that $\tilde{g}(\cdot) \geq 0$ on $\mathbb{R}^m$, so one can write $\tilde{g}(u)$ for $[\tilde{g}(u)]_+$). Invoking Theorem 4.2, we see that (4.14) holds by (4.19).

Notice that $\tilde{g}(\bar{u}) = 0 < \tilde{g}(u)$ for every $u \neq \bar{u}$. Assume that there exist $\gamma \in (0, 1]$ and $\tau > 0$ such that $\|u - \bar{u}\| \leq \tau [\tilde{g}(u)]^\gamma$ for every $u \in \mathbb{R}^m$. For $\gamma = 1$, we obtain that $1 \leq \tau \|u\|^{-1} (\tilde{g}(\bar{u} + u) - \tilde{g}(\bar{u}))$ for every $u \neq 0$. Since $\nabla \tilde{g}(\bar{u}) = 0$, by taking the limit in the preceding inequality for $u \rightarrow 0$ we get the contradiction $1 \leq 0$. For $\gamma < 1$, we get a contradiction by taking $\|u\| \rightarrow \infty$ in the inequality $1 \leq \tau \|u\|^{-1} (\tilde{g}(\bar{u} + u))^\gamma$. $\square$

The only remaining case that we have not yet considered is $a_i = 0$ for each $i > m$, $\|a\| = 1$, and $m \geq 2$.

THEOREM 4.5. *Suppose that assumptions (4.5) and (4.6) are satisfied. If $\|a\| = 1$ and $m \geq 2$, then g has a Robinson–Ursescu type error bound but does not have a Υ -error bound for any $\Upsilon \in \Gamma$. Moreover, S_g is unbounded.*

Proof. By assumption, the function $\tilde{g}$ in (4.10) is of the same type studied in Proposition 2.1 and Theorem 2.1. Thus, $S_{\tilde{g}}$ is unbounded (see the proof of Proposition 2.1) and $\tilde{g}$ has a Robinson–Ursescu type error bound but does not have a Υ -error bound for any $\Upsilon \in \Gamma$. Consequently, by Corollary 4.1, we prove our assertion. $\square$

COROLLARY 4.2. *Let $g := \sqrt{\zeta} - l$, where l is an affine function and ζ is a quadratic function on $\mathbb{R}^n$ such that $\zeta(x) \geq 0$ for each $x \in \mathbb{R}^n$. Suppose that g does not have a Lipschitz error bound and S_g is a singleton. Then the conclusions in Theorem 4.4 hold.*

Proof. As already noted, Theorems 4.1, 4.3, 4.4, and 4.5 exhaust all possible cases for g . But the cases of Theorems 4.1 and 4.3 cannot occur for g (because it does not have the Lipschitz error bound) and the case of Theorem 4.5 cannot occur (because S_g is a singleton). Thus, we have only the case of Theorem 4.4. $\square$

5. Error bounds for inequality system with the second-order cone. For convenience, we now summarize the notation we have used and will continue to use throughout this entire section. Let X be a Banach space. Let $A : X \rightarrow \mathbb{R}^n$ be a continuous linear operator, $b \in \mathbb{R}^n$, and let $T(x) := Ax + b$ for every $x \in X$. Let $M = \text{Im}(A)$. Then $\text{Im}(T) = M + b$. Let m denote the dimension of M . Let $\Theta : \mathbb{R}^m \rightarrow M + b$ be the function defined by

$$(5.1) \quad \Theta(v) = \hat{v} \quad \forall v \in \mathbb{R}^m, \hat{v} \in M + b \text{ such that } \hat{v} = \sum_{i=1}^m v_i w^i + b,$$

where $\{w^i\}_{i=1}^m$ is an orthonormal basis of M . Let $\mathcal{Q}$ be defined by (2.2). We use $\mathcal{K}$ to denote the second-order cone in $\mathbb{R}^n$. Let φ be defined by

$$(5.2) \quad \varphi(x) = \text{dist}(Ax + b, \mathcal{K}) \quad \forall x \in X,$$

and let ψ be defined by

$$(5.3) \quad \psi(y) = \text{dist}(y, \mathcal{K}) \quad \forall y \in \text{Im}(T).$$

We assume that S_φ is nonempty, or equivalently,

$$(5.4) \quad [\text{Im}(A) + b] \cap \mathcal{K} \neq \emptyset.$$

Noting $\varphi(-x) = \text{dist}(Ax - b, -\mathcal{K})$, we see that it follows from [15, Thms. 5.7, 5.8, 5.9] that the function φ has a Lipschitz error bound if and only if one of the following cases occurs:

- (1) $M \cap \text{int}(\mathcal{K}) \neq \emptyset$.
- (2) $M \cap \mathcal{K} = \{0\}$ and $(M + b) \cap \text{int}(\mathcal{K}) \neq \emptyset$.
- (3) $M \cap \mathcal{K} = \{0\}$ and $b \in M$.
- (4) $\{0\} \neq M \cap \mathcal{K} \subset \text{bd}(\mathcal{K})$ and $\dim(M) = 1$.

Then there are two remaining cases to exhaust all the possibilities:

- (a) $\{0\} \neq M \cap \mathcal{K} \subset \text{bd}(\mathcal{K})$ and $\dim(M) \geq 2$.
- (b) $M \cap \mathcal{K} = \{0\}$, $(M + b) \cap \text{int}(\mathcal{K}) = \emptyset$, and $b \notin M$.

These two cases will be studied here (respectively, in Theorems 5.1 and 5.2). Before we present the two theorems, we need a preparatory lemma.

LEMMA 5.1. *For any orthonormal basis $\{w^i\}_{i=1}^m$, let $\Theta : \mathbb{R}^m \rightarrow M + b$ be defined by (5.1). Let ϱ , φ , and ψ be as already defined. Let $g : \mathbb{R}^m \rightarrow \mathbb{R}$ be defined by $g = \varrho \circ \Theta$. For any $v \in \mathbb{R}^m$, $\hat{v} \in M + b$ such that $\hat{v} = \sum_{i=1}^m v_i w^i + b$, the following assertions hold:*

- (i) $\frac{g(v)}{\sqrt{2}} \leq \psi(\hat{v}) \leq g(v)$.
- (ii) $\text{dist}(v, S_g) = \text{dist}(\hat{v}, S_\psi)$.
- (iii) φ has an error bound of any type ((a), (b), (c), or (d) from Definition 2.1) if and only if g has the corresponding one.
- (iv) g can be expressed in the form (4.1) with an affine function l and a quadratic function ξ satisfying (4.2).

Proof. By Lemma 2.1(iii), (i) follows directly. It is easy to see that $\Theta S_g = S_\psi$. This and the isometry of Θ imply (ii). Since (i) and (ii) hold, ψ has an error bound of any type ((a), (b), (c), or (d) from Definition 2.1) if and only if g has the corresponding one, where ψ can be replaced by φ thanks to Theorem 3.1. Next we show that (iv) holds. Let $e = (0, \dots, 0, 1)$. One has that

$$(5.5) \quad g(v) = \varrho(\Theta v) = \sqrt{\|\Theta(v)\|^2 - \langle \Theta(v), e \rangle^2} - \langle \Theta(v), e \rangle \quad \forall v \in \mathbb{R}^m.$$

Thus

$$(5.6) \quad g(v) = \sqrt{\zeta(v)} - l(v) \quad \forall v \in \mathbb{R}^m,$$

where

$$(5.7) \quad \begin{aligned} \zeta(v) &:= \|\Theta(v)\|^2 - \langle \Theta(v), e \rangle^2 \quad \forall v \in \mathbb{R}^m, \\ l(v) &:= \langle \Theta(v), e \rangle \quad \forall v \in \mathbb{R}^m. \end{aligned}$$

It is easy to see that l is an affine function and ζ is a quadratic function such that $\zeta(v) \geq 0$ for each $v \in \mathbb{R}^m$. $\square$

THEOREM 5.1. *Suppose that (5.4) holds and*

$$(5.8) \quad \{0\} \neq M \cap \mathcal{K} \subset \text{bd}(\mathcal{K}) \quad \text{and} \quad \dim(M) \geq 2.$$

Then φ has a Robinson–Ursescu type error bound but does not have a Υ -error bound for any $\Upsilon \in \Gamma$.

Proof. Let P_M denote the orthogonal projection onto M . We assume without loss of generality that b is perpendicular to M (replace b with $b - P_M b$ if necessary). Our proof given below covers the case when $b = 0$. Let m be the dimension of M ; then $2 \leq m < n$. Pick a unit vector $e^1 \in M \cap \mathcal{K}$, and pick $e^2, \dots, e^m \in M$ such that $\{e^i : 1 \leq i \leq m\}$ is an orthonormal base of M . By assumption (5.8) it follows from Lemma 2.1(i) that

$$(5.9) \quad M \cap \mathcal{K} = M \cap \text{bd}(\mathcal{K}) = \{\alpha e^1 : \alpha \geq 0\};$$

in particular e^i , $e^1 + \alpha e^i \notin \mathcal{K} \forall \alpha \in \mathbb{R}$ and $i \geq 2$. Let

$$e^1 := (u_1, \dots, u_{n-1}, u_n) \quad \text{and} \quad s := (u_1, \dots, u_{n-1}, -u_n).$$

By Lemma 2.1(ii), it follows that $\langle s, \pm e^i \rangle \geq 0$; that is, $\langle s, e^i \rangle = 0 \forall i \geq 2$. Denoting the unit vector $(0, \dots, 0, 1)$ in $\mathbb{R}^n$ by e , it follows that

$$(5.10) \quad \langle e, e^i \rangle = \left\langle \frac{e^1 - s}{2u_n}, e^i \right\rangle = 0 \quad \forall i \geq 2.$$

Let ϱ be defined by (2.2). Since $e^1 \in \text{bd}(\mathcal{K})$, it follows that

$$0 = \varrho(e^1) = \sqrt{1 - \langle e^1, e \rangle^2} - \langle e^1, e \rangle,$$

which implies that

$$(5.11) \quad \langle e^1, e \rangle = \frac{\sqrt{2}}{2}.$$

Each $\hat{v}$ in $M + b$ uniquely corresponds to an element $v \in \mathbb{R}^m$ such that $\hat{v} = \sum_{i=1}^m v_i e^i + b$, where $v = (v_1, \dots, v_m)$. Define Θ by (5.1) with $w^i := e^i \forall i$, namely,

$$(5.12) \quad \Theta(v) = \hat{v} = \sum_{i=1}^m v_i e^i + b \quad \forall v = (x_1, \dots, v_m) \in \mathbb{R}^m.$$

Let g be as in Lemma 5.1. By (5.5) and since $\{b, e^1, \dots, e^m\}$ is orthogonal, one has that, for each $v \in \mathbb{R}^m$, $\hat{v} \in M + b$ such that $\hat{v} = \Theta(v)$,

$$(5.13) \quad g(v) = \varrho(\hat{v}) = \sqrt{\left(\sum_{i=1}^m v_i^2 + \|b\|^2 \right) - \left(\sum_{i=1}^m v_i \langle e^i, e \rangle + \langle b, e \rangle \right)^2} - \left(\sum_{i=1}^m v_i \langle e^i, e \rangle + \langle b, e \rangle \right).$$

By (5.10) and (5.11), it follows that

$$(5.14) \quad \begin{aligned} g(v) &= \sqrt{\left(\sum_{i=1}^m v_i^2 + \|b\|^2 \right) - \left(\frac{1}{\sqrt{2}} v_1 + \langle b, e \rangle \right)^2} - \left(\frac{1}{\sqrt{2}} v_1 + \langle b, e \rangle \right) \\ &= \sqrt{\frac{1}{2} v_1^2 + \sum_{i=2}^m v_i^2 + \|b\|^2 - \sqrt{2} v_1 \langle b, e \rangle - \langle b, e \rangle^2} - \frac{1}{\sqrt{2}} v_1 - \langle b, e \rangle \\ &= \sqrt{\left(\frac{1}{\sqrt{2}} v_1 - \langle b, e \rangle \right)^2 + \sum_{i=2}^m v_i^2 + \eta} - \left(\frac{1}{\sqrt{2}} v_1 - \langle b, e \rangle \right) - \nu, \end{aligned}$$

where $\eta := \|b\|^2 - 2\langle b, e \rangle^2$ and $\nu := 2\langle b, e \rangle$. By performing a homeomorphic affine transform (if necessary), the function g can be deemed of the type considered in Theorem 4.5. Hence the conclusions of that theorem are enjoyed by this g , and thus by φ thanks to Lemma 5.1(iii). Thus φ has a Robinson–Ursescu type error bound but does not have a Υ -error bound for any $\Upsilon \in \Gamma$. $\square$

THEOREM 5.2. Suppose that (5.4) holds and that

$$(5.15) \quad b \notin M, \quad M \cap \mathcal{K} = \{0\}, \quad \text{and} \quad (M + b) \cap \text{int}(\mathcal{K}) = \emptyset.$$

Then $(M + b) \cap \mathcal{K}$ is a singleton and the following assertions hold:

(i) There exists $\tau > 0$ such that

$$(5.16) \quad \text{dist}(x, S_\varphi) \leq \tau([\varphi(x)]_+^{1/2} + [\varphi(x)]_+) \quad \forall x \in X.$$

(ii) There exists $\kappa > 0$ such that

$$(5.17) \quad \text{dist}(y, S_\psi) \leq \kappa([\psi(y)]_+^{1/2} + [\psi(y)]_+) \quad \forall y \in M + b,$$

where ψ is defined by (5.3).

(iii) φ and ψ have neither a Lipschitz error bound nor a Hölder error bound.

Proof. Let $\{w^i\}_{i=1}^m$ be an orthonormal basis of M , and let Θ and g be defined as in Lemma 5.1. When we combine that lemma and Theorem 3.1, the functions φ , ψ , and g have the equivalent error bound properties in the sense that one of the functions has an error bound of any type ((a), (b), (c), or (d) from Definition 2.1) if and only if the other two functions have the corresponding one. On the other hand, by (5.4) and (5.15), $(M + b) \cap \mathcal{K}$ is a singleton and so is the set S_ψ . Since $S_\psi = \Theta S_g$ and Θ is one-to-one, S_g is also a singleton. In view of Lemma 5.1(iv), one can apply Corollary 4.2 to conclude that g enjoys the conclusions of Theorem 4.4. With the help of Lemma 5.1, now it is a routine matter to verify the results. For instance, (5.17) follows from (4.14) and Lemma 5.1(i), (ii). $\square$

Acknowledgments. We thank Professor J.-S. Pang for helpful comments and valuable suggestions. We also thank the two referees, whose suggestions led to many improvements in our presentation and shortened proofs of many results. One referee pointed out that, for the convex case, one can use notions of recession cones/functions and apply some known error bound results such as Deng's theorem ([4, Thm. 2.3]; see also the recent book by Zălinescu [19], especially section 3.10) to provide different proofs of Theorem 4.1 and parts of Theorem 4.3.

REFERENCES

- [1] H. H. BAUSCHKE, *Projection Algorithms and Monotone Operators*, Ph.D. thesis, Simon Fraser University, Department of Mathematics, Burnaby, British Columbia, Canada, 1996. Available online at <http://www.cecm.sfu.ca/preprints/1996pp.html>
- [2] H. H. BAUSCHKE AND J. M. BORWEIN, *On projection algorithms for solving convex feasibility problems*, SIAM Rev., 38 (1996), pp. 367–426.
- [3] J. V. BURKE AND P. TSENG, *A unified analysis of Hoffman's bound via Fenchel duality*, SIAM J. Optim., 6 (1996), pp. 265–282.
- [4] S. DENG, *Computable error bounds for convex inequality systems in reflexive Banach spaces*, SIAM J. Optim., 7 (1997), pp. 274–279.
- [5] M. FUKUSHIMA, Z.-Q. LUO, AND P. TSENG, *Smoother functions for second-order-cone complementarity problems*, SIAM J. Optim., 12 (2001), pp. 436–460.
- [6] J.-B. HIRIART-URRUTY AND C. LEMARÉCHAL, *Convex Analysis and Minimization Algorithms, I: Fundamentals*, Springer-Verlag, New York, 1993.
- [7] A. J. HOFFMAN, *On approximate solutions of systems of linear inequalities*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 263–265.
- [8] R. B. HOLMES, *Geometric Functional Analysis and Its Applications*, Graduate Texts in Math. 24, Springer-Verlag, New York, Heidelberg, 1975.
- [9] A. LEWIS AND J.-S. PANG, *Error bounds for convex inequality systems*, in Generalized Convexity, Generalized Monotonicity: Recent Results, Proceedings of the Fifth Symposium on Generalized Convexity, Luminy, June 1996, J.-P. Crouzeix, J.-E. Martinez-Legaz, and M. Volle, eds., Kluwer Academic Publishers, Dordrecht, 1997, pp. 75–100.

- [10] M. S. LOBO, L. VANDENBERGHE, S. BOYD, AND H. LEBRET, *Applications of second-order cone programming*, Linear Algebra Appl., 284 (1998), pp. 193–228.
- [11] Z. Q. LUO AND J. F. STURM, *Error bounds for quadratic inequality systems*, in High Performance Optimization, H. Frenk et al., eds., Kluwer Academic Publishers, Dordrecht, 2000, pp. 383–404.
- [12] Z. Q. LUO AND J. F. STURM, *Error Analysis*, in Handbook of Semidefinite Programming. Theory, Algorithms and Applications, H. Wolkowicz, R. Saigal, and L. Vandenberg, eds., Kluwer Academic Publishers, Dordrecht, 2000, pp. 163–189.
- [13] K. F. NG AND X. Y. ZHENG, *Error bounds for lower semicontinuous functions in normed spaces*, SIAM J. Optim., 12 (2001), pp. 1–17.
- [14] K. F. NG AND X. Y. ZHENG, *Global weak sharp minima on Banach spaces*, SIAM J. Control Optim., 41 (2003), pp. 1868–1885.
- [15] K. F. NG AND W. H. YANG, *Error bounds for abstract linear inequality systems*, SIAM J. Optim., 13 (2002), pp. 24–43.
- [16] K. F. NG AND W. H. YANG, *Regularities and their relations to error bounds*, Math. Program. Ser. A, 99 (2004), pp. 521–538.
- [17] J.-S. PANG, *Error bounds in mathematical programming*, Math. Program., 79 (1997), pp. 299–332.
- [18] R. T. ROCKAFELLAR AND R. J.-B. WETS, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
- [19] C. ZĂLINESCU, *Convex Analysis in General Vector Spaces*, World Scientific, Singapore, 2002.

AN ALGORITHM MODEL FOR MIXED VARIABLE PROGRAMMING*

S. LUCIDI[†], V. PICCIALLI[†], AND M. SCIANDRONE[‡]

Abstract. In this paper we consider a particular class of nonlinear optimization problems involving both continuous and discrete variables. The distinguishing feature of this class of nonlinear mixed variable optimization problems is that the structure and the number of variables of the problem depend on the values of some discrete variables. In particular, we define a general algorithm model for the solution of this class of problems, that draws inspiration from the approach recently proposed by Audet and Dennis [*SIAM J. Optim.*, 11 (2001), pp. 573–594], and is based on the strategy of combining in a suitable way a local search with respect to the continuous variables and a local search with respect to the discrete variables. We prove global convergence of the algorithm model without specifying the local continuous search, but only identifying some reasonable requirements. Moreover, we define a particular derivative-free algorithm for solving mixed variable programming problems where the continuous variables are linearly constrained and derivative information is not available. Finally, we report numerical results obtained by the proposed algorithm in solving a real optimal design problem. These results show the effectiveness of the approach.

Key words. nonlinear optimization, mixed variable programming, derivative-free methods

AMS subject classifications. 65K05, 90C30, 90C56, 90C11

DOI. 10.1137/S1052623403429573

1. Introduction. Optimization problems involving both continuous and discrete variables are able to describe many real world problems. A particular class of such mixed variable optimization problems, which is important but difficult to solve, is characterized by the following distinguishing features:

- (i) the problem involves a special kind of discrete variables (called categorical variables), which identify an element of an unordered set (for example, colors, shapes, or materials) and affect the structure of the optimization problem. These categorical variables can be represented as a set of discrete numbers, but they cannot assume intermediate values since for such values the objective and/or the constraints functions cannot be evaluated. This implies that in a minimization procedure their discreteness must be always satisfied;
- (ii) the dimensions of the problem are not fixed and are themselves decision variables. In particular, they can be represented by a vector of integer variables (called dimensional variables). Each value of these variables determines the number of the other variables, the number of constraints and the structure of the problem. Also in this case the discreteness of these variables cannot be relaxed. Moreover, the presence of these dimensional variables complicates considerably the analysis and the solution of the optimization problem;
- (iii) the objective function and/or the constraints do not satisfy any convexity assumption. This makes the minimization process difficult and implies, among

*Received by the editors June 6, 2003; accepted for publication (in revised form) August 31, 2004; published electronically July 26, 2005. This work was supported by MIUR, FIRB Research Program *Large-Scale Nonlinear Optimization*, Rome, Italy.

<http://www.siam.org/journals/siopt/15-4/42957.html>

[†]Dipartimento di Informatica e Sistemistica “A. Ruberti,” Università degli Studi di Roma “La Sapienza,” via Buonarroti 12, 00185 Roma, Italy (lucidi@dis.uniroma1.it, piccialli@dis.uniroma1.it).

[‡]Istituto di Analisi dei Sistemi ed Informatica, Consiglio Nazionale delle Ricerche, viale Manzoni 30, 00185 Roma, Italy (sciandro@iasi.rm.cnr.it).

other things, that an efficient bounding technique for fixed values of the integer variables cannot be defined.

In order to formally describe the considered problem, we introduce the vector of decision variables, which has the form

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix},$$

where $z \in \mathcal{Z}^{n_z}$ is the vector of dimensional variables, $y \in \mathcal{Z}^{n_y(z)}$ is the vector of discrete variables (to include categorical), and $x \in \mathbb{R}^{n_x(z)}$ is the vector of continuous variables.

Then we consider problems with the following mathematical formulation:

$$\begin{aligned} (1) \quad & \min f(x, y, z), \\ & z \in \mathcal{F}_z, \\ & y \in \mathcal{F}_y(z), \\ & x \in \mathcal{F}_x(y, z), \end{aligned}$$

where $\mathcal{F}_z \subseteq \mathcal{Z}^{n_z}$, $\mathcal{F}_y(z) \subseteq \mathcal{Z}^{n_y(z)}$, $\mathcal{F}_x(y, z) \subseteq \mathbb{R}^{n_x(z)}$, and $f(\cdot, z) : \mathbb{R}^{n_x(z)} \times \mathcal{Z}^{n_y(z)} \rightarrow \mathbb{R}$. We note that the dimensional variables determine both the dimensions of the other variables of the problem and the structure of the feasible set of the discrete variables y , while the feasible set of the continuous variables depends both on the dimensional variables z and on the discrete variables y .

Standard solution approaches for mixed variable optimization problems are not able to tackle efficiently problem (1), which presents features (i)–(iii). Therefore, it is worthwhile to study and define new solution methods for this class of problems.

Recently, a globally convergent algorithm for a class of problems with similar features has been proposed in [2] and then further developed in [1]. The method described in [2] considers the special case where the domain of the continuous variables $\mathcal{F}_x$ is bound constrained, and extends the pattern search approach introduced in [7]. In particular, the idea is to generate a sequence of feasible iterates with nonincreasing objective value. At any iteration, the objective function is evaluated at a finite number of points on a mesh in order to identify a point with objective value lower than the current solution. The mesh is constructed in such a way as to take into account the variation of both continuous and discrete variables, and the exploration of the mesh (constructed around the current point) is performed in three phases. First, the discrete variables are fixed and a pattern-search-type local search is performed. Second, some “discrete neighbors” of the current point are evaluated. Third, a further exploration with respect to the continuous variables is conducted around a subset of the “discrete neighbors” which are considered promising on the basis of their objective value. In [1] the approach is extended to the case of nonsmooth functions and more general feasible sets.

The method proposed in [2] can be reinterpreted as a suitable combination of a local continuous search (the pattern search algorithm) and a local search with respect to the discrete variables. Our purpose has been to investigate the possibility to embed different continuous minimization techniques in a mixed variable minimization algorithm. In particular, we are interested in algorithms derived from the linesearch approach, where a sufficient decrease of the objective function (instead of a simple

decrease) is required, but the trial points are not restricted to lie on a mesh. This latter feature makes the approach more modular and flexible, in the sense that different continuous and discrete local searches can be used.

In this paper we introduce a new algorithm model for solving the entire class of problems described by problem (1). The global convergence properties of this algorithm model are proved without specifying the local continuous search, but instead by identifying some reasonable requirements that it must satisfy.

Furthermore, it does not refer to a particular structure of the feasible sets and it handles explicitly the presence of dimensional variables. The proposed algorithm model can be used as a basis for developing new algorithms which exploit the structure of the considered instance of problem (1).

In particular, as an example, in this paper we use this algorithm model to propose a derivative-free algorithm to solve problem (1) when the continuous variables are linearly constrained and derivative information is not available. Moreover, the same algorithm model is used in [11] as a framework to define derivative-based algorithms for solving problem (1) in the case where the continuous variables are unconstrained and their number is very large.

The paper is organized as follows. In section 2 we give some definitions and assumptions. In section 3 we introduce, as an illustrative example, a mixed variable optimization problem concerning the design of a magnetic resonance device. We describe our algorithm in section 4 and study its properties in section 5. In section 6 we propose a derivative-free algorithm well suited to solve the application of section 3, namely, one that is able to tackle problem (1) in the case where the continuous variables are linearly constrained and derivative information is not available. Finally, in section 7 we report the numerical results obtained by applying the algorithm described in section 6 on the optimal design application.

In what follows, we will denote by $\|x\|$ the standard Euclidean norm of x and by $\mathcal{B}(x, \epsilon)$ the ball of radius ϵ centered at x .

2. Definitions and assumptions. We start by recalling the definitions of global and local minima of problem (1).

Obviously, the global optimum of problem (1) is a point which satisfies the following definition.

DEFINITION 2.1. *A feasible point (x^*, y^*, z^*) is said to be a global minimizer of problem (1) if*

$$(2) \quad f(x^*, y^*, z^*) \leq f(x, y, z) \quad \forall z \in \mathcal{F}_z, \forall y \in \mathcal{F}_y(z), \forall x \in \mathcal{F}_x(y, z).$$

Less immediate is the definition of local minimum point. This notion refers to the behavior of the objective function in a “suitable neighborhood” of a given point. While a neighborhood of a continuous variable is well represented by a continuous ball, the neighborhood of a discrete variable must be defined taking into account the structure of the particular problem. Furthermore, the discrete neighborhood must represent the fact that variations of the discrete variables can imply also variations of the continuous variables (see for example next section).

Following Audet and Dennis [2], we can characterize a local solution (x^*, y^*, z^*) of problem (1) as a point such that there are no better feasible solutions in the balls centered at the points belonging to the discrete neighborhood of (x^*, y^*, z^*) .

DEFINITION 2.2. *A feasible point (x^*, y^*, z^*) is said to be a local minimizer of problem (1) with respect to the feasible discrete neighborhood $\mathcal{N}(x^*, y^*, z^*)$ if there*

exists an $\epsilon > 0$ such that for all $(\hat{x}, \hat{y}, \hat{z}) \in \mathcal{N}(x^*, y^*, z^*)$

$$(3) \quad f(x^*, y^*, z^*) \leq f(x, \hat{y}, \hat{z}) \quad \forall x \in \mathcal{B}(\hat{x}, \epsilon) \cap \mathcal{F}_x(\hat{y}, \hat{z}),$$

where $\mathcal{N}(x^*, y^*, z^*)$ is a finite set of feasible points.

This definition depends on the choice of the discrete neighborhoods. In fact, a larger discrete neighborhood $\mathcal{N}(x^*, y^*, z^*)$ should give a better local minimizer, but this may increase the computational effort needed to locate the solution.

As is common in nonlinear programming algorithms, the cost of locating a global or local solution of problem (1) can be prohibitive. In practice, it is more reasonable to determine a point (usually called stationary point) which satisfies suitable necessary optimality conditions.

The definition of stationary points of problem (1) relies on the definition of stationary points of the continuous problem formed by fixing the discrete variables in problem (1); namely, for fixed $\bar{z} \in \mathcal{F}_z$ and $\bar{y} \in \mathcal{F}_y(\bar{z})$,

$$(4) \quad \begin{aligned} &\min f(x, \bar{y}, \bar{z}), \\ &x \in \mathcal{F}_x(\bar{y}, \bar{z}). \end{aligned}$$

DEFINITION 2.3. A feasible point (x^*, y^*, z^*) is said to be a stationary point of problem (1) with respect to the feasible discrete neighborhood $\mathcal{N}(x^*, y^*, z^*)$ if

(i) the point x^* is a stationary point of the continuous problem

$$(5) \quad \begin{aligned} &\min f(x, y^*, z^*), \\ &x \in \mathcal{F}_x(y^*, z^*); \end{aligned}$$

(ii) every $(\hat{x}, \hat{y}, \hat{z}) \in \mathcal{N}(x^*, y^*, z^*)$ satisfies $f(\hat{x}, \hat{y}, \hat{z}) \geq f(x^*, y^*, z^*)$;

(iii) for every $(\hat{x}, \hat{y}, \hat{z}) \in \mathcal{N}(x^*, y^*, z^*)$ such that $f(\hat{x}, \hat{y}, \hat{z}) = f(x^*, y^*, z^*)$, the point $\hat{x}$ is a stationary point of the continuous problem

$$(6) \quad \begin{aligned} &\min f(x, \hat{y}, \hat{z}), \\ &x \in \mathcal{F}_x(\hat{y}, \hat{z}). \end{aligned}$$

Note that in the above definition since we want to treat problem (1) in its general form, we generically refer to stationary points of problem (4), namely, to points satisfying suitable optimality conditions.

Remark. It is easy to prove (see [12]) that a local minimum point of problem (1) is a stationary point of problem (1) according to Definition 2.3. Moreover, if any stationary point of problem (4) is also a local minimum point of problem (4), then any stationary point of problem (1) is also a local minimum point of problem (1).

Any minimization algorithm for solving problem (1) produces vectors with both continuous and discrete components which can have different dimension at each iteration. For this reason we need to specify the notion of convergence and accumulation point in a mixed variable domain.

DEFINITION 2.4. A sequence $\{x_k, y_k, z_k\}$ converges to a point $(\bar{x}, \bar{y}, \bar{z})$ if for any $\epsilon > 0$ there exists an index k_ϵ such that for all $k \geq k_\epsilon$ we have that $y_k = \bar{y}$, $z_k = \bar{z}$ and $\|x_k - \bar{x}\| < \epsilon$.

DEFINITION 2.5. A point $(\bar{x}, \bar{y}, \bar{z})$ is an accumulation point of the sequence $\{x_k, y_k, z_k\}$ if there exists an infinite subset $K \subseteq \{0, 1, \dots\}$ such that the subsequence $\{x_k, y_k, z_k\}_K$ converges to $(\bar{x}, \bar{y}, \bar{z})$.

We note that similar definitions have been introduced in [1]. In what follows we suppose that the following assumptions are verified.

Assumptions

A1. The set $\mathcal{F}_z$ contains a finite number of elements.

A2. Given $z_0 \in \mathcal{F}_z$, $y_0 \in \mathcal{F}_y(z_0)$, $x_0 \in \mathcal{F}_x(y_0, z_0)$, and a scalar $\xi > 0$, for each $z \in \mathcal{F}_z$, the level set

$$(7) \quad \mathcal{L}_z(x_0, y_0, z_0) = \{(x, y) \in \mathcal{F}_x(y, z) \times \mathcal{F}_y(z) : f(x, y, z) \leq f(x_0, y_0, z_0) + \xi, \xi > 0\}$$

is compact.

A3. Given $z_0 \in \mathcal{F}_z$, $y_0 \in \mathcal{F}_y(z_0)$, $x_0 \in \mathcal{F}_x(y_0, z_0)$, for each $z \in \mathcal{F}_z$ and $y \in \mathcal{F}_y(z)$, the function $f(\cdot, y, z)$ is continuous in a neighborhood of $\{x : (x, y) \in \mathcal{L}_z(x_0, y_0, z_0)\}$.

A4. Let $\{x_k, y_k, z_k\}$ be a sequence converging to $(\bar{x}, \bar{y}, \bar{z})$. Then, for any $(\hat{x}, \hat{y}, \hat{z}) \in \mathcal{N}(\bar{x}, \bar{y}, \bar{z})$, there exists a sequence $\{(\hat{x}_k, \hat{y}_k, \hat{z}_k)\}$ converging to $(\hat{x}, \hat{y}, \hat{z})$ such that $(\hat{x}_k, \hat{y}_k, \hat{z}_k) \in \mathcal{N}(x_k, y_k, z_k)$.

Assumption A1 is a reasonable assumption and requires that the vector of dimensional variables can assume a finite number of values. Assumption A2 requires the standard compactness of the level sets of the objective function over the feasible set. The positive scalar ξ which appears in Assumption A2 will be used in the algorithm model defined later. Assumption A3 is a minimal smoothness requirement on the objective function. Assumptions A1–A3 ensure the existence of a solution of the problem, as shown in the next theorem. Assumption A4 is a mild continuity assumption on the discrete neighborhoods and is equivalent to the lower semicontinuity of a point-to-set function as defined in [4].

THEOREM 2.6. *Under Assumptions A1–A3 the objective function $f(x, y, z)$ admits a global minimizer on the feasible set.*

Proof. By Assumption A1 the vector z belongs to the finite set $\{z^1, \dots, z^m\}$. Given a feasible point (x_0, y_0, z_0) , for each feasible vector z^i , $i = 1, \dots, m$, consider the level set $\mathcal{L}_{z^i}(x_0, y_0, z_0)$. By Assumption A2, this set is compact and hence the vector of discrete variables y can assume a finite number of values on it, namely, $y \in \{y_1^i, \dots, y_p^i\}$, where the superscript i means that these values depend on the vector z^i . Now, for a fixed value y_j^i of the vector y , consider the set

$$\mathcal{L}_{z^i}^{y_j^i}(x_0, y_0, z_0) = \{x \in \mathcal{F}_x(y_j^i, z^i) : f(x, y_j^i, z^i) \leq f(x_0, y_0, z_0) + \xi, \xi > 0\}.$$

This set is compact since it is a closed subset of the set $\mathcal{L}_{z^i}(x_0, y_0, z_0)$, which is compact by Assumption A2. Moreover, by Assumption A3, $f(x, y_j^i, z^i)$ is continuous on $\mathcal{L}_{z^i}^{y_j^i}(x_0, y_0, z_0)$, and hence it admits a global minimum point $(x_{i,j}^*, y_j^i, z^i)$ on this set. Thus we can conclude that the objective function $f(x, y, z)$ admits a global minimum point on the feasible set, given by

$$(x^*, y^*, z^*) = \arg \min_{\substack{i=1, \dots, m \\ j=1, \dots, p}} f(x_{i,j}^*, y_j^i, z^i). \quad \square$$

3. An illustrative application. In order to illustrate our approach, we consider an application in which the aim is to find the optimal design of a small magnetic resonance device (see [12] for the details) with the following features:

- a high uniformity of the magnetic field, to obtain high resolution images,
- a large uniformity region, to be able to scan an area which is large with respect to the dimension of the apparatus,
- low weight, to make the apparatus transportable.

The realization of such an apparatus would greatly benefit the diagnosis and prognosis of many pathologies. One possible technique for building a low field dedicated magnet makes use of permanent magnets surrounded by an iron yoke to amplify the magnetic field.

We consider a model of the multipolar magnet, which is cylindrical and consists of elliptical rings having a certain number of small magnets screwed onto each of them.

What characterizes the structure is the number of rings and the number of magnets screwed onto each ring. These two elements are the dimensional variables of the corresponding mixed variable programming problem.

For practical considerations, we consider only the case in which all rings are equal with respect to the number and positioning of the magnets, and all magnets are equal as well. While this may seem to restrict the flexibility of the design, it actually eases the industrial building process. Each ring is 30 mm wide and 40 mm thick (like the one depicted in Figure 1). The semiaxes of the rings are not fixed, and each ring can have different semiaxes with values belonging to two fixed intervals centered in two nominal values $a = 236$ mm and $b = 180$ mm. For each ring there are two fixed magnets directed along the short semiaxis, whose role is to strengthen the magnetic field without worsening the uniformity. The magnets have a cylindrical shape, with a height of 30 mm (as shown in Figure 2), and a residual magnetization of $M_r = 1.35$ Tesla. Their characteristics are therefore fixed, except for the radius r , which is a variable of the optimization problem.

Moreover, a simplifying assumption is made on the whole structure. In particular, we assume that the whole structure is symmetric with respect to the semiaxes of the rings and with respect to a plane parallel to the rings which divides the model into two halves (see Figure 3, where an example of half multipolar magnet is depicted with six rings and eighteen magnets), so that we can consider only one eighth of the multipolar magnet and get the other parts by reflection. This implies that we consider as dimensional variables the number of rings nr which constitute half magnet and the number of magnets nm which are screwed on a quarter of rings. Let XYZ be a system of Cartesian coordinates with origin at the center of the structure, X -axis parallel to the cylinder axis, and Z - and Y -axes directed, respectively, along the shortest and longest semiaxis of the elliptical base. The continuous variables of the problem are the position d_i of the i th ring along the X -axis (see Figure 4, point (a)), the angular position φ_j of the magnet j , which is the same on each ring (see Figure 4, point (c)), one variable b_i for each ring, which represents the difference between the effective length of the two semiaxes of the i th ring, and two nominal values a and b (see Figure 4, point (b)).

As already stated, the key feature of every magnetic resonance apparatus is the uniformity of the magnetic field within a specified region in space. In our case, this so-called target region is a cylinder (12 cm long) having an elliptical base with major and minor semiaxes 6 cm and 5 cm long, respectively. These dimensions (which identify a large region with respect to the dimension of the multipolar magnet itself) have been

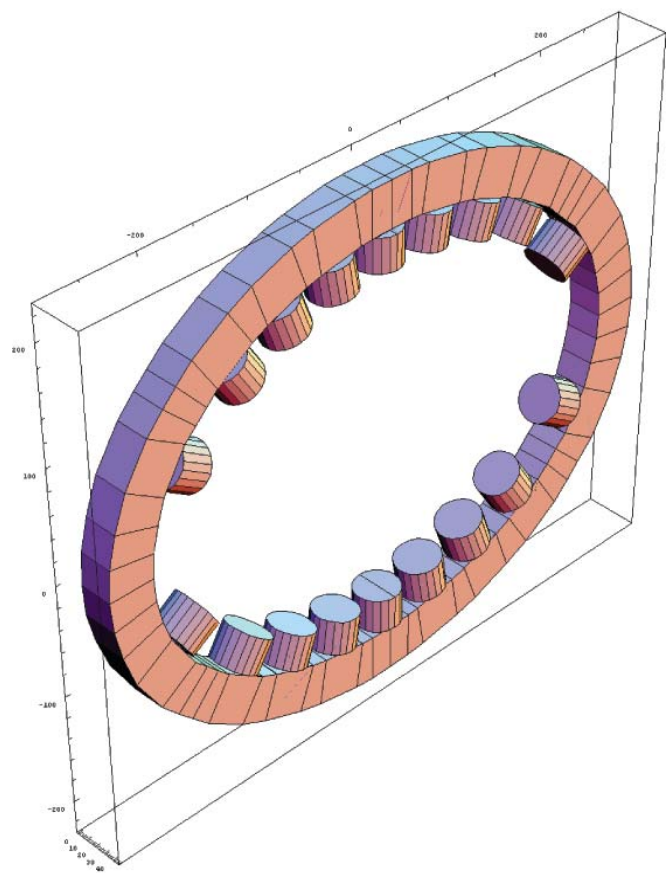


FIG. 1. *Example of a ring.*

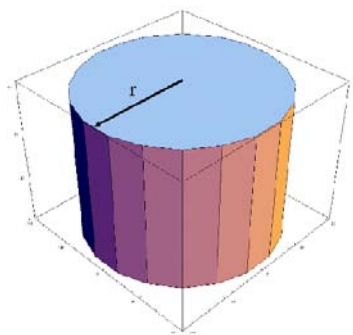
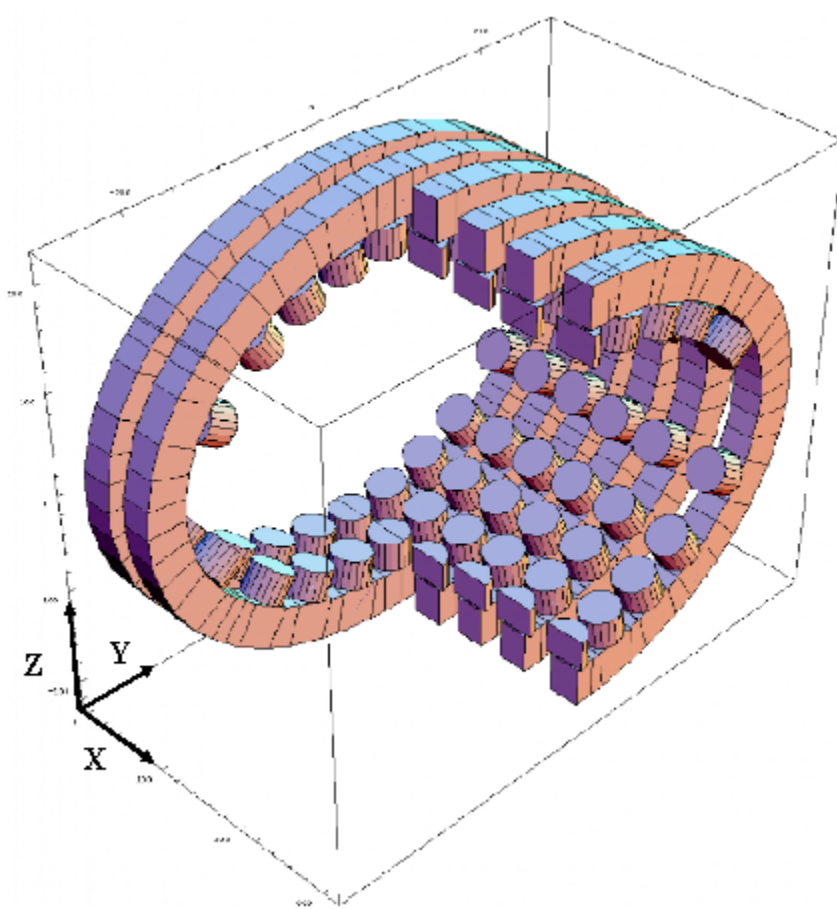


FIG. 2. *Small magnet.*

FIG. 3. *Example of a half magnet.*

established by magnetic resonance experts and are the same of standard commercial apparatus. To summarize, we consider as optimization variables:

- the number nr of rings which constitute the half structure,
- the number of magnets nm on each quarter of a ring (all rings have the same number of magnets) which does not include the magnets fixed on each ring,
- the position d_i of the i th ring of the half structure along the X -axis,
- the angular position φ_j of the j th magnet on the quarter of a ring, which is the same on each ring,
- one variable b_i for each ring, which represents the difference between the effective length of the two semiaxes of the i th ring and two nominal values a and b ,
- the radius r of the magnets.

Thus, in the formulation of the optimization problem, the discrete dimensional variables are the number of rings nr and the number of magnets nm ; the continuous variables are the positions d_i and offsets b_i of the rings, $i = 1, \dots, nr$, and the angular positions φ_j of the magnets, for $j = 1, \dots, nm$.

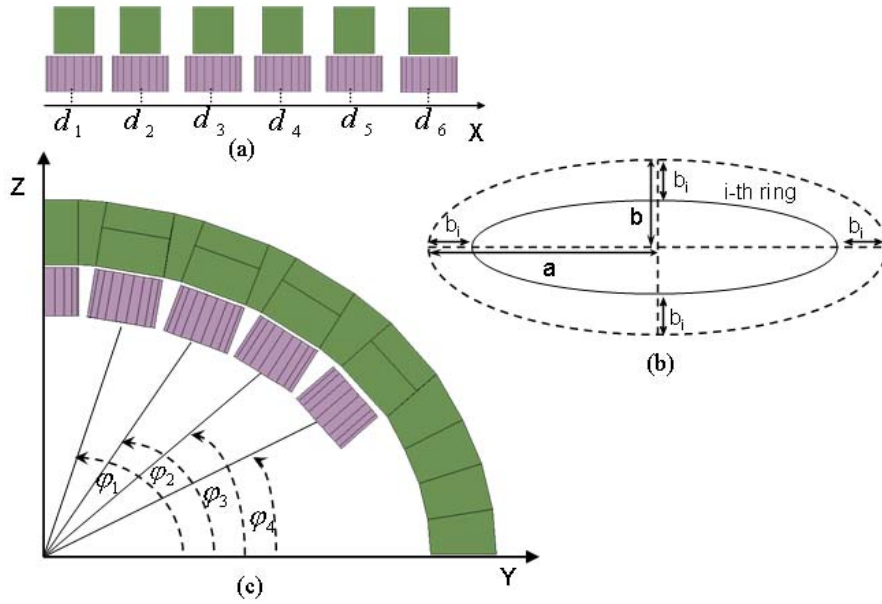


FIG. 4. Variables of the problem with 6 rings and 4 magnets.

Finally, since the small magnets commercially available belong to a finite list, their radius r can assume only some integer values in a fixed range and hence is a discrete variable.

We note that the dimension of the continuous variables depends on the dimensional variables. Summarizing and using the notation introduced in the previous section, the decision variables of the problem are the following:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix},$$

where

$$x = (d_1, \dots, d_{nr}, b_1, \dots, b_{nr}, \varphi_1, \dots, \varphi_{nm})^T, \quad y = (r), \quad \text{and} \quad z = (nr, nm)^T$$

with $x \in \mathbb{R}^{2nr+nm}$, $y \in [l_y, u_y] \cap \mathcal{Z}$, and $z \in [l_z, u_z] \cap \mathcal{Z}^2$.

The objective function to be minimized must be a measure of the nonuniformity of the magnetic field, since our aim is to obtain that the magnetic field $\vec{B}$ generated by the multipolar magnet be as uniform as possible within the target region and directed along the Z -axis. Let N_p be the number of points uniformly distributed on a grid inside the cylindrical target region of interest. Let $B_X^{(i)}(x, y, z)$, $B_Y^{(i)}(x, y, z)$, and $B_Z^{(i)}(x, y, z)$ be the three components of the magnetic field measured at the i th point

of the grid. The objective function is

$$f(x, y, z) = \frac{\sum_{i=1}^{N_p} \left[(B_Z^{(i)}(x, y, z) - \bar{B}_Z(x, y, z))^2 + B_X^{(i)}(x, y, z)^2 + B_Y^{(i)}(x, y, z)^2 \right]}{\bar{B}_Z(x, y, z)^2},$$

where $\bar{B}_z(x, y, z)$ is the average Z -component of the magnetic field probed on the grid points. It can be easily verified that the objective function $f(x, y, z)$ attains its global minimum of 0 when $B_Z^{(i)}(x, y, z) = \bar{B}_Z(x, y, z)$, $B_X^{(i)}(x, y, z) = 0$, and $B_Y^{(i)}(x, y, z) = 0$, $i = 1, \dots, N_p$; that is, when the magnetic field is uniform and directed along the Z -axis on every point of the control grid.

In spite of its simple structure, and due to the high nonlinearities introduced by the iron yoke, the field generated by the multipolar magnet model cannot be analytically computed. For this reason, the multipolar magnet model, corresponding to a feasible choice of the above-mentioned variables, is computer rendered within a field simulation program that predicts the behavior of the magnetic field by means of a sophisticated finite element analysis. In particular, we employ RADIA (see [6, 5]) a reliable finite element simulator released by the European Synchrotron Radiation Facility (ESRF) at Grenoble (France). The simulation process carried out by this code can take up to a minute of CPU time on a Pentium IV processor based computer.

The constraints can be divided into four groups.

The first one consists of the following box constraints on the variables:

$$\begin{aligned} 0 &\leq d_i \leq 300, \quad i = 1, \dots, nr, \\ 0 &\leq \varphi_j \leq 90, \quad j = 1, \dots, nm, \\ -10 &\leq b_i \leq 20, \quad i = 1, \dots, nr. \end{aligned}$$

The second group ensures that the rings are separated from one another; namely, the following inequalities must be satisfied:

$$(8) \quad \begin{aligned} -d_1 + r &\leq -\epsilon \\ d_{i-1} - d_i + 2r &\leq -\epsilon, \quad i = 2, \dots, nr, \end{aligned}$$

where r is the magnet radius. The third group of constraints ensures that the ordering of the angular positions of the magnets is maintained; namely,

$$(9) \quad \varphi_{i+1} - \varphi_i \leq -\epsilon, \quad i = 1, \dots, nm.$$

Finally, there is a group of constraints which are meant to avoid magnets overlapping. These constraints are continuously differentiable but nonconvex and their description can be found in [12]. In our computational experiments, we have substituted these nonlinear constraints with simpler but more restrictive box constraints.

For the considered problem, we define the discrete neighborhood of a given point as follows:

- the two points obtained, respectively, by increasing and decreasing by 1 mm the radius r of the magnets, while keeping fixed the number of rings nr and the number of magnets nm ;
- the two points obtained, respectively, by adding ($nr := nr + 1$) and deleting ($nr := nr - 1$) one ring, while keeping fixed the number of magnets nm and the radius r of the magnets;
- the two points obtained, respectively, by adding ($nm := nm + 1$) and deleting ($nm := nm - 1$) one small magnet, while keeping fixed the number of rings nr and the radius r of the magnets.

We have to note that whenever we obtain a new point by changing the values of any discrete variables, it may also be necessary to change the values of some continuous variables. For example, if we increase by 1 mm the radius r of the magnets, without changing angular positions, the resulting design may have two overlapping magnets, which is infeasible.

4. Algorithm model MIVAM. We now describe our algorithm model for solving problem (1). In defining an algorithm model, it is necessary to account for the presence of both continuous and discrete variables, which need different minimization procedures. This algorithm is based on the idea to alternate two phases:

- an attempt to update the continuous variables by using a local continuous search (Phase 1),
- an attempt to update the discrete variables by using a local search in the discrete neighborhood of the current point (Phase 2).

Phase 1:

Given the current feasible point (x_k, y_k, z_k) , we keep fixed the discrete variables (y_k, z_k) and consider the following continuous optimization problem:

$$(10) \quad \begin{aligned} \min f(x, y_k, z_k), \\ x \in \mathcal{F}_x(y_k, z_k). \end{aligned}$$

Starting from x_k , we perform a local continuous search (*LCS*) with the goal of finding a new vector $\tilde{x}_k$ which is, roughly speaking, a better approximation of a stationary point of problem (10).

Phase 2:

In this phase we try to update the discrete variables by considering the points belonging to the discrete neighborhood $\mathcal{N}(\tilde{x}_k, y_k, z_k)$ of the point $(\tilde{x}_k, y_k, z_k)$ produced by Phase 1.

First, we simply evaluate the objective function at the points belonging to $\mathcal{N}(\tilde{x}_k, y_k, z_k)$. If one of these points produces a sufficient decrease from $f(\tilde{x}_k, y_k, z_k)$, then it becomes the current point, and a new iteration is performed.

If none of the points belonging to $\mathcal{N}(\tilde{x}_k, y_k, z_k)$ produces a sufficient decrease with respect to $f(\tilde{x}_k, y_k, z_k)$, it does not necessarily mean that the trial points are not promising. For any point in $\mathcal{N}(\tilde{x}_k, y_k, z_k)$ that is not significantly worse in function value than the current point, we can still perform a suitable number of local continuous searches starting with the selected point, with the expectation of finding a new iterate with sufficiently lower function value.

The proposed algorithm model is formally stated as follows.

Mixed integer variable algorithm model (MIVAM)

Data: $z_0 \in \mathcal{F}_z$, $y_0 \in \mathcal{F}_y(z_0)$, $x_0 \in \mathcal{F}_x(y_0, z_0)$, $\xi \geq 0$, $\theta \in (0, 1)$, $\eta_0 > 0$, $\mu^{in} > 0$.

Step 0: Set $k = 0$, $\mu_k^0 = \mu^{in}$.

Step 1: Compute $\tilde{x}_k$ and μ_k by $LCS(x_k, \mu_k^0, \tilde{x}_k, \mu_k, y_k, z_k)$ and set $\mu_{k+1}^0 = \mu_k$.

Step 2: If there exists a $(\hat{x}_k, \hat{y}_k, \hat{z}_k) \in \mathcal{N}(\tilde{x}_k, y_k, z_k)$ such that

$$f(\hat{x}_k, \hat{y}_k, \hat{z}_k) \leq f(\tilde{x}_k, y_k, z_k) - \eta_k,$$

set $x_{k+1} = \hat{x}_k$, $y_{k+1} = \hat{y}_k$, $z_{k+1} = \hat{z}_k$, $\eta_{k+1} = \eta_k$, and go to Step 5.

Step 3: Define $W_k = \{(x, y, z) \in \mathcal{N}(\tilde{x}_k, y_k, z_k) : f(x, y, z) \leq f(\tilde{x}_k, y_k, z_k) + \xi\}$.

3.1: If $W_k \neq \emptyset$, choose $(x', y', z') \in W_k$, set $j = 1$, $x^j = x'$, $\mu^{j-1} = \mu_k$.
Otherwise go to Step 4.

3.2: Compute x^{j+1} and μ^j by $LCS(x^j, \mu^{j-1}, x^{j+1}, \mu^j, y', z')$.

3.3: If $f(x^{j+1}, y', z') \leq f(\tilde{x}_k, y_k, z_k) - \eta_k$, set $x_{k+1} = x^{j+1}$, $y_{k+1} = y'$,
 $z_{k+1} = z'$, $\eta_{k+1} = \eta_k$, and go to Step 5.

3.4: If $\mu^j > \mu_k$, set $j = j + 1$, and go to 3.2.
Otherwise set $W_k = W_k \setminus \{(x', y', z')\}$, and go to 3.1.

Step 4: Set $x_{k+1} = \tilde{x}_k$, $y_{k+1} = y_k$, $z_{k+1} = z_k$. If $f(x_{k+1}, y_{k+1}, z_{k+1}) \leq f(\tilde{x}_k, y_k, z_k) - \eta_k$, set $\eta_{k+1} = \eta_k$. Otherwise set $\eta_{k+1} = \theta\eta_k$.

Step 5: Set $k = k + 1$, and go to Step 1.

At Step 1, Phase 1 is performed by applying the local continuous search procedure $LCS(x_k, \mu_k^0, \tilde{x}_k, \mu_k, y_k, z_k)$. This procedure should guarantee, in the limit, stationarity with respect to problem (10). The procedure LCS uses the scalar μ_k^0 , which is derived from previous iterations and gives, roughly speaking, an initial estimate of the expected improvement of the objective function obtainable from the point x_k . The procedure LCS tries to produce a new point $\tilde{x}_k$, where the objective function is sufficiently decreased. The information obtained during this process can be used to update the estimate μ_k^0 , producing a new scalar μ_k . A suitable choice of the scalar μ_k (and μ_k^0) and its updating rule must imply that the smaller μ_k is, the better x_k approximates a stationary point. In particular, if the procedure LCS is not able to produce a sufficient decrease of the objective function, the point $\tilde{x}_k$ is set equal to x_k , and the scalar μ_k must be set to a value sufficiently smaller than μ_k^0 .

Phase 2 is performed in Steps 2 and 3. In particular, at Step 2 the objective function is evaluated at each point in $\mathcal{N}(\tilde{x}_k, y_k, z_k)$. If one of these points produces a decrease with respect to $f(\tilde{x}_k, y_k, z_k)$ greater than or equal to η_k , then it becomes the current point and a new iteration is performed. Otherwise, the discrete neighborhood

is further investigated in Step 3. In particular, a set $W_k \subseteq \mathcal{N}(\tilde{x}_k, y_k, z_k)$ of points with objective value not significantly worse than $f(\tilde{x}_k, y_k, z_k)$ is selected. Each of these points $(x', y', z') \in W_k$ is considered promising, and the algorithm tries to determine if it is worth replacing (y_k, z_k) with (y', z') . In particular, starting from each point $(x', y', z') \in W_k$, the local continuous search is repeated until

- (a) it finds a significantly better point than $(\tilde{x}_k, y_k, z_k)$, or
- (b) the test at Step 3.4 fails.

In case (a), the new point becomes the current iterate (with new discrete variables (y', z')), and a new iteration is performed. In case (b), we reject the discrete variables (y', z') because a sufficient decrease of the objective function has not been achieved.

At Step 4 the point $(\tilde{x}_k, y_k, z_k)$ becomes the new current point and, if neither the local continuous search nor the discrete search has been able to produce a decrease of the objective function greater than or equal to η_k , then this parameter is reduced before starting the next iteration.

In order to complete the description of the algorithm, we should specify the procedure *LCS*, but, as already said in the introduction, we prefer to identify only those properties it must satisfy to ensure the global convergence of this algorithm model.

In particular, we require that procedure $LCS(x_k, \mu_k^0, \tilde{x}_k, \mu_k, y_k, z_k)$ satisfies the following two properties.

Property A

Given a feasible point (x_h, y_h, z_h) and a scalar $\mu_h^0 > 0$, the local continuous search produces a point $\tilde{x}_h$ and a scalar $\mu_h > 0$ such that

- (i) $\tilde{x}_h \in \mathcal{F}_x(y_h, z_h)$ and $f(\tilde{x}_h, y_h, z_h) \leq f(x_h, y_h, z_h)$,
- (ii) - either $f(\tilde{x}_h, y_h, z_h) \leq f(x_h, y_h, z_h) - \sigma(\mu_h)$
 (where the function $\sigma(\cdot) \geq 0$ is such that if $\sigma(t_h) \rightarrow 0$, then $t_h \rightarrow 0$);
 - or $\mu_h \leq \delta \mu_h^0$, $\delta \in (0, 1)$.

Property B

Let $\{(x_h, y_h, z_h)\}$ be a sequence of feasible points. For all h , suppose *LCS* is applied to (x_h, y_h, z_h) to obtain a point $(\tilde{x}_h, y_h, z_h)$ and a scalar $\mu_h > 0$. If the sequence $\{\mu_h\}$ converges to zero, then for every accumulation point $(\bar{x}, \bar{y}, \bar{z})$ of the sequence $\{(x_h, y_h, z_h)\}$, $\bar{x}$ is a stationary point of the problem

$$(11) \quad \begin{aligned} & \min f(x, \bar{y}, \bar{z}), \\ & x \in \mathcal{F}_x(\bar{y}, \bar{z}). \end{aligned}$$

Properties A and B are the minimal requirements that procedure *LCS*, when embedded in the MIVAM algorithm, must satisfy in order to ensure in the limit the stationarity with respect to the continuous variables. The scalar μ_h can be viewed as a nonstationarity measure of the point x_h . Point (ii) of Property A and Property B imply that there must exist a strict relationship between the obtained decrease of

the objective function and the stationarity of the point x_h . In fact, if procedure *LCS* produces a point $\tilde{x}_h$ different from x_h , Properties A and B imply that the produced point must guarantee a sufficient decrease of the objective function if the point x_h is not a “good” approximation of a stationary point. Moreover, a small decrease of the objective function must indicate that the point x_h is close to a stationary point. Since we want to include the case where the local continuous search does not use derivative information, we must consider also the possibility that it is not able to produce a point $\tilde{x}_h$ different from x_h . In this case, Properties A and B require that a failure of procedure *LCS* must indicate that the point x_h approximates a stationary point better than expected. Therefore procedure *LCS* must produce a parameter μ_h which is strictly smaller than μ_h^0 , where the parameter μ_h^0 is derived from the previous iterations.

We note that our strategy to ensure the convergence is different from the one used in [2] and [1]. In fact, the algorithms proposed in [2] and [1] follow the pattern search approach, where the trial points must lie on a mesh and a new point is accepted whenever there is a simple decrease of the objective function. In our approach, a new point is accepted only when it produces a “sufficient decrease” of the objective function, but this allows us to remove the requirement that the trial points lie on a prefixed mesh. This implies less restrictions both on the sampling technique with respect to the continuous variables and on the definition of the discrete neighborhoods. Moreover, the tests at Steps 2 and 3 cannot be too restrictive, since the parameter η_k is reduced by the algorithm whenever a sufficient decrease is not obtained. On the other hand, a suitable value of η_k should prevent the discrete variables from changing value too early and oscillating between different values.

5. Theoretical analysis. In this section we analyze the theoretical properties of the MIVAM algorithm model.

First, we prove that the algorithm is well posed; namely, that the algorithm cannot cycle at Step 3.

PROPOSITION 5.1. *For each iteration k of the MIVAM algorithm, Step 3 terminates in a finite number of steps.*

Proof. Suppose by contradiction that Steps 3.1–3.4 generate an infinite loop, so that an infinite sequence of points $\{x^j\}$ and an infinite sequence of scalars $\{\mu^j\}$ are produced, and

$$(12) \quad \mu^j > \mu_k \quad \forall j.$$

By Property A of *LCS*, we have two possibilities for each j ; namely,

$$(13) \quad f(x^{j+1}, y, z) - f(x^j, y, z) \leq -\sigma(\mu^j)$$

or

$$(14) \quad \mu^j \leq \delta \mu^{j-1}.$$

Now we split the inner iterations of Step 3 in two subsets, J_1 and J_2 , which contain the iteration indices for which (13) and (14) hold, respectively.

If $j \in J_1$, by (13) we have

$$(15) \quad |f(x^{j+1}, y, z) - f(x^j, y, z)| \geq \sigma(\mu^j).$$

By Property A of procedure *LCS* the sequence $\{f(x^j, y, z)\}$ is nonincreasing. Moreover, it follows from Assumption A2 that the sequence $\{f(x^j, y, z)\}$ is bounded from

below. Therefore, the sequence $\{f(x^j, y, z)\}$ converges. This implies that, if J_1 is infinite, the left-hand side of (13) tends to zero, and hence

$$(16) \quad \lim_{j \rightarrow \infty, j \in J_1} \sigma(\mu^j) = 0.$$

By (ii) of Property A (property of $\sigma(\cdot)$), we have

$$(17) \quad \lim_{j \rightarrow \infty, j \in J_1} \mu^j = 0.$$

We now consider the set J_2 . For each $j \in J_2$, let $m_j \in J_1$ be the largest index satisfying $m_j < j$, with the convention that $m_j = 0$ if J_1 is empty. Then

$$(18) \quad \mu^j \leq (\delta)^{j-m_j} \mu^{m_j}.$$

If J_2 is finite, then J_1 is infinite, and recalling (16), we can write

$$\lim_{j \rightarrow \infty} \sigma(\mu^j) = \lim_{j \rightarrow \infty, j \in J_1} \sigma(\mu^j) = 0,$$

which implies that $\mu^j \rightarrow 0$.

Then assume that J_2 is infinite. As $j \rightarrow \infty$ and $j \in J_2$, either $m_j \rightarrow \infty$ (if J_1 is an infinite subset) or $(j - m_j) \rightarrow \infty$. In the first case, since $\delta \in (0, 1)$, it follows from (17) and (18) that $\mu^j \rightarrow 0$. In the second case, it follows from (18) that $\mu^j \rightarrow 0$.

Thus we have proved that $\mu^j \rightarrow 0$, but this contradicts (12). $\square$

In order to analyze the properties of the algorithm, we need to characterize the outer iterations (k) of the algorithm. First, we partition the set of indices of the iterates into two subsets: the set of *unsuccessful* iterates K_u and the set of *successful* iterates K_s , which is the complement of K_u . An iteration is defined *unsuccessful* when it does not produce a point where the objective function is sufficiently decreased; namely,

$$K_u = \{k : \eta_{k+1} < \eta_k\}.$$

Furthermore, for every $z \in \mathcal{F}_z$, we introduce the following subset of iteration indices:

$$(19) \quad K(z) = \{k \in K_u : z_k = z_{k+1} = z\}.$$

The next proposition describes some properties of the sequences produced by the algorithm, that will be used later in the convergence analysis.

PROPOSITION 5.2. *Let $\{(x_k, y_k, z_k)\}$, $\{\eta_k\}$, $\{\mu_k\}$ be the sequences produced by the algorithm. Then*

- (i) *the sequence of function values $\{f(x_k, y_k, z_k)\}$ is nonincreasing and convergent;*
- (ii) *the sequence $\{(x_k, y_k, z_k)\}$ is bounded;*
- (iii) *the subset of indices of unsuccessful iterates K_u is infinite;*
- (iv) *the sequence $\{\eta_k\}$ is such that $\lim_{k \rightarrow \infty} \eta_k = 0$;*
- (v) *the sequence $\{\mu_k\}$ is such that $\lim_{k \rightarrow \infty} \mu_k = 0$.*

Proof. (i) The instructions of the algorithm and Property A of procedure *LCS* imply that the sequence $\{f(x_k, y_k, z_k)\}$ is nonincreasing, and Theorem 2.6 ensures that $\{f(x_k, y_k, z_k)\}$ is bounded below. Hence, the sequence $\{f(x_k, y_k, z_k)\}$ is convergent.

(ii) By Assumption A1 we have that the vector z_k belongs to the finite set $\{z^1, \dots, z^m\}$. The instructions of the algorithm imply that each point (x_k, y_k, z_k)

belongs to the set $\mathcal{L}_{z_k}(x_0, y_0, z_0)$, which is compact by Assumption A2. Thus the whole sequence $\{(x_k, y_k, z_k)\}$ is contained in the set $\bigcup_{i=1, \dots, m} \mathcal{L}_{z^i}(x_0, y_0, z_0)$, which is the union of a finite number of compact sets and hence is compact. Then we can conclude that the sequence $\{(x_k, y_k, z_k)\}$ is bounded.

(iii) Suppose that K_u is finite. Then there exists $\bar{k} > 0$ such that all iterates satisfying $k > \bar{k}$ are successful. In this case, we have

$$f(x_{k+1}, y_{k+1}, z_{k+1}) \leq f(x_k, y_k, z_k) - \eta_k,$$

and $\eta_{k+1} = \eta_k = \eta$ for all $k \geq \bar{k}$. Since $\eta > 0$, this would mean that $\{f(x_k, y_k, z_k)\}$ diverges to $-\infty$ in the limit, which violates point (i).

(iv) If $k \in K_u$, then $\eta_{k+1} = \theta \eta_k$, $\theta \in (0, 1)$; otherwise, $\eta_{k+1} = \eta_k$. Since K_u is infinite, it follows that $\eta_k \rightarrow 0$ for $k \rightarrow \infty, k \in K_u$, and thus $\eta_k \rightarrow 0$.

(v) Let $\tilde{x}_k$ and μ_k be the point and the scalar produced by procedure *LCS* at Step 1 of iteration k . By Property A of *LCS*, we have two possibilities:

- either

$$(20) \quad f(\tilde{x}_k, y_k, z_k) - f(x_k, y_k, z_k) \leq -\sigma(\mu_k),$$

- or

$$(21) \quad \mu_{k+1}^0 = \mu_k \leq \delta \mu_k^0.$$

In the first case, by the instructions of the algorithm and by (20), since $f(x_{k+1}, y_k, z_k) \leq f(\tilde{x}_k, y_k, z_k)$, we have

$$(22) \quad f(x_{k+1}, y_k, z_k) - f(x_k, y_k, z_k) \leq -\sigma(\mu_k).$$

Now we split the iterations of the algorithm in two subsets, K_a and K_b , which contain the iteration indices for which (20) and (21) hold, respectively. If $k \in K_a$, by (22) we have

$$(23) \quad |f(x_{k+1}, y_k, z_k) - f(x_k, y_k, z_k)| \geq \sigma(\mu_k).$$

By point (i) the sequence $\{f(x_k, y_k, z_k)\}$ converges. This implies that, if K_a is infinite, then the left-hand side of (23) tends to zero, and hence

$$(24) \quad \lim_{k \rightarrow \infty, k \in K_a} \sigma(\mu_k) = 0.$$

By (ii) of Property A (property of $\sigma(\cdot)$), we have

$$(25) \quad \lim_{k \rightarrow \infty, k \in K_a} \mu_k = 0.$$

We now consider the set K_b . If K_b is finite, then the desired result follows from (25). Now assume that K_b is infinite. For each $k \in K_b$, let $m_k \in K_a$ be the largest index satisfying $m_k < k$, with the convention that $m_k = 0$ if K_a is empty. Then we have

$$(26) \quad \mu_k \leq \delta^{k-m_k} \mu_{m_k}.$$

If K_a is finite, then the desired results follows from (26), since $\delta \in (0, 1)$ and $k - m_k$ grows unbounded. If K_a is infinite, the results follows from applying (25) to the right-hand side of (26), while noting that $\delta^{k-m_k} \in (0, 1)$. $\square$

Before stating the first main convergence result, we report the following proposition.

PROPOSITION 5.3. *There exists at least one vector $z^* \in \mathcal{F}_z$ such that the subset $K(z^*)$ of iteration indices is infinite. For any such subset $K(z^*)$,*

- the sequence $\{x_k, y_k, z_k\}_{K(z^*)}$ admits at least one accumulation point;
- for any accumulation point (x^*, y^*, z^*) of the sequence $\{x_k, y_k, z_k\}_{K(z^*)}$, every $(\hat{x}, \hat{y}, \hat{z}) \in \mathcal{N}(x^*, y^*, z^*)$ is an accumulation point of a sequence $\{(\hat{x}_k, \hat{y}_k, \hat{z}_k)\}_{K(z^*)}$, where $(\hat{x}_k, \hat{y}_k, \hat{z}_k) \in \mathcal{N}(x_k, y_k, z_k)$.

Proof. By Assumption A1, z belongs to the finite set $\{z^1, \dots, z^m\}$, and hence, by definition (19), we can write $K_u = \bigcup_{i=1, \dots, m} K(z^i)$. Since the set K_u is infinite (see point (iii) of Proposition 5.2), we can conclude that there exists at least one vector $z^* \in \{z^1, \dots, z^m\}$ such that $K(z^*)$ is infinite.

Then point (ii) of Proposition 5.2 implies that the sequence $\{x_k, y_k, z_k\}_{K(z^*)}$ admits at least one accumulation point. Assumption A4 implies that every $(\hat{x}, \hat{y}, \hat{z}) \in \mathcal{N}(x^*, y^*, z^*)$ is an accumulation point of a sequence $\{(\hat{x}_k, \hat{y}_k, \hat{z}_k)\}_{K(z^*)}$, where $(\hat{x}_k, \hat{y}_k, \hat{z}_k) \in \mathcal{N}(x_k, y_k, z_k)$. $\square$

The main result can now be proved.

THEOREM 5.4. *Let (x_k, y_k, z_k) be the sequence of iterates produced by the algorithm. For any $z^* \in \mathcal{F}_z$ at which $K(z^*)$ is infinite, every accumulation point of the sequence $\{(x_k, y_k, z_k)\}_{K(z^*)}$ is a stationary point of problem (1).*

Proof. Proposition 5.3 states that there exists at least one vector $z^* \in \mathcal{F}_z$ at which $K(z^*)$ is infinite and that the sequence $\{(x_k, y_k, z_k)\}_{K(z^*)}$ admits accumulation points. Let (x^*, y^*, z^*) be any accumulation point of any of these subsequences. We must prove that it satisfies conditions (i), (ii), (iii) stated in Definition 2.3.

Condition (i): Since any iterate (x_k, y_k, z_k) generated by the algorithm is a starting point of procedure *LCS* at Step 1, the result follows from point (v) of Proposition 5.2 and from Property B.

Condition (ii): Since $K(z^*)$ is an infinite subset of the set of unsuccessful iterates and $k \in K(z^*)$, the tests at Steps 2 and 4 fail at iteration k , and hence

$$(27) \quad f(\hat{x}_k, \hat{y}_k, \hat{z}_k) > f(\tilde{x}_k, y_k, z_k) - \eta_k$$

for all $(\hat{x}_k, \hat{y}_k, \hat{z}_k) \in \mathcal{N}(\tilde{x}_k, y_k, z_k)$. Then, recalling that $x_{k+1} = \tilde{x}_k$, $y_{k+1} = y_k$ and $z_{k+1} = z_k$, and by using (i) of Proposition 5.2 we can write

$$(28) \quad f(x^*, y^*, z^*) \leq f(x_{k+1}, y_{k+1}, z_{k+1}) < f(\hat{x}_k, \hat{y}_k, \hat{z}_k) + \eta_k$$

for all $(\hat{x}_k, \hat{y}_k, \hat{z}_k) \in \mathcal{N}(x_k, y_k, z_k)$. Then, the result is obtained in the limit by (iv) of Proposition 5.2, Proposition 5.3, and the continuity of f .

Condition (iii): Note that point (i) of Proposition 5.2 and Assumption A3 ensure the existence of $f^* \in \mathbb{R}$ satisfying

$$(29) \quad \lim_{k \rightarrow \infty} f(x_k, y_k, z_k) = f(x^*, y^*, z^*) = f^*.$$

Consider any $(\hat{x}, \hat{y}, \hat{z}) \in \mathcal{N}(x^*, y^*, z^*)$ such that

$$(30) \quad f(\hat{x}, \hat{y}, \hat{z}) = f^*.$$

We observe that Proposition 5.3 implies that $(\hat{x}, \hat{y}, \hat{z})$ is an accumulation point of a sequence $\{(\hat{x}_k, \hat{y}_k, \hat{z}_k)\}_{K(z^*)}$, where $(\hat{x}_k, \hat{y}_k, \hat{z}_k) \in \mathcal{N}(x_k, y_k, z_k)$. Since $k \in K(z^*)$, we have $x_{k+1} = \tilde{x}_k$, $y_{k+1} = y_k$, $z_{k+1} = z_k$, so that, by (29) and (30), we get, for sufficiently large values of k ,

$$f(\hat{x}_k, \hat{y}_k, \hat{z}_k) < f(\tilde{x}_k, y_k, z_k) + \xi.$$

Therefore, for such values of k , $(\hat{x}_k, \hat{y}_k, \hat{z}_k) \in W_k$, and Steps 3.2–3.4 produce the points $x_k^2, \dots, x_k^{j_k^*}$ (where j_k^* is the finite number of repetitions of Steps 3.2–3.4 until the test at Step 3.4 fails), which, by the instructions at Step 3.2 and by point (i) of Property A, satisfy the following inequalities:

$$(31) \quad f(\hat{x}_k, \hat{y}_k, \hat{z}_k) \geq f(x_k^2, \hat{y}_k, \hat{z}_k) \geq \dots \geq f(x_k^{j_k^*}, \hat{y}_k, \hat{z}_k).$$

Recalling again that $k \in K(z^*)$, which implies that the test at Step 3.3 fails, we can write

$$(32) \quad f(x_k^{j_k^*}, \hat{y}_k, \hat{z}_k) > f(\tilde{x}_k, y_k, z_k) - \eta_k = f(x_{k+1}, y_{k+1}, z_{k+1}) - \eta_k.$$

Moreover, as the sequence $\{\hat{x}_k, \hat{y}_k, \hat{z}_k\}_{K(z^*)}$ converges to the point $(\hat{x}, \hat{y}, \hat{z})$, by (29), (30), (31), (32), and by point (iv) of Proposition 5.2, we obtain

$$(33) \quad \begin{aligned} f^* &= \lim_{k \rightarrow \infty, k \in K(z^*)} f(\hat{x}_k, \hat{y}_k, \hat{z}_k) = \lim_{k \rightarrow \infty, k \in K(z^*)} f(x_k^2, \hat{y}_k, \hat{z}_k) \\ &= \lim_{k \rightarrow \infty} f(x_{k+1}, y_{k+1}, z_{k+1}) = f^*. \end{aligned}$$

Finally, by Property A(ii) of LCS, either

$$(34) \quad f(x_k^2, \hat{y}, \hat{z}) - f(\hat{x}_k, \hat{y}, \hat{z}) \leq -\sigma(\mu_k^1),$$

or

$$(35) \quad \mu_k^1 \leq \delta \mu_k.$$

Since $\{\mu_k\} \rightarrow 0$, it follows easily from (33)–(35) that the sequence $\{\mu_k^1\}$ tends to zero, and hence the result follows again by Property B. $\square$

The convergence analysis just presented proves that the MIVAM algorithm converges to a point satisfying the necessary conditions for optimality introduced in section 2.

We now establish some additional properties of certain limit points of subsequences produced by Step 3.2. This is strongly related to some results stated in [2] (see in particular Theorem 4.8).

PROPOSITION 5.5. *Let (x^*, y^*, z^*) be an accumulation point of the sequence $\{x_k, y_k, z_k\}_{K(z^*)}$, and let $(\hat{x}, \hat{y}, \hat{z}) \in \mathcal{N}(x^*, y^*, z^*)$ be an accumulation point of a sequence $\{(\hat{x}_k, \hat{y}_k, \hat{z}_k)\}_{K(z^*)}$, with $(\hat{x}_k, \hat{y}_k, \hat{z}_k) \in \mathcal{N}(x_k, y_k, z_k)$ (see Proposition 5.3). Then, if $(\hat{x}, \hat{y}, \hat{z}) \in \mathcal{N}(x^*, y^*, z^*)$ satisfies $f(\hat{x}, \hat{y}, \hat{z}) < f(x^*, y^*, z^*) + \xi$, there exists a feasible accumulation point $(\bar{x}, \bar{y}, \bar{z})$ of the subsequence $\{(x_k^{j_k^*}, \hat{y}_k, \hat{z}_k)\}_{\bar{K}(z^*)}$, where*

$$\bar{K} = \{k \in K(z^*) : (\hat{x}_k, \hat{y}_k, \hat{z}_k) \in W_k\},$$

and $x_k^{j_k^*}$ is obtained at iteration k by starting from $(\hat{x}_k, \hat{y}_k, \hat{z}_k)$ and repeating Steps 3.2–3.4 j_k^* times until the test at Step 3.4 fails. Moreover, the point $(\bar{x}, \bar{y}, \bar{z})$ satisfies

$$(36) \quad f(x^*, y^*, z^*) \leq f(\bar{x}, \bar{y}, \bar{z}) \leq f(\hat{x}, \hat{y}, \hat{z})$$

and $\bar{x}$ is a stationary point of the problem

$$(37) \quad \begin{aligned} \min f(x, \hat{y}, \hat{z}), \\ x \in \mathcal{F}_x(\hat{y}, \hat{z}). \end{aligned}$$

Proof. Under Assumption A3, Proposition 5.3 ensures that every point $(\hat{x}, \hat{y}, \hat{z}) \in \mathcal{N}(x^*, y^*, z^*)$ satisfying $f(\hat{x}, \hat{y}, \hat{z}) < f(x^*, y^*, z^*) + \xi$ is an accumulation point of a sequence $\{(\hat{x}_k, \hat{y}_k, \hat{z}_k)\}_{K(z^*)}$, where $(\hat{x}_k, \hat{y}_k, \hat{z}_k) \in \mathcal{N}(x_k, y_k, z_k)$. Since $k \in K(z^*)$, we have $x_{k+1} = \tilde{x}_k$, $y_{k+1} = y_k$, $z_{k+1} = z_k$, so that, as $f(x_k, y_k, z_k)$ converges to the value $f(x^*, y^*, z^*)$ (by (i) of Proposition 5.2), we get, for sufficiently large values of k ,

$$(38) \quad f(\hat{x}_k, \hat{y}_k, \hat{z}_k) \leq f(\tilde{x}_k, y_k, z_k) + \xi.$$

Thus the set $\bar{K}(z^*)$ is not empty. Then Assumption A2 implies (see proof of point (ii) of Proposition 5.2) that the sequence $\{(x_k^{j_k^*}, \hat{y}_k, \hat{z}_k)\}_{\bar{K}(z^*)}$ admits at least one accumulation point, which is feasible by point (i) of Property A of *LCS*.

Since $k \in K(z^*)$, the test of Step 3.3 fails at the k th iteration; that is,

$$f(x_k^{j_k^*}, \hat{y}_k, \hat{z}_k) > f(\tilde{x}_k, y_k, z_k) - \eta_k = f(x_{k+1}, y_{k+1}, z_{k+1}) - \eta_k.$$

Then by points (i) and (iv) of Proposition 5.2, for $k \in \bar{K}(z^*)$ and $k \rightarrow \infty$,

$$f(\hat{x}, \hat{y}, \hat{z}) \geq f(\bar{x}, \hat{y}, \hat{z}) \geq f(x^*, y^*, z^*),$$

where the left inequality follows from Property A of *LCS* applied to Step 3.2.

Furthermore, since the instructions at Step 3.4 yield $\mu_k^{j_k^*} \leq \mu_k$ for all k , it follows from point (v) of Proposition 5.2 that

$$(39) \quad \lim_{k \rightarrow \infty} \mu_k^{j_k^*} = 0$$

and hence, by Property B, the thesis. $\square$

The property stated in the above proposition can be particularly interesting under some stronger assumptions on the continuous subproblem (4), as stated in the next corollary.

COROLLARY 5.6. *Assume that, for any $\bar{z} \in \mathcal{F}_z$ and for any $\bar{y} \in \mathcal{F}_y(\bar{z})$, each stationary point of problem (4) is also a global minimum of problem (4). Let (x^*, y^*, z^*) be an accumulation point of the sequence $\{x_k, y_k, z_k\}_{K(z^*)}$. Then, for any $(\hat{x}, \hat{y}, \hat{z}) \in \mathcal{N}(x^*, y^*, z^*)$ such that $f(\hat{x}, \hat{y}, \hat{z}) < f(x^*, y^*, z^*) + \xi$, we have*

$$f(x^*, y^*, z^*) \leq f(x, \hat{y}, \hat{z}) \quad \forall x \in \mathcal{F}_x(\hat{y}, \hat{z}).$$

Proof. Consider a point $(\hat{x}, \hat{y}, \hat{z}) \in \mathcal{N}(x^*, y^*, z^*)$ such that $f(\hat{x}, \hat{y}, \hat{z}) < f(x^*, y^*, z^*) + \xi$. Proposition 5.5 implies that there exists a feasible point $(\bar{x}, \hat{y}, \hat{z})$ such that

$$(40) \quad f(x^*, y^*, z^*) \leq f(\bar{x}, \hat{y}, \hat{z})$$

and $\bar{x}$ is a stationary problem of the problem

$$(41) \quad \min f(x, \hat{y}, \hat{z}), \\ x \in \mathcal{F}_x(\hat{y}, \hat{z}).$$

By assumption, $\bar{x}$ is a global minimum point of problem (41), so that the thesis follows from (40). $\square$

6. A derivative-free algorithm. We consider here a subclass of problem (1), which includes the optimal design problem described in section 3. In particular, we assume that the feasible set $\mathcal{F}_x$ is defined by linear constraints and that the derivatives of $f(x, y, z)$ with respect to continuous variables x can be neither calculated nor approximated explicitly. More formally, the problem we consider is

$$(42) \quad \begin{aligned} \min & f(x, y, z), \\ & z \in \mathcal{F}_z, \\ & y \in \mathcal{F}_y(z), \\ & A(y, z)x \leq b(y, z), \\ & z \in \mathcal{Z}^{n_z}, \quad y \in \mathcal{Z}^{n_y(z)}, \end{aligned}$$

where $A(y, z) \in \mathbb{R}^{m(y, z) \times n_x(z)}$ and $b(y, z) \in \mathbb{R}^{m(y, z)}$,

We suppose that Assumptions A1, A2, A4 are verified and we replace Assumption A3 with the following assumption.

A3'. For each $z \in \mathcal{F}_z$ and $y \in \mathcal{F}_y(z)$, the function $f(\cdot, y, z)$ is continuously differentiable in a neighborhood of $\{x : (x, y) \in \mathcal{L}_z(x_0, y_0, z_0)\}$.

Given a feasible point $(\bar{x}, \bar{y}, \bar{z})$, we define the set of indices of active constraints

$$I(\bar{x}, \bar{y}, \bar{z}) = \{i \in \{1, \dots, m(\bar{y}, \bar{z})\} : a_i(\bar{y}, \bar{z})^T \bar{x} = b_i(\bar{y}, \bar{z})\}$$

and the cone of feasible directions

$$T(\bar{x}, \bar{y}, \bar{z}) = \{d \in \mathbb{R}^{n_x(\bar{z})} : a_i(\bar{y}, \bar{z})^T d \leq 0 \quad \forall i \in I(\bar{x}, \bar{y}, \bar{z})\}.$$

In order to apply Definition 2.3 to problem (42), we refer to a stationary point of the continuous problem,

$$(43) \quad \begin{aligned} \min & f(x, \bar{y}, \bar{z}), \\ & A(\bar{y}, \bar{z})x \leq b(\bar{y}, \bar{z}), \end{aligned}$$

where $\bar{z} \in \mathcal{F}_z$, $\bar{y} \in \mathcal{F}_y(\bar{z})$ are fixed, as any point $(\bar{x}, \bar{y}, \bar{z})$ satisfying

$$(44) \quad \nabla_x f(\bar{x}, \bar{y}, \bar{z})^T d \geq 0 \quad \forall d \in T(\bar{x}, \bar{y}, \bar{z}).$$

We describe a particular instance of the MIVAM algorithm by specifying procedure *LCS*, which takes into account the features of problem (42). In order to overcome the lack of gradient information, we draw inspiration from the derivative-free method proposed in [10], which is an iterative method that generates a sequence of feasible points with nonincreasing objective function value. At each iteration the objective function is evaluated at a finite number of points along a suitable set of search directions in order to try to find one that provides a sufficient decrease of the objective function value. Similarly here we define a procedure based on a sampling of the objective function along sets of search directions with theoretical properties analogous to the ones of the set of directions used in [8], [10], and [3]. In particular, we assume that for every feasible point (x, y, z) there exists a set of search directions

$$D(x, y, z) = \{d^1, \dots, d^r : d^i \in \mathbb{R}^{n_x(z)}, \|d^i\| = 1, i = 1, \dots, r\}$$

that satisfies the following condition.

Condition 1

If a sequence of points $\{x_k, y_k, z_k\}$ converges to a point $(\bar{x}, \bar{y}, \bar{z})$, then

- i) $\text{cone}(D(x_k, y_k, z_k) \cap T(\bar{x}, \bar{y}, \bar{z})) = T(\bar{x}, \bar{y}, \bar{z})$ for k sufficiently large;
- ii) $\cup_k D(x_k, y_k, z_k)$ is finite.

A technique for constructing a set of directions satisfying Condition 1 can be found in [8].

We present below the derivative-free (DF) procedure which plays the role of procedure *LCS* in the proposed instance of the MIVAM algorithm. We assume that the input point (x, y, z) is feasible.

Procedure DF $(x, \mu^0, \tilde{x}, \mu, y, z)$

Data: $\gamma \geq 0$, $\delta_1 \in (0, 1)$, $\delta_2 \in (0, 1)$.

Step 1: Determine a set of directions $D(x, y, z) = \{d^1, \dots, d^r\}$ satisfying Condition 1.

Step 2: Set $i = 1$, $x^i = x$.

Step 3: Compute the maximum steplength $\bar{\alpha}^i$ such that $x^i + \bar{\alpha}^i d^i \in \mathcal{F}_x(y, z)$.
Set $\hat{\alpha}^i = \min\{\bar{\alpha}^i, \mu^0\}$.

Step 4: If $\hat{\alpha}^i = 0$ or $f(x^i + \hat{\alpha}^i d^i, y, z) > f(x^i, y, z) - \gamma(\hat{\alpha}^i)^2$, set $\alpha^i = 0$ and go to Step 6.

Step 5: Set $\alpha = \hat{\alpha}^i$

5.1: Let $\tilde{\alpha} = \min\{\bar{\alpha}^i, \frac{\alpha}{\delta_2}\}$.

5.2: If $\tilde{\alpha} = \bar{\alpha}^i$ or $f(x^i + \tilde{\alpha} d^i, y, z) > f(x^i, y, z) - \gamma \tilde{\alpha}^2$, set $\alpha^i = \alpha$, and go to Step 6.

5.3: Set $\alpha = \tilde{\alpha}$ and go to Step 5.1.

Step 6: Set $x^{i+1} = x^i + \alpha^i d^i$. If $i < r$, set $i = i + 1$, and go to Step 3.

Step 7: Set $\tilde{x} = x^{i+1}$. Set $\mu = \max\{\max_{i=1, \dots, r} \{\alpha^i\}, \delta_1 \mu^0\}$ and stop.

First, we show that Procedure DF is well defined.

PROPOSITION 6.1. *Assume that $f(x, y, z)$ is bounded from below on $\mathcal{L}_z(x, y, z)$. Then Procedure DF is well defined; i.e., Step 5 terminates in a finite number of steps.*

Proof. Suppose by contradiction that Step 5 in Procedure DF fails to terminate and let j be the subiteration counter associated with the loop in Steps 5.1–5.3. Then there exists an index $i \in \{1, \dots, r\}$ such that for all j

$$(45) \quad \hat{\alpha}^i > 0,$$

$$(46) \quad x^i + \frac{\alpha}{\delta_2^j} d^i \in \mathcal{F}_x(y, z),$$

$$(47) \quad f\left(x^i + \frac{\alpha}{\delta_2^j} d^i, y, z\right) \leq f(x^i, y, z) - \gamma \left(\frac{\alpha}{\delta_2^j}\right)^2, \quad j = 0, 1, \dots$$

Taking limits in (47) for $j \rightarrow \infty$, we obtain that the objective function is unbounded from below on the level set $\mathcal{L}_z(x, y, z)$, but this contradicts the assumption. $\square$

We note that the hypothesis of Proposition 6.1 holds under Assumptions A2 and A3'.

The next theorems show that Procedure DF satisfies the properties needed to ensure the global convergence of the sequence of points generated by the MIVAM algorithm.

THEOREM 6.2. *Procedure DF satisfies Property A.*

Proof. The instructions of Procedure DF and the convexity of the feasible set $\mathcal{F}_x(y, z)$ imply that the point $\tilde{x}$ produced by Procedure DF is feasible; furthermore, $f(\tilde{x}, y, z) \leq f(x, y, z)$, so that point (i) of Property A holds.

Regarding point (ii), by the instructions of DF, either $\mu = \delta_1 \mu^0$ or $\mu = \max_{i=1, \dots, r} \{\alpha^i\}$. In the latter case, $\tilde{x} \neq x$ and

$$(48) \quad f(x^{i+1}, y, z) \leq f(x^i, y, z) - \gamma(\alpha^i)^2, \quad i = 1, \dots, r,$$

and hence

$$(49) \quad f(\tilde{x}, y, z) \leq f(x, y, z) - \gamma \sum_{i=1}^r (\alpha^i)^2.$$

Since $\tilde{x} \neq x$, there exists at least one index i such that $\alpha^i \neq 0$, so that (49) implies

$$(50) \quad f(\tilde{x}, y, z) \leq f(x, y, z) - \gamma \left(\max_{i=1, \dots, r} \{\alpha^i\}\right)^2 = f(x, y, z) - \gamma(\mu)^2,$$

and hence point (ii) holds. $\square$

THEOREM 6.3. *Procedure DF satisfies Property B.*

Proof. Let $\{(x_h, y_h, z_h)\}$ be a sequence of feasible points. We need to prove that, if we apply, for all h , Procedure DF starting from (x_h, y_h, z_h) , and if the corresponding sequence $\{\mu_h\}$ satisfies

$$(51) \quad \lim_{h \rightarrow \infty} \mu_h = 0,$$

then every accumulation point $(\bar{x}, \bar{y}, \bar{z})$ of the sequence $\{(x_h, y_h, z_h)\}$ is such that $\bar{x}$ is a stationary point of the continuous problem

$$(52) \quad \begin{aligned} &\min f(x, \bar{y}, \bar{z}), \\ &x \in \mathcal{F}_x(\bar{y}, \bar{z}); \end{aligned}$$

that is,

$$(53) \quad \nabla_x f(\bar{x}, \bar{y}, \bar{z})^T d \geq 0 \quad \forall d \in T(\bar{x}, \bar{y}, \bar{z}).$$

We suppose by contradiction that there exists a direction $\bar{d} \in T(\bar{x}, \bar{y}, \bar{z})$ such that

$$(54) \quad \nabla_x f(\bar{x}, \bar{y}, \bar{z})^T \bar{d} < 0.$$

We denote by d_h^i , $i = 1, \dots, r$, the directions used by DF when applied at x_h , with corresponding point x_h^i and scalars α_h^i and $\bar{\alpha}_h^i$. Condition 1(i) implies that there exists a direction $d_h^{i(h)}$ such that, for h sufficiently large, $d_h^{i(h)} \in D(x_h, y_h, z_h) \cap T(\bar{x}, \bar{y}, \bar{z})$ and

$$(55) \quad \nabla_x f(\bar{x}, \bar{y}, \bar{z})^T d_h^{i(h)} < 0.$$

By using Condition 1(ii) and by relabelling the sequence $\{x_h, y_h, z_h\}$ (if necessary), we can assume $d_h^{i(h)} = d^\ell$ for all h , with $\ell \in \{1, \dots, r\}$, and hence we have

$$(56) \quad \nabla_x f(\bar{x}, \bar{y}, \bar{z})^T d^\ell < 0.$$

Since

$$x_h^\ell = x_h + \sum_{j=1}^{\ell-1} \alpha_h^j d_h^j,$$

we can write

$$(57) \quad \begin{aligned} \|x_h^\ell - x_h\| &= \left\| \sum_{j=1}^{\ell-1} \alpha_h^j d_h^j \right\| \leq \sum_{j=1}^{\ell-1} \|\alpha_h^j d_h^j\| \\ &= \sum_{j=1}^{\ell-1} \alpha_h^j \leq r \max_{j=1, \dots, r} \{\alpha_h^j\} \leq r\mu_h, \end{aligned}$$

where the last inequality follows from the definition of μ_h at Step 7. It then follows from (51) that

$$(58) \quad \lim_{h \rightarrow \infty} x_h^\ell = \bar{x}.$$

Since $\alpha_h^\ell \leq \mu_h$ (also by Step 7), (51) implies that

$$(59) \quad \lim_{h \rightarrow \infty} \alpha_h^\ell = 0.$$

Now we show that, for h sufficiently large,

$$(60) \quad 0 \leq \frac{\alpha_h^\ell}{\delta_2} < \bar{\alpha}_h^\ell.$$

Given (60), it is sufficient to show that $\bar{\alpha}_h^\ell$ is bounded away from zero. Note that, for h sufficiently large, the sets of indices of active constraints $I(x_h, y_h, z_h)$ is contained in the set $I(\bar{x}, \bar{y}, \bar{z})$, i.e., $I(x_h, y_h, z_h) \subseteq I(\bar{x}, \bar{y}, \bar{z})$, which implies $T(\bar{x}, \bar{y}, \bar{z}) \subseteq T(x_h, y_h, z_h)$. Thus, for h sufficiently large, $d^\ell \in T(x_h, y_h, z_h)$.

The definition of $\bar{\alpha}_h^\ell$ implies either $\bar{\alpha}_h^\ell = \infty$ or the existence of an index $j(h) \notin I(x_h, y_h, z_h)$ such that

$$(61) \quad a_{j(h)}^T (x_h^\ell + \bar{\alpha}_h^\ell d^\ell) = b_{j(h)}.$$

Since constraint $j(h)$ is not binding, (61) yields $0 < a_{j(h)}^T d^\ell$, and since $d^\ell \in T(\bar{x}, \bar{y}, \bar{z})$, it follows that

$$(62) \quad b_{j(h)} - a_{j(h)}^T \bar{x} = \epsilon > 0.$$

Solving (61) for $\bar{\alpha}_h^\ell$ yields

$$\begin{aligned}\bar{\alpha}_h^\ell &= \frac{b_{j(h)} - a_{j(h)}^T x_h^\ell}{a_{j(h)}^T d^\ell} = \frac{b_{j(h)} - a_{j(h)}^T x_h^\ell + a_{j(h)}^T \bar{x} - a_{j(h)}^T \bar{x}}{a_{j(h)}^T d^\ell} \\ &\geq \frac{b_{j(h)} - a_{j(h)}^T \bar{x} - \|a_{j(h)}\| \|x_h^\ell - \bar{x}\|}{a_{j(h)}^T d^\ell}.\end{aligned}$$

By (62) we get

$$(63) \quad \bar{\alpha}_h^\ell \geq \frac{\epsilon - \|a_{j(h)}\| \|x_h^\ell - \bar{x}\|}{\|a_{j(h)}\|}.$$

Since, by assumption, $\|a_{j(h)}\|$ is bounded, (63) and (58) imply that $\bar{\alpha}_h^\ell$ is bounded away from zero and (60) holds.

Then we have two possibilities: either $\alpha_h^\ell > 0$ or $\alpha_h^\ell = 0$.

Let

$$H_1(\ell) = \{h : \alpha_h^\ell > 0\}$$

and

$$H_2(\ell) = \{h : \alpha_h^\ell = 0\}.$$

For all sufficiently large $h \in H_1(\ell)$, the definition of α_h^ℓ and (60) imply

$$(64) \quad f\left(x_h^\ell + \frac{\alpha_h^\ell}{\delta_2} d_h^\ell, y_h, z_h\right) - f(x_h^\ell, y_h, z_h) > -\gamma \left(\frac{\alpha_h^\ell}{\delta_2}\right)^2.$$

By the mean value theorem, we obtain

$$(65) \quad \nabla_x f\left(x_h^\ell + \lambda \frac{\alpha_h^\ell}{\delta_2} d_h^\ell, y_h, z_h\right)^T d_h^\ell > -\gamma \left(\frac{\alpha_h^\ell}{\delta_2}\right)^2,$$

where $\lambda \in (0, 1)$. Taking limits in (65) as $h \rightarrow \infty$, $h \in H_1(\ell)$, recalling that $(x_h^\ell, y_h^\ell, z_h^\ell) \rightarrow (\bar{x}, \bar{y}, \bar{z})$ with $\alpha_h^\ell \rightarrow 0$, we can write

$$(66) \quad \nabla_x f(\bar{x}, \bar{y}, \bar{z})^T \bar{d}^\ell = \lim_{h \rightarrow \infty, h \in H_1(\ell)} \nabla_x f(x_h^\ell, y_h^\ell, z_h^\ell)^T d_h^\ell \geq 0,$$

which contradicts (56).

For all $h \in H_2(\ell)$, the instructions of the algorithm and $\bar{\alpha}_h^\ell > 0$ by (60) imply that

$$(67) \quad f(x_h^\ell + \hat{\alpha}_h^\ell d_h^\ell, y_h, z_h) - f(x_h^\ell, y_h, z_h) > -\gamma (\hat{\alpha}_h^\ell)^2.$$

By the mean value theorem, we obtain

$$(68) \quad \nabla_x f(x_h^\ell + \tilde{\lambda} \hat{\alpha}_h^\ell d_h^\ell, y_h, z_h)^T d_h^\ell > -\gamma (\hat{\alpha}_h^\ell)^2,$$

where $\tilde{\lambda} \in (0, 1)$. By repeating similar reasoning used above and recalling that the instructions of Procedure DF imply

$$\hat{\alpha}_h^\ell \leq \mu_h^0 \leq \frac{\mu_h}{\delta_1},$$

we obtain

$$(69) \quad \nabla_x f(\bar{x}, \bar{y}, \bar{z})^T \bar{d}^\ell = \lim_{h \rightarrow \infty, h \in H_2(\ell)} \nabla_x f(x_h, y_h, z_h)^T d_h^\ell \geq 0,$$

which contradicts (56). $\square$

Theorems 6.2 and 6.3 show that the particular instance of the MIVAM algorithm where Procedure DF is used as procedure *LCS* maintains all the convergence properties established in section 5 for the general case.

Remark. Note that the description of Procedure DF looks somewhat similar to one iteration of a pattern search algorithm for linearly constrained continuous minimization problems (see [8]). However, Procedure DF is derived from a linesearch approach and hence it is based on a sufficient decrease condition, while pattern search requires a simple decrease of the objective function on a lattice to accept the new point. The stronger descent condition on the objective function, along with the relaxation of the requirement to have rational search directions, allows steps in the chosen directions to be independent of each other. This feature permits to take into account the different behavior of the objective function along the search directions.

7. Optimal design of a magnetic resonance device: Numerical results.

In this section we report computational results for the optimal design of a small magnetic resonance device described in section 3.

We applied the MIVAM algorithm to this problem using the definition of discrete neighborhood described in section 3 and parameter values set at $\xi = 10^{-2}$, $\theta = 0.5$, $\eta_0 = 10^{-5}$, $\mu^{in} = 0.5$. These choices seemed reasonable on the basis of the range of the objective function values.

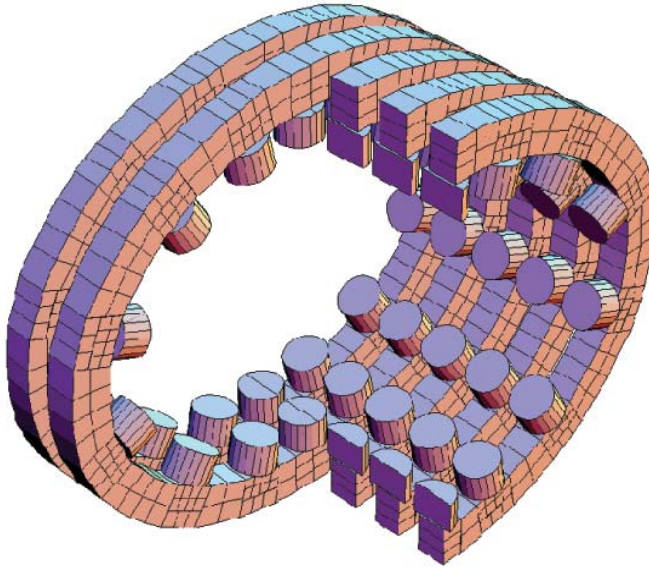
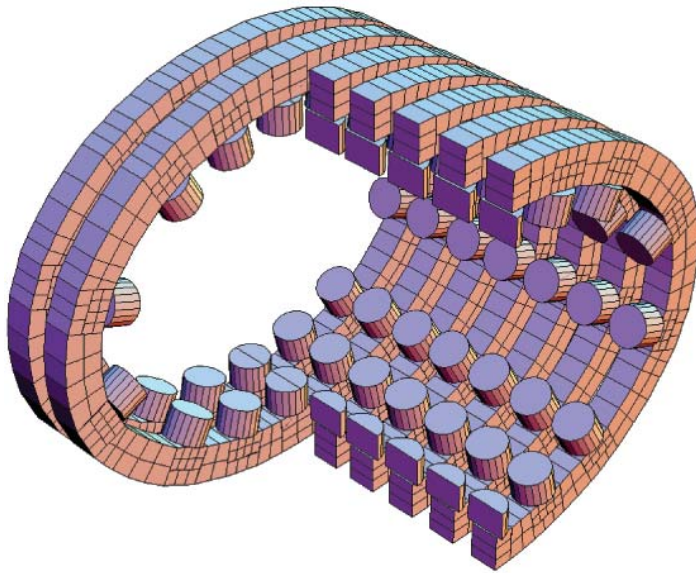
For *LCS*, we employed Procedure DF introduced in section 6 with $\gamma = 10^{-6}$, $\delta_1 = \delta_2 = 0.5$, which are typical parameter values for derivative-free linesearch methods. The directions determined at Step 1 of Procedure DF have been computed as described in [10]. In particular, the set of directions is computed by performing a QR factorization of the matrix corresponding to the current set of ϵ -active constraints.

Finally, we terminated the MIVAM algorithm when μ_k fell below 10^{-6} .

The device structure corresponding to the starting point is reported in Figure 5. This structure, which was considered reasonable by magnetic resonance experts, is made of five rings and three 22 mm magnets. The initial objective function value was $f^0 = 0.3715$. The corresponding value of the continuous variables is

$$\begin{aligned} \varphi_1 &= 22.50^\circ, \\ \varphi_2 &= 49.50^\circ, \\ \varphi_3 &= 77.50^\circ, \\ d_1 &= 27.50 \text{ mm}, & b_1 &= 2.50 \text{ mm}, \\ d_2 &= 82.50 \text{ mm}, & b_2 &= 2.50 \text{ mm}, \\ d_3 &= 137.50 \text{ mm}, & b_3 &= 2.50 \text{ mm}, \\ d_4 &= 192.50 \text{ mm}, & b_4 &= 2.50 \text{ mm}, \\ d_5 &= 247.50 \text{ mm}, & b_5 &= 2.50 \text{ mm}. \end{aligned}$$

The best solution $f^* = 6.03 \times 10^{-3}$ was found by the MIVAM algorithm after 2162 objective function evaluations. Depicted in Figure 6, this structure is made of seven rings and three 22 mm magnets. The corresponding continuous variables are

FIG. 5. *Structure corresponding to the starting point.*FIG. 6. *Structure corresponding to the best point found by the MIVAM algorithm.*

the following:

$$\begin{aligned}\varphi_1 &= 23.49^\circ, \\ \varphi_2 &= 46.02^\circ, \\ \varphi_3 &= 77.16^\circ,\end{aligned}$$

$$\begin{aligned}
d_1 &= 27.49 \text{ mm}, & b_1 &= 15.00 \text{ mm}, \\
d_2 &= 82.50 \text{ mm}, & b_2 &= 4.51 \text{ mm}, \\
d_3 &= 138.02 \text{ mm}, & b_3 &= 4.67 \text{ mm}, \\
d_4 &= 189.96 \text{ mm}, & b_4 &= 2.51 \text{ mm}, \\
d_5 &= 248.51 \text{ mm}, & b_5 &= 2.50 \text{ mm}, \\
d_6 &= 302.45 \text{ mm}, & b_6 &= 2.50 \text{ mm}, \\
d_7 &= 359.51 \text{ mm}, & b_7 &= 2.50 \text{ mm}.
\end{aligned}$$

We note that uniformity of the magnetic field has been greatly improved by a relatively small number of function evaluations. Moreover, by comparing Figures 5 and 6 and the design variable values, we can observe that the optimization procedure modified the initial configuration in a significant way.

In order to further evaluate possible advantages of the proposed mixed optimization algorithm, we performed additional experiments.

In particular, we applied a different strategy that does not alternate continuous and discrete searches as the MIVAM algorithm does. In this approach, we instead treated the dimensional and categorical variables as parameters, whose values were chosen by considering the points belonging to the discrete neighborhood of the initial vector. Then, starting from each one of these points and keeping fixed the discrete variables, the corresponding nonlinear continuous problem was solved; i.e., a stationary point with respect to the continuous variables was produced by applying iteratively Procedure DF.

Therefore, the comparison between the proposed mixed algorithm and the above-described strategy appears reasonable.

In particular, the implemented strategy found the best solution $f^* = 7.54 \times 10^{-3}$ (with 6 rings, 3 22 mm magnets), which is significantly worse (with respect to the considered application) than the one determined by the MIVAM algorithm. Moreover, it required 5579 function evaluations, which is much greater than the MIVAM algorithm needed.

8. Conclusions. In this paper, we consider mixed integer nonlinear programming problems with the distinguishing feature that some of the discrete variables cannot be relaxed into continuous ones since they may characterize the structure of the problem or determine the number of variables and constraints. In particular, we propose the algorithm model MIVAM which combines in a suitable way a local continuous search and a local search with respect to the discrete variables. The convergence of the algorithm model is proved without specifying the structure of the problem and without specifying the local continuous search technique, which must only satisfy some reasonable requirements. As an example of algorithms which can be defined starting from the algorithm model MIVAM, we describe a derivative-free algorithm for solving linearly constrained minimization problems and we test it on an optimal design problem. The obtained results indicate that the proposed approach is promising, even though further work is needed on the specific optimal design problem considered. In particular, in order to obtain a better solution, two aspects may deserve further attention: more sophisticated strategies for defining the discrete neighborhoods, and the possible application of global optimization techniques in the continuous minimization process.

Acknowledgments. The authors would like to thank two anonymous referees and the guest editor for their helpful comments, suggestions, and corrections which greatly contributed in improving the paper.

REFERENCES

- [1] M. A. ABRAMSON, *Pattern Search Algorithms for Mixed Variable General Constrained Optimization Problems*, Ph.D. thesis, Department of Computational and Applied Mathematics, Rice University, 2002.
- [2] C. AUDET AND J. E. DENNIS, JR., *Pattern search algorithms for mixed variable programming*, SIAM J. Optim., 11 (2001), pp. 573–594.
- [3] C. AUDET AND J. E. DENNIS, JR., *Analysis of generalized pattern searches*, SIAM J. Optim., 13 (2003), pp. 889–903.
- [4] C. BERGE, *Topological Spaces*, MacMillan, New York, 1963.
- [5] O. CHUBAR, P. ELLEAUME, AND J. CHAVANNE, *A 3d magnetostatics computer code for insertion devices*, J. Synchrotron Radiation, 5 (1998), pp. 481–484.
- [6] P. ELLEAUME, O. CHUBAR, AND J. CHAVANNE, *Computing 3d magnetic field from insertion devices*, in Proceedings of the PAC97 Conference, Vancouver, BC, Canada, 1997, pp. 3509–3511.
- [7] R. M. LEWIS AND V. TORCZON, *Pattern search algorithms for bound constrained minimization*, SIAM J. Optim., 9 (1999), pp. 1082–1099.
- [8] R. M. LEWIS AND V. TORCZON, *Pattern search methods for linearly constrained minimization*, SIAM J. Optim., 10 (2000), pp. 917–941.
- [9] G. LIUZZI, S. LUCIDI, G. PLACIDI, AND A. SOTGIU, *A Magnetic Resonance Device Designed via Global Optimization Techniques*, Technical report, 09-02, Department of Computer and Systems Science “Antonio Ruberti,” University of Rome “La Sapienza,” 2002.
- [10] S. LUCIDI, M. SCIANDRONE, AND P. TSENG, *Objective-derivative-free methods for constrained optimization*, Math. Program., 92 (2002), pp. 37–59.
- [11] S. LUCIDI AND V. PICCIALLI, *A derivative-based algorithm for a particular class of mixed variable optimization problems*, Optim. Methods Softw., 19 (2004), pp. 371–387.
- [12] V. PICCIALLI, *Methods for Solving Mixed Variable Programming Problems*, Ph.D. thesis, Department of Computer and Systems Science “Antonio Ruberti,” University of Rome “La Sapienza,” 2003.

EXTREME POINTS OF SETS OF RANDOMIZED STRATEGIES IN CONSTRAINED OPTIMIZATION AND CONTROL PROBLEMS*

JUAN GONZÁLEZ-HERNÁNDEZ[†] AND ONÉSIMO HERNÁNDEZ-LERMA[‡]

Abstract. This paper concerns the existence and characterization of optimal randomized strategies for some constrained optimization and control problems. We first present a characterization of the extreme points of a set of randomized strategies that satisfy n moment-like constraints. Conditions are given under which those extreme points are randomizations of at most $n + 1$ *deterministic* strategies. This result is then applied to obtain the existence and characterization of optimal strategies for a class of deterministic, allocation-like, optimization problems and their Young relaxations. Similar results are obtained for constrained Markov control processes in Borel spaces.

Key words. randomized strategies (relaxed controls), extreme points of convex sets, Young’s relaxation technique, allocation problems, constrained Markov control processes

AMS subject classifications. 90C25, 46A55, 90C40

DOI. 10.1137/040605345

1. Introduction. In some areas—e.g., optimal control, game theory, and statistical decision theory, to name just a few—there are optimization problems in which one has to deal with “randomized” decisions or control actions, known as *randomized strategies*. (Other names: Young measures, parametrized measures, relaxed controls, randomized control policies, etc.; see [3, 5, 7, 8, 9, 10, 11, 12, 13, 14, 16, 17, 18, 20, 21, 22, 23, 27, 28, 29, 30].) There are, on the other hand, “deterministic” optimization problems which can be *relaxed* via randomized strategies. For these classes of optimization problems, an important issue is of course to find conditions for the *existence* and *characterization* of optimal randomized strategies, which are the questions addressed in this paper.

To motivate our general approach and main results, we shall first present a brief description of Young’s relaxation technique, which can be traced back to the 1930s (see, e.g., [9, 30, 42] for earlier references), originally introduced for problems in the calculus of variations.

Young’s relaxation technique. Consider two topological spaces X and A , endowed with their Borel σ -algebras $\mathcal{B}(X), \mathcal{B}(A)$, and let c_0 be a real-valued *cost* function on $X \times A$. Moreover, let ν be a probability measure (p.m.) on $\mathcal{B}(X)$, and let $\mathbb{F}_0$ be the set of all (Borel) measurable functions from X to A . A function f in $\mathbb{F}_0$ will be referred to as a *deterministic strategy* or decision rule. The problem then is

$$(1.1) \quad \text{minimize } \int_X c_0(x, f(x)) \nu(dx) \text{ over all } f \in \mathbb{F}_0.$$

Note that this simple-looking problem is in fact far from being trivial because, without suitable (restrictive) hypotheses, one can use *none* of the usual continuity-compactness arguments for optimization problems.

*Received by the editors March 17, 2004; accepted for publication August 23, 2004; published electronically July 26, 2005. This research was partially supported by Consejo Nacional de Ciencia y Tecnología (CONACYT) grant 37355-E.

<http://www.siam.org/journals/siopt/15-4/60534.html>

[†]Departamento de Probabilidad y Estadística, IIMAS-UNAM, A. Postal 20-726, Mexico D.F. 01000, Mexico (juan@sigma.iimas.unam.mx).

[‡]Corresponding author. Departamento de Matemáticas, CINVESTAV-IPN, A. Postal 14-740, Mexico D.F. 0700, Mexico (ohernand@math.cinvestav.mx).

Young's relaxation technique to solve (1.1) consists basically of two steps. The first one is to replace each control or decision function $f \in \mathbb{F}_0$ with the Dirac measure (or "unit mass") $\delta_{f(x)}(\cdot)$ concentrated at $f(x)$ for each $x \in X$ (i.e., $\delta_{f(x)}(D) := 1$ if $f(x)$ is in D , and 0 otherwise, for all $D \in \mathcal{B}(A)$). Actually, the set of all these Dirac measures is not sufficiently large and so, more precisely, we have to replace $\mathbb{F}_0$ with the set Φ_0 defined as follows.

DEFINITION 1.1. A randomized strategy φ from X to A (or on A given X) is a mapping $(x, D) \mapsto \varphi(D|x)$ from $X \times \mathcal{B}(A)$ to the interval $[0, 1]$ such that

- (a) $\varphi(\cdot|x)$ is a p.m. on $\mathcal{B}(A)$ for every fixed $x \in X$, and
- (b) $\varphi(D|\cdot)$ is a measurable function on X for every fixed $D \in \mathcal{B}(A)$.

We shall denote by Φ_0 the set of all these randomized strategies.

Remark. For the reader unfamiliar with "deterministic" and "randomized" decision strategies, we mention that a deterministic strategy $f \in \mathbb{F}_0$ prescribes the action $f(x) \in A$ whenever the state of nature is $x \in X$, whereas for a randomized strategy $\varphi \in \Phi_0$ the corresponding action is an A -valued random variable with probability distribution $\varphi(\cdot|x)$ for each $x \in X$.

Given a deterministic strategy $f \in \mathbb{F}_0$, we may identify each $f(x) \in A$ with the Dirac measure $\varphi(\cdot|x) := \delta_{f(x)}(\cdot)$, which of course defines a randomized strategy. With this identification we may then write

$$(1.2) \quad \mathbb{F}_0 \subset \Phi_0.$$

Moreover, defining

$$(1.3) \quad c_0(x, \varphi(x)) := \int_A c_0(x, a) \varphi(da|x) \quad \forall \varphi \in \Phi_0, x \in X,$$

the relaxation of problem (1.1) can be expressed as

$$(1.4) \quad \text{minimize } \int_X c_0(x, \varphi(x)) \nu(dx) \text{ over all } \varphi \in \Phi_0.$$

We may now note that (1.4) is a convex problem because it is evident that the mapping

$$(1.5) \quad \varphi \mapsto \int_X c_0(x, \varphi(x)) \nu(dx)$$

is linear on the convex set Φ_0 . We have thus transformed, or "relaxed", (1.1) into the convex problem (1.4), which completes the first step in Young's relaxation technique.

In the second (and final) step, which is the difficult one, we have to show that (under some assumptions)

- (a) problem (1.4) is solvable and that
- (b) the existence of optimal solutions to (1.4) yields the existence of optimal solutions (in $\mathbb{F}_0$) to the original problem (1.1).

These are the questions we are concerned with in this paper. Part (a) is more or less standard. On the other hand, if (a) holds, we then obtain (b) as follows: we show that (1.4) has an optimal solution that is an extreme point of Φ_0 , which, therefore, is in $\mathbb{F}_0$ because it turns out that $\text{ex}(\Phi_0)$, the set of extreme points of Φ_0 , coincides with $\mathbb{F}_0$. See Lemma 3.3 for a more general version of the latter statement. (In passing, we note that the problem of minimizing the so-called Bayes risk [29] can be expressed as in (1.4). In this case, c_0 is called a "loss" function rather than a "cost.")

Here, we are in fact interested in a *constrained* version of (1.1) which is essentially the typical allocation problem in economics [6, 41], namely,

$$(1.6) \quad \text{minimize } \int_X c_0(x, f(x)) \nu(dx)$$

over the set of deterministic strategies $f \in \mathbb{F}_0$ for which

$$(1.7) \quad \int_X c_i(x, f(x)) \nu(dx) \leq b_i \quad \forall i = 1, \dots, n,$$

where $c_1, \dots, c_n$ are given cost functions on $X \times A$, and $b_1, \dots, b_n$ are given numbers. The corresponding relaxation is obtained as in (1.3), (1.4), replacing $f \in \mathbb{F}_0$ with $\varphi \in \Phi_0$; in particular, (1.7) becomes

$$(1.8) \quad \int_X c_i(x, \varphi(x)) \nu(dx) \leq b_i \quad \forall i = 1, \dots, n, \quad \varphi \in \Phi_0.$$

Contributions of this paper. The remainder of this paper is divided in two main parts. The first one, which consists of sections 2 and 3, concerns the *characterization* of the extreme points of the family, call it Δ , of randomized strategies $\varphi \in \Phi_0$ that satisfy (1.8). The main result in this part, Theorem 2.6, gives conditions under which a randomized strategy φ in $ex(\Delta)$ is a *randomization of at most $n+1$ functions f_j* in $\mathbb{F}_0$; i.e., there is a positive integer $m \leq n+1$ such that

$$(1.9) \quad \varphi(\cdot | x) = \sum_{j=1}^m \alpha_j \delta_{f_j(x)}(\cdot) \quad \forall x \in X,$$

where $\alpha_1, \dots, \alpha_m$ are positive numbers with $\alpha_1 + \dots + \alpha_m = 1$.

The second part of the paper, sections 4, 5, and 6, deals with applications of Theorem 2.6. In section 4 we study the unconstrained problems (1.1) and (1.4), in which one can already appreciate the main technical tools to obtain (a) and (b) above. For instance, to get (a) it all boils down to reducing (1.4) to a problem of minimizing a lower semicontinuous (l.s.c.) function (see (4.22) or (4.23)) on a (relatively) compact set of p.m.'s. On the other hand, to get (b) one can either use an “elimination of randomization” result (Lemmas 4.8 and 4.9) or directly analyze the set of optimal solutions for (1.4) (see Theorem 4.12). Essentially the same approach and tools are used in section 5 to study the *constrained* problem (1.6), (1.7) and its relaxation.

In section 6 we turn our attention to a class of constrained stochastic control problems called *Markov control processes* (MCPs). It turns out that a constrained MCP can be transformed into an optimization problem with an objective function as in (1.4), with constraints of the form (1.8) together with an additional, suitable, constraint (see (6.7)). We then use the approach briefly described in the previous paragraph to show, under very mild assumptions, the existence of optimal randomized strategies of the form (1.9).

Related literature. Our Theorem 2.6 extends to randomized strategies a result by Winkler [40, Theorem 2.1], which in turn is an extension of Karr's [24, Theorem 2.1]. Winkler considers a set P of p.m.'s on a measurable space $(S, \mathcal{S})$, and the subset P_0 of p.m.'s μ in P for which

$$(1.10) \quad \int_S |g_i| d\mu < \infty \quad \text{and} \quad \int_S g_i d\mu \leq b_i \quad \forall i = 1, \dots, n,$$

where $g_1, \dots, g_n$ are given measurable functions on S , and $b_1, \dots, b_n$ are fixed real numbers. He then proceeds to characterize the extreme points of P_0 assuming that P is a (Choquet-) simplex of p.m.'s whose extreme points are Dirac measures. Here, we show that the set Φ_0 of randomized strategies is a simplex whose set of extreme points is $\mathbb{F}_0$ (Lemma 3.3), and so specializing our Theorem 2.6 to “constant” kernels $\varphi(\cdot|x) \equiv \mu(\cdot)$ we obtain Winkler’s, and hence Karr’s, main results. For applications of Winkler’s Theorem 2.1 to problems in optimal control and decision analysis see [8] and [36], respectively. The latter reference, and also (e.g.) [11], [26], considers problems in which, for some given function g_0 on S , one wishes to optimize integrals $\int g_0 d\mu$ over all μ in P_0 . These problems include versions of the well-known “Markov moment problem” [26].

On the other hand, the allocation problem (1.6), (1.7) is well known in the economics literature—see [6, 41], for instance. Our approach here, via its Young relaxation and Theorem 2.6, is new.

Finally, our main result (Theorem 6.2) on constrained MCPs extends some facts for MCPs with a *countable* state space X [3, 14]. In this paper, the state and control (or action) spaces X and A are both *Borel* spaces (a Borel space is a Borel subset of a *Polish*—i.e., a complete and separable metric—space) and, moreover, all of the cost functions $c_0, \dots, c_n$ can be *unbounded*. Hence our Theorem 6.2 also extends recent results in, e.g., [28, Theorem 8] in which X is a Borel space, but the control set A is required to be *compact*, the constraint functions $c_1, \dots, c_n$ are *bounded*, and, in addition, either c_0 is *bounded* or a *Slater condition* is satisfied. Our hypotheses are considerably weaker (see Assumption 6.1). For results related to our Theorem 6.2 see [12, 13].

We conclude in section 7 with some comments on related open problems.

2. Constrained sets of randomized strategies. In many applications—e.g., in optimal control and game theory—the feasible actions $a \in A$ depend on the current state $x \in X$; that is, a is restricted to a subset $A(x)$ of A . Therefore, in most of what follows, instead of the sets $\mathbb{F}_0$ and Φ_0 in section 1 we shall consider subsets $\mathbb{F} \subset \mathbb{F}_0$ and $\Phi \subset \Phi_0$ defined below. Moreover, we need Φ to be a convex set in a suitable linear space, and so we will also consider *signed* transition measures or “signed kernels” (as in [20, section 7.2] or [35], for instance).

Let X and A be topological spaces with Borel σ -algebras $\mathcal{B}(X)$ and $\mathcal{B}(A)$, respectively. For each $x \in X$, we denote by $A(x) \in \mathcal{B}(A)$ the (nonempty) set of feasible actions in x . We shall assume that the set

$$(2.1) \quad \mathbb{K} := \{(x, a) | x \in X, a \in A(x)\}$$

of feasible state-action pairs is a *closed* (hence a Borel) subset of $X \times A$, and that, in addition, $\mathbb{K}$ contains the graph of a measurable function from X to A . (The latter condition ensures that the set $\mathbb{F}$ in Definition 2.2, below, is nonempty.) Observe that if $A(x) = A$ for all $x \in X$, then $\mathbb{K} = X \times A$.

DEFINITION 2.1. A randomized strategy φ (restricted to $\mathbb{K}$) is a mapping $(x, D) \mapsto \varphi(D|x)$ from $X \times \mathcal{B}(A)$ to the interval $[0, 1]$ such that

- (a) for every fixed $x \in X$, $\varphi(\cdot|x)$ is a p.m. on $\mathcal{B}(A)$ concentrated on $A(x)$, i.e.,

$$(2.2) \quad \varphi(A(x)^c|x) = 0 \quad \forall x \in X,$$

where $A(x)^c$ denotes the complement of $A(x)$; and

- (b) $\varphi(D|\cdot)$ is a measurable function on X for every fixed $D \in \mathcal{B}(A)$.

We shall denote by Φ the family of randomized strategies (restricted to $\mathbb{K}$).

We will also consider the convex cone $\mathcal{D}_+ := \mathbb{R}_+ \cdot \Phi$ of so-called *transition measures* (restricted to $\mathbb{K}$), and the linear space $\mathcal{D} := \mathcal{D}_+ - \mathcal{D}_+$ (with the obvious definitions of sum and scalar multiplication) of *signed transition measures*.

DEFINITION 2.2. $\mathbb{F}$ denotes the family of measurable functions f from X to A such that $f(x)$ is in $A(x)$ for all $x \in X$. The functions in $\mathbb{F}$ are called *deterministic strategies from X to A (restricted to $\mathbb{K}$)*. (A function in $\mathbb{F}$ is sometimes called a *selector of the set-valued mapping $x \mapsto A(x)$* .)

If $A(x) = A$ for all $x \in X$, then $\mathbb{K} = X \times A$ and the sets Φ and $\mathbb{F}$ coincide, respectively, with the sets Φ_0 and $\mathbb{F}_0$ in section 1. Further, as in (1.2), identifying $f(x) \in A(x)$ with $\varphi(\cdot|x) := \delta_{f(x)}(\cdot)$, we have

$$(2.3) \quad \mathbb{F} \subset \Phi.$$

Incidentally, if $A(x) = A$ for all $x \in X$ and $A = X$, then a randomized strategy is usually called a *Markov transition probability*.

DEFINITION 2.3. A randomized strategy $\varphi \in \Phi$ is said to be a *randomization of at most $n+1$ deterministic strategies, or simply an $(n+1)$ -randomization*, if there exist a positive integer $m \leq n+1$, functions $f_1, \dots, f_m$ in $\mathbb{F}$, and positive numbers $\alpha_1, \dots, \alpha_m$ with $\alpha_1 + \dots + \alpha_m = 1$ for which

$$(2.4) \quad \varphi(D|x) = \sum_{j=1}^m \alpha_j \delta_{f_j(x)}(D) \quad \forall D \in \mathcal{B}(A), x \in X.$$

We denote by $\mathcal{R}_{n+1}$ the family of $(n+1)$ -randomizations.

We need the following regularity concepts that are briefly discussed in Remarks 2.5 and 3.4. (For further details see, e.g., Topsoe [37, section 1] or Vakhania, Tarieladze and Chobanyan [38, section I.3].)

DEFINITION 2.4. Let μ be a finite (nonnegative) measure on $\mathcal{B}(A)$. Then μ is said to be

- (a) *regular* if $\mu(D) = \sup\{\mu(F) | F \subset D \text{ is closed}\}$ for each Borel set $D \in \mathcal{B}(A)$.
- (b) τ -*smooth* if for every decreasing net $\{F_\alpha\}$ of closed subsets of A we have $\mu(\cap_\alpha F_\alpha) = \inf_\alpha \mu(F_\alpha)$.

REMARK 2.5. (a) If A is a Hausdorff (or T_2) space, then every Radon measure on A is τ -smooth, and if A is regular (or T_3), then every τ -smooth measure is regular (see, e.g., Proposition I.3.1 in [38]).

(b) If A is strongly Lindelöf (which is the case, e.g., if A is a Suslin space—see Schwartz [34, p. 104]), then every finite measure on $\mathcal{B}(A)$ is τ -smooth. The latter fact and (a) yield, in particular, the following.

(c) As an example, if A is a locally compact and separable metric space, then each p.m. on $\mathcal{B}(A)$ is both τ -smooth and regular. (See also Remark 3.4.)

We can finally state our first main result, which is proved in section 3, and applied to optimization and control problems in sections 4, 5, 6. We will use the notation (1.3) replacing Φ_0 and A with Φ and $A(x)$, respectively. Concerning the hypothesis on the space A in the following theorem see Remark 3.4.

THEOREM 2.6. Let X be an arbitrary topological space, and A a topological space such that every p.m. on $\mathcal{B}(A)$ is τ -smooth and regular. Let $\mathbb{K}$ be as in (2.1). Fix an arbitrary p.m. ν on $\mathcal{B}(X)$, real-valued measurable functions $c_1, \dots, c_n$ on $\mathbb{K}$, and real numbers $b_1, \dots, b_n$. Consider the set $\Delta \subset \Phi$ that consists of all the randomized

strategies $\varphi \in \Phi$ for which

$$(2.5) \quad \int_X \int_{A(x)} |c_i(x, a)| \varphi(da|x) \nu(dx) < \infty \quad \forall i = 1, \dots, n$$

and

$$(2.6) \quad \int_X c_i(x, \varphi(x)) \nu(dx) \leq b_i \quad \forall i = 1, \dots, n.$$

Let $ex(\Delta)$ be the set of extreme points of Δ . Then

(a) Δ is convex and

$$(2.7) \quad ex(\Delta) \subset \mathcal{R}_{n+1}^0,$$

where $\mathcal{R}_{n+1}^0$ is the set of all the $(n+1)$ -randomizations $\varphi \in \mathcal{R}_{n+1}$ as in (2.4) for which the vectors

$$(2.8) \quad \left(\int c_1(x, f_j(x)) \nu(dx), \dots, \int c_n(x, f_j(x)) \nu(dx), 1 \right) \in \mathbb{R}^{n+1}$$

for $j = 1, \dots, m$ are linearly independent.

(b) If equality holds in (2.6), then the sets in (2.7) are equal.

REMARK 2.7. A randomized strategy φ in Φ (or in Φ_0) is said to be constant if there is a p.m. μ on A such that $\varphi(\cdot|x) \equiv \mu(\cdot)$ for all $x \in X$. Now let $g_1, \dots, g_n$ be given measurable functions on A , and suppose that the functions c_i in (2.5), (2.6) are of the form $c_i(x, a) := g_i(a)$ for all $i = 1, \dots, n$. Then comparing (2.5), (2.6) with (1.10), we can view Winkler's [40, Theorem 2.1] as a special case of our Theorem 2.6 restricted to constant kernels, and similarly for Theorem 2.1 in Karr [24].

3. Proof of Theorem 2.6.

We begin by noting the following.

REMARK 3.1. The convex set Φ of randomized strategies from X to A (restricted to $\mathbb{K}$) is the base of the convex cone $\mathcal{D}_+ = \mathbb{R}_+ \cdot \Phi$; that is, for each $q \in \mathcal{D}_+$ there exist a unique $\alpha \geq 0$ and $\varphi \in \Phi$ such that $q = \alpha\varphi$. On the other hand, the linear space $\mathcal{D} = \mathcal{D}_+ - \mathcal{D}_+$ with the partial order induced by $\mathcal{D}_+$ (i.e., $q_1 \geq q_2$ if and only if $q_1 - q_2$ is in $\mathcal{D}_+$) is a vector lattice.

DEFINITION 3.2. (see Phelps [31, p. 59]) Let S be a convex subset of a real linear space L , and suppose that S is contained in a closed hyperplane that misses the origin. Then S is called a (Choquet-)simplex if it is the base of a cone $\tilde{S}$ such that the space $\tilde{S} - \tilde{S}$ is a vector lattice in the ordering induced by $\tilde{S}$.

The above definition of a simplex coincides with the usual one in case S is finite dimensional. (See [31, Proposition 9.11].)

LEMMA 3.3. The convex set Φ is a simplex, and the set of its extreme points coincides with $\mathbb{F}$, i.e.,

$$(3.1) \quad ex(\Phi) = \mathbb{F}.$$

Proof. The first statement follows from Remark 3.1 (and Definition 3.2). To prove (3.1) we will show in fact that

$$(3.2) \quad \mathbb{F} = \Phi^d = ex(\Phi),$$

where $\Phi^d \subset \Phi$ is the set of randomized strategies $\varphi \in \Phi$ such that $\varphi(\cdot|x)$ is a Dirac measure for each $x \in X$. Clearly $\mathbb{F} \subset \Phi^d \subset ex(\Phi)$, and so to get (3.2) it remains to show that

$$(3.3) \quad ex(\Phi) \subset \Phi^d \subset \mathbb{F}.$$

To this end, suppose first that $ex(\Phi)$ is not contained in Φ^d , and let φ be a randomized strategy in $ex(\Phi)$ that is not in Φ^d . Then, by the hypothesis on A and Theorem 11.1 in Topsoe [37], there exists $z \in X$ such that $\varphi(\cdot|z)$ is not a Dirac measure; that is, there is a set C in $\mathcal{B}(A)$ with $0 < \varphi(C|z) < 1$. Now let $\beta := \varphi(C|z)$, and for each $x \in X$ and $D \in \mathcal{B}(A)$ define

$$(3.4) \quad \begin{aligned} \varphi_1(D|x) &:= \varphi(D|x) \text{ if } x \neq z \\ &:= \varphi(D \cap C|x)/\beta \text{ if } x = z \end{aligned}$$

and

$$(3.5) \quad \begin{aligned} \varphi_2(D|x) &:= \varphi(D|x) \text{ if } x \neq z \\ &:= \varphi(D \cap C^c|x)/(1-\beta) \text{ if } x = z, \end{aligned}$$

where C^c stands for the complement of C . We have thus stochastic kernels φ_1 and φ_2 such that $\varphi = \beta\varphi_1 + (1-\beta)\varphi_2$, and, therefore, φ is not an extreme point of Φ . This contradiction yields the first relation in (3.3).

To prove the second relation in (3.3) choose an arbitrary φ in Φ^d . It follows that for each $x \in X$ there exists $a \in A$ such that $\varphi(D|x) = \delta_a(D)$ for all $D \in \mathcal{B}(X)$ and, moreover, by (2.2), a is in $A(x)$. Writing a as $f(x)$ we thus obtain a function f from X to A , with $f(x)$ in $A(x)$ for each $x \in X$, and so to prove that f is in $\mathbb{F}$ we need to show that f is measurable. This, however, is obvious because by part (b) of Definition 2.1, for any set $D \in \mathcal{B}(A)$

$$f^{-1}(D) = \{x \in X | \varphi(D|x) = 1\} = \varphi^{-1}(D|\cdot)(1)$$

is in $\mathcal{B}(X)$. Hence, as $\varphi \in \Phi^d$ was arbitrary, we conclude that $\Phi^d \subset \mathbb{F}$. This completes the proof of (3.3), and (3.1) follows. $\square$

Completion of the proof. Having Lemma 3.3, the proof of Theorem 2.6 can now be completed exactly as the proof of Winkler's [40, Theorem 2.1], with the appropriate changes. Namely, we define

- a convex set $C \subset \Phi$ as the collection of randomized strategies $\varphi \in \Phi$ that satisfy (2.5);
- an affine map $J : C \rightarrow \mathbb{R}^n$ as $J(\varphi) = (J_1(\varphi), \dots, J_n(\varphi))$ with components

$$(3.6) \quad J_i(\varphi) := \int_X c_i(x, \varphi(x)) \nu(dx) \text{ for } i = 1, \dots, n;$$

- and a convex set $W \subset J(C)$ as

$$W := J(C) \bigcap \prod_{i=1}^n (-\infty, b_i] \text{ for part (a),}$$

$$W := J(C) \bigcap \{(b_1, \dots, b_n)\} \text{ for part (b).}$$

Moreover, we note that if $\varphi - \varphi'$ is in the cone $\mathbb{R}_+ \cdot C$, with φ in $\mathbb{R}_+ \cdot C$ and φ' in $\mathcal{D}_+ = \mathbb{R}_+ \cdot \Phi$, then

$$\int \int |c_i(x, a)| \varphi'(da|x) \nu(dx) \leq \int \int |c_i(x, a)| \varphi(da|x) \nu(dx) < \infty$$

for each $i = 1, \dots, n$, and so φ' is in $\mathbb{R}_+ \cdot C$; in other words, $\mathbb{R}_+ \cdot C$ is a hereditary subcone of $\mathcal{D}_+$. This implies (by Lemma 9.4 in [31]) that $\mathbb{R}_+ \cdot C$ is a lattice cone in its own order, which in turn (by the main theorem and paragraph 4 (p. 369) in Kendall [25]) yields that C is linearly compact. Therefore, the conclusions (a) and (b) in Theorem 2.6 follow from Proposition 2.1 in Winkler [40]. (The number 1 that appears in the vectors (2.8) ensures linear independence rather than the affine independence in [40, Proposition 2.1]; see also Karr [24, Theorem 2.1].) $\square$

REMARK 3.4. (a) *Theorem 2.6 requires the action set A to be a topological space such that*

$$(3.7) \quad \text{every p.m. on } \mathcal{B}(A) \text{ is } \tau\text{-smooth and regular.}$$

This condition ensures that, as in the proof of Lemma 3.3, the set of extreme points of $\mathbb{P}(A)$, the space of p.m.'s on A , coincides with the set of Dirac measures δ_a for all $a \in A$ [37, Theorem 11.1]. Without (3.7) we can only guarantee that the extreme points of $\mathbb{P}(A)$ are zero-one p.m.'s [1], that is, nontrivial p.m.'s μ such that $\mu(B) = 0$ or 1 for all $B \in \mathcal{B}(A)$. For examples of zero-one p.m.'s that are not Dirac measures see [15] or the references in [1]. On the other hand, we can obtain (3.7) as follows (see also Remark 2.5).

(b) *A p.m. μ on A is said to be tight if for every $\varepsilon > 0$ there exists a compact subset K of A such that $\mu(K) \geq 1 - \varepsilon$. Moreover, if a p.m. is tight, then it is τ -smooth and regular (see p. xiii in [37]). It follows that to obtain (3.7) it suffices to give conditions on A so that every p.m. on $\mathcal{B}(A)$ is tight. This is the case if, for instance, A is (i) a σ -compact Hausdorff space, or (ii) a Polish space, or (iii) a locally compact separable metric space. (See [2, 29, 34, 37].)*

4. The unconstrained optimization problems. In this section we consider the unconstrained problem (1.1) and its Young relaxation (1.4) but replacing $\mathbb{F}_0$ and Φ_0 with the sets $\mathbb{F}$ and Φ in Definitions 2.2 and 2.1, respectively. For notational ease, let (as in (3.6))

$$(4.1) \quad J_0(\varphi) := \int_X c_0(x, \varphi(x)) \nu(dx) \quad \forall \varphi \in \Phi \supset \mathbb{F}, \quad x \in X,$$

where (as in (1.3))

$$(4.2) \quad c_0(x, \varphi(x)) := \int_{A(x)} c_0(x, a) \varphi(da|x).$$

Thus the problems addressed in this section are (using the abbreviation “s.t.” for “subject to”)

$$(4.3) \quad \text{minimize } J_0(f) \quad \text{s.t. } f \in \mathbb{F},$$

$$(4.4) \quad \text{minimize } J_0(\varphi) \quad \text{s.t. } \varphi \in \Phi.$$

If $A(x) = A$ for all $x \in X$, then (4.3), (4.4) reduce to the problems (1.1), (1.4).

To fix ideas we shall impose the following assumptions but in Remark 4.10 we mention other possibilities.

ASSUMPTION 4.1. (a) *The spaces X and A are both Borel spaces. (Recall that the set $\mathbb{K}$ in (2.1) is supposed to be closed.)*

(b) *c_0 is nonnegative.*

(c) *c_0 is inf-compact; i.e., the level set $\{(x, a) \in \mathbb{K} | c_0(x, a) \leq r\}$ is compact for each $r \in \mathbb{R}$.*

Let ρ^* and $\rho^\#$ be the values of problems (4.3) and (4.4), respectively; that is,

$$(4.5) \quad \rho^* := \inf \{J_0(f) | f \in \mathbb{F}\}$$

and similarly for $\rho^\#$. By (2.3),

$$(4.6) \quad \rho^\# \leq \rho^*.$$

The next theorem shows, in particular, that the equality holds in (4.6).

THEOREM 4.2. *Under Assumption 4.1,*

(a) *the unconstrained relaxed problem (4.4) is solvable; i.e., there exists φ^* in Φ such that*

$$(4.7) \quad \rho^\# = J_0(\varphi^*) = \int_X c_0(x, \varphi^*(x)) \nu(dx);$$

(b) *(4.3) is also solvable and, moreover, $\rho^* = \rho^\#$.*

Before proving Theorem 4.2 we will recall some well-known concepts and facts from probability theory, which are stated here for completeness and ease of reference.

DEFINITION 4.3. *Let Y be a metric space, and $\mathbb{P}(Y)$ the family of p.m.'s on Y . A subset M of $\mathbb{P}(Y)$ is said to be*

(a) *tight if for each $\epsilon > 0$ there is a compact set $K = K_\epsilon$ in Y such that*

$$\inf_{\mu \in M} \mu(K) \geq 1 - \epsilon;$$

(b) *relatively compact if for each sequence $\{\mu_n\}$ in M there is a subsequence $\{\mu_m\}$ of $\{\mu_n\}$ and a p.m. μ on Y (but not necessarily in M) such that μ_m converges weakly to μ , i.e.,*

$$(4.8) \quad \int_Y u d\mu_m \rightarrow \int_Y u d\mu \quad \forall u \in C_b(Y),$$

where $C_b(Y)$ denotes the space of continuous bounded functions on Y .

Throughout the following $\mathbb{P}(Y)$ is endowed with the topology of weak convergence (as in (4.8)). It follows in particular that if Y is a Borel space, then so is $\mathbb{P}(Y)$.

By Prohorov's Theorem [9, 32, 35],

$$(4.9) \quad \text{tightness} \Rightarrow \text{relative compactness}.$$

(The converse of (4.9) holds if Y is a Polish space.) The Assumption 4.1—in particular the inf-compactness condition in part (c)—will be used in combination with the following lemma.

LEMMA 4.4. *Let Y and $M \subset \mathbb{P}(Y)$ be as in Definition 4.3. If there exists an inf-compact function $u : Y \rightarrow \mathbb{R}_+$ such that*

$$\sup_{\mu \in M} \int_Y u d\mu < \infty,$$

then M is tight, and hence (by (4.9)) relatively compact.

Lemma 4.4 is a direct consequence of the definitions of inf-compactness and tightness. (Again, the converse of the first statement in Lemma 4.4 holds under suitable assumptions—e.g., if the metric space Y is σ -compact.)

LEMMA 4.5. *Let Y be a metric space, and $u : Y \rightarrow \mathbb{R}$ l.s.c. and bounded below. If a sequence $\{\mu_n\}$ in $\mathbb{P}(Y)$ converges weakly to μ (i.e., as in (4.8)), then*

$$\liminf_{n \rightarrow \infty} \int_Y u \, d\mu_n \geq \int_Y u \, d\mu.$$

Lemma 4.5 is well known; see, e.g., statement (12.3.37) in [20].

If μ is a p.m. on $X \times A$, we shall denote by $\hat{\mu}$ its *marginal* (or projection) on X , i.e., $\hat{\mu}(B) := \mu(B \times A)$ for all B in $\mathcal{B}(X)$.

LEMMA 4.6 (disintegration of measures). *Let X , A , and $\mathbb{K}$ be as in Assumption 4.1(a). If μ is a p.m. on $X \times A$ concentrated on $\mathbb{K}$ (i.e., $\mu(\mathbb{K}^c) = 0$), then there exists a randomized strategy $\varphi \in \Phi$ such that*

$$(4.10) \quad \mu(B \times D) = \int_B \varphi(D|x) \hat{\mu}(dx) \quad \forall B \in \mathcal{B}(X), D \in \mathcal{B}(A).$$

For a proof of Lemma 4.6 see, e.g., Hinderer [23, p. 89]. Note, on the other hand, that if φ is in Φ and ν is a p.m. on X , then

$$(4.11) \quad \mu(B \times D) := \int_B \varphi(D|x) \nu(dx) \quad \forall B \in \mathcal{B}(X), D \in \mathcal{B}(A)$$

defines a p.m. on $X \times A$ that (by (2.2)) is concentrated on $\mathbb{K}$. The measures μ in (4.10) and (4.11) will be written as $\mu = \hat{\mu} \cdot \varphi$ and $\mu = \nu \cdot \varphi$, respectively. (The p.m. $\nu \cdot \varphi$ is sometimes called the ν -mixture of φ [35].)

DEFINITION 4.7. $\mathbb{P}(\mathbb{K})$ denotes the family of p.m.'s on $X \times A$ that are concentrated on $\mathbb{K}$. If ν is a p.m. on X and Φ' is a subfamily of randomized strategies in Φ , then

$$\nu \cdot \Phi' := \{\nu \cdot \varphi \mid \varphi \in \Phi'\} \subset \mathbb{P}(\mathbb{K}).$$

Throughout the following we use a notation similar to (4.2), namely,

$$(4.12) \quad v(x, \varphi(x)) := \int_{A(x)} v(x, a) \varphi(da|x) \quad \forall \varphi \in \Phi, x \in X,$$

for any measurable (possibly extended) real-valued function v for which the integral in (4.12) is well defined. Moreover, as usual, $v^- := -\min\{v, 0\} \geq 0$ stands for the “negative part” of v .

LEMMA 4.8 (see [23, Lemma 15.1]). *Suppose that X and A are as in Assumption 4.1(a). Let φ be a randomized strategy in Φ , and v an extended real-valued measurable function on $\mathbb{K}$. If the function $x \mapsto v^-(x, \varphi(x))$ on X is finite-valued, then there exists a deterministic strategy $f \in \mathbb{F}$ such that*

$$(4.13) \quad v(x, \varphi(x)) \geq v(x, f(x)) \quad \forall x \in X.$$

In statistical decision theory, a result such as Lemma 4.8 is called an *elimination of randomization* theorem [7, 10, 39] because of (4.13). We will use Lemma 4.8 in the proof of Theorem 4.2, which concerns an unconstrained problem. In the *constrained*

case, however, we shall need a different setting, as in the following “multidimensional” elimination of randomization result.

LEMMA 4.9 (see [7, Theorem 2.3 and Corollary 2.5]). *Let $(X, \mathcal{X}, \nu)$ be a probability space in which the p.m. ν is nonatomic, and let A be a Polish space. Suppose that $A(x) = A$ for all $x \in X$, so that the set $\mathbb{K}$ in (2.1) coincides with $X \times A$ and, moreover, $\mathbb{F} = \mathbb{F}_0$ and $\Phi = \Phi_0$.*

- (a) *If the functions $c_0, \dots, c_n$ in (1.6)–(1.8) are nonnegative and each $c_i(x, \cdot)$ is l.s.c. on A for each $x \in X$, then for every $\varphi \in \Phi_0$ there exists $f \in \mathbb{F}_0$ such that*

$$(4.14) \quad \int_X c_i(x, \varphi(x)) \nu(dx) \geq \int_X c_i(x, f(x)) \nu(dx) \quad \forall i = 0, 1, \dots, n.$$

- (b) *If the functions $c_0, \dots, c_n$ are uniformly bounded and each $c_i(x, \cdot)$ is continuous on A for each $x \in X$, then for every $\varphi \in \Phi_0$ there exists $f \in \mathbb{F}_0$ for which (4.14) holds with equality.*

Proof of Theorem 4.2. Theorem 4.2 is trivial if $J_0(\varphi) = +\infty$ for all φ in Φ . Hence, without loss of generality, we may—and will—assume that

$$J_0(\widehat{\varphi}) = \int_X c_0(x, \widehat{\varphi}(x)) \nu(dx) < \infty$$

for some randomized strategy $\widehat{\varphi}$ in Φ . Let $\rho^\# := \inf\{J_0(\varphi) | \varphi \in \Phi\}$ be the value of problem (4.4), and let $\{\varphi_n\} \subset \Phi$ be a minimizing sequence, i.e.,

$$(4.15) \quad J_0(\varphi_n) \downarrow \rho^\#.$$

Thus for each $\epsilon > 0$ there exists $N(\epsilon)$ such that

$$(4.16) \quad J_0(\varphi_n) \leq \rho^\# + \epsilon \quad \forall n \geq N(\epsilon).$$

For each n , let $\mu_n = \nu \cdot \varphi_n$ be the p.m. in $\mathbb{P}(\mathbb{K})$ defined as in (4.11), i.e.,

$$(4.17) \quad \mu_n(B \times D) := \int_B \varphi_n(D|x) \nu(dx)$$

for all $B \in \mathcal{B}(X)$, $D \in \mathcal{B}(A)$, and $n = 1, 2, \dots$. Therefore we can express $J_0(\varphi_n)$ as

$$(4.18) \quad J_0(\varphi_n) = \int_{\mathbb{K}} c_0(x, a) \mu_n(d(x, a)).$$

We can now see (from (4.16), Assumption 4.1—in particular part (c)—and Lemma 4.4) that

$$(4.19) \quad \text{the sequence } M := \{\mu_n, n \geq N(\epsilon)\} \text{ is relatively compact.}$$

Therefore there is a subsequence $\{\mu_m\}$ of M and a p.m. μ^* on $X \times A$, which is concentrated on $\mathbb{K}$ (by the closedness of $\mathbb{K} \subset X \times A$), i.e., μ^* is in $\mathbb{P}(\mathbb{K})$, such that

$$(4.20) \quad \mu_m \rightarrow \mu^* \text{ weakly,}$$

that is, as in (4.8). In turn (4.20) implies the following.

- (i) The marginals $\hat{\mu}_m \rightarrow \hat{\mu}^*$ weakly; hence $\hat{\mu}^* = \nu$ because (by (4.17)) $\hat{\mu}_m = \nu$ for all m . Thus, by Lemma 4.6, there exists φ^* in Φ such that

$$(4.21) \quad \mu^* = \hat{\mu}^* \cdot \varphi^* = \nu \cdot \varphi^*.$$

- (ii) As Assumptions 4.1(b), (c) yield that c_0 is nonnegative and l.s.c., from (4.20) and Lemma 4.5 we have that

$$(4.22) \quad \liminf_{n \rightarrow \infty} \int c_0 d\mu_m \geq \int c_0 d\mu^*;$$

equivalently, by (4.21) and (4.18),

$$(4.23) \quad \liminf_{m \rightarrow \infty} J_0(\varphi_m) \geq J_0(\varphi^*) \quad (\geq \rho^\#).$$

This concludes the proof of part (a) because (4.23) and (4.15) give (4.7).

- (b) By (4.7) and Lemma 4.8, there exists f^* in $\mathbb{F}$ such that

$$\rho^\# \geq \int_X c_0(x, f^*(x)) \nu(dx) = J_0(f^*) \geq \rho^*.$$

The latter inequality and (4.6) yield that $f^* \in \mathbb{F}$ is an optimal solution for problem (4.3) and that $\rho^* = \rho^\#$. This completes the proof of Theorem 4.2. $\square$

REMARK 4.10. *It is clear that the nonnegativity of c_0 in Assumption 4.1(b) can be replaced with the following: c_0 is bounded below. It is just as clear that the inf-compactness of c_0 (Assumption 4.1(c)) was used in the above proof to get that (a) c_0 is l.s.c., which, together with (4.20), gives (4.22)–(4.23); and (b) to obtain (4.19). The condition (a) seems to be unavoidable: either we assume it holds or we get it from some stronger condition, as in our case. On the other hand, to get the relative compactness in (b) we can replace the inf-compactness of c_0 with, for instance, the following condition:*

- (b.1) *There exists $\hat{\varphi}$ in Φ such that the set $\{\mu = \nu \cdot \varphi \mid J_0(\varphi) \leq J_0(\hat{\varphi})\} \subset \mathbb{P}(\mathbb{K})$ is tight (or just relatively compact; recall (4.9)).*

Of course, a trivial case to get (b) would be

- (b.2) *X and A are both compact metric spaces,*

because then the closed set $\mathbb{K} \subset X \times A$ would be compact, and so would be $\mathbb{P}(\mathbb{K})$ in the weak topology defined by (4.8).

Remark 4.10, with obvious changes, is valid for the problems addressed in sections 5 and 6.

In the following sections we use the fact that part (b) in Theorem 4.2 can be obtained from part (a) and Lemma 3.3. To show that this is indeed the case we need to introduce some terminology and a preliminary result.

Let $\mathcal{D}$ be the linear space introduced after Definition 2.1, and $M(\mathbb{K})$ the linear space of finite signed measures on $X \times A$ concentrated on $\mathbb{K}$. Consider the linear mapping $\ell : \mathcal{D} \rightarrow M(\mathbb{K})$ defined by $\ell(\varphi) := \nu \cdot \varphi$ and the quotient space $\overline{\mathcal{D}} := \mathcal{D} / \text{Ker}(\ell)$, where $\text{Ker}(\ell) := \{\varphi \in \mathcal{D} \mid \ell(\varphi) = 0\}$ is the kernel of ℓ . For each $\varphi \in \mathcal{D}$, let $\overline{\varphi} := \{\varphi' \in \mathcal{D} \mid \nu \cdot \varphi' = \nu \cdot \varphi\}$ be the corresponding equivalence class in $\overline{\mathcal{D}}$. The quotient sets $\overline{\mathbb{F}}, \overline{\Phi}, \overline{\Delta}, \overline{\mathcal{R}}_{n+1}^0$ are defined similarly. For instance, $\overline{\Delta} := \{\overline{\varphi} \mid \varphi \in \Delta\}$.

LEMMA 4.11. (a) *If $\nu \cdot \varphi$ is an extreme point of $\nu \cdot \Phi$ (resp., $\nu \cdot \Delta$), then $\overline{\varphi}$ is an extreme point of $\overline{\Phi}$ (resp., $\overline{\Delta}$).*

- (b) *If $\overline{\varphi}$ is an extreme point of $\overline{\Phi}$, then $\overline{\varphi}$ has an extreme point f in $\mathbb{F} \cap \overline{\varphi}$.*

(c) If $\bar{\varphi}$ is an extreme point of $\bar{\Delta}$, with Δ as in Theorem 2.6, then $\bar{\varphi}$ has an extreme point φ^* of Δ with φ^* in $\bar{\mathcal{R}}_{n+1}^0 \cap \bar{\varphi}$.

Proof. (a) Let $\nu \cdot \varphi$ be an extreme point of $\nu \cdot \Phi$ and suppose that $\bar{\varphi} = \beta \bar{\varphi}_1 + (1-\beta) \bar{\varphi}_2$ for some $\beta \in (0, 1)$ and $\bar{\varphi}_1, \bar{\varphi}_2$ in $\bar{\Phi}$. It follows that $\nu \cdot \varphi = \beta \nu \cdot \varphi_1 + (1-\beta) \nu \cdot \varphi_2$ and by the extremality of $\nu \cdot \varphi$ we get $\nu \cdot \varphi_1 = \nu \cdot \varphi_2 = \nu \cdot \varphi$. Hence $\bar{\varphi}_1 = \bar{\varphi}_2 = \bar{\varphi}$. The same argument holds if Φ is replaced with Δ .

(b) This part follows from Theorem 2 in [32], or Theorem 10 in [33, pp. 83–89].

(c) This is obtained exactly as the proof of Theorem 2.6 but dealing with equivalence classes $\bar{\varphi} \in \bar{\Delta}$ rather than randomized strategies $\varphi \in \Delta$. $\square$

Now let C be a convex set, and recall that an *extremal subset* of C is any nonempty subset F of C with the property that if $x = \alpha y + (1-\alpha)z$ is in F , where $0 < \alpha < 1$ and $y, z \in C$, then y, z are in F . A convex extremal subset of C is called a *face* of C [2]. Finally, let

$$(4.24) \quad \Phi^\# := \{\varphi \in \Phi \mid J_0(\varphi) = \rho^\#\}$$

be the set of optimal solutions for problem (4.4), and let $\nu \cdot \Phi^\# \subset \mathbb{P}(\mathbb{K})$ be as in Definition 4.7, i.e., $\nu \cdot \Phi^\# := \{\nu \cdot \varphi \mid \varphi \in \Phi^\#\}$. By Theorem 4.2(a), $\nu \cdot \Phi^\#$ is nonempty. We also have the following.

THEOREM 4.12. *Under Assumption 4.1, the set $\nu \cdot \Phi^\#$ is*

(a) *a face of $\nu \cdot \Phi$.*

Moreover, $\nu \cdot \Phi^\#$ is

(b) *relatively compact and*

(c) *weakly closed (i.e., closed in the weak topology defined by (4.8)); hence*

(d) *$\nu \cdot \Phi^\#$ is weakly compact.*

(e) *$\nu \cdot \Phi^\#$ has an extreme point $\nu \cdot \varphi$ with φ in $\text{ex}(\Phi) = \mathbb{F}$; that is (by Lemma 3.3), $\varphi(\cdot|x) = \delta_{f^\#}(\cdot)$ for some $f^\#$ in $\mathbb{F}$, and $J_0(f^\#) = \rho^\# = \rho^*$.*

Proof. Part (a) follows from the linearity of the mapping $\varphi \mapsto J_0(\varphi)$, whereas (b) and (c), which give (d), can be obtained as in (4.19)–(4.21).

(e) By (a)–(d), $\nu \cdot \Phi^\#$ is a (weakly) compact extremal subset of $\nu \cdot \Phi$ and, therefore, $\nu \cdot \Phi^\#$ contains an extreme point, say $\nu \cdot \varphi$, of $\nu \cdot \Phi$ (see Lemma 5.114 in [2]). Consequently (by Lemma 4.11(a)), $\bar{\varphi}$ is an extreme point of $\bar{\Phi}$ and so (by Lemma 4.11(b)) there exists $f^\#$ in $\mathbb{F} \cap \bar{\varphi}$, that is, $\nu \cdot \varphi = \nu \cdot f^\#$. Finally, as $\rho^\# = J_0(f^\#) \geq \rho^*$, the last statement in (e) follows from (4.6). $\square$

REMARK 4.13. *In the proof of Theorems 4.2 and 4.12 we have shown “indirectly” the relative compactness of some subsets of $\nu \cdot \Phi$ via their relative compactness in $\mathbb{P}(\mathbb{K})$. However, there are ways of defining “directly” a topology on Φ with respect to which sets such as $\Phi^\#$ in (4.24) are compact; see, e.g., Yushkevich [43] and the references therein. (This typically requires the sets $A(x) \subset A$ to be compact.) Therefore, as usually it is not difficult to get that $\varphi \mapsto J_0(\varphi)$ is l.s.c. (see (4.22), (4.23)), the last part of both Theorems 4.2 and 4.12 would be a direct consequence of Lemma 3.3 and Bauer’s extremum principle, which states the following [2]: If $\mathbb{K}$ is a compact convex subset of a locally convex Hausdorff topological vector space, then every l.s.c. concave function on $\mathbb{K}$ achieves its minimum at an extreme point.*

5. The constrained optimization problems. We next consider the constrained optimization problems in (1.6)–(1.8) with $\mathbb{F}$ and Φ in lieu of $\mathbb{F}_0$ and Φ_0 , respectively. That is, extending the notation (4.1), (4.2) to all the cost functions $c_0, \dots, c_n$, we have the “deterministic” problem

$$(5.1) \quad \text{minimize } J_0(f) \text{ s.t. } J_i(f) \leq b_i \quad \forall i = 1, \dots, n, f \in \mathbb{F},$$

and its Young relaxation

$$(5.2) \quad \text{minimize } J_0(\varphi) \text{ s.t. } J_i(\varphi) \leq b_i \quad \forall i = 1, \dots, n, \varphi \in \Phi.$$

In the remainder of this section we suppose the following.

- ASSUMPTION 5.1. (a) X and A are Borel spaces (and $\mathbb{K} \subset X \times A$ is closed).
 (b) The functions $c_i : \mathbb{K} \rightarrow \mathbb{R}$ ($i = 0, 1, \dots, n$) are nonnegative and l.s.c.
 (c) At least one of the functions $c_0, \dots, c_n$ is inf-compact.
 (d) There exists $\hat{\varphi}$ in Φ such that $J_0(\hat{\varphi}) < \infty$ and $J_i(\hat{\varphi}) \leq b_i$ for all $i = 1, \dots, n$.
 Let ρ_c^* and $\rho_c^\#$ be the values of the problems (5.1) and (5.2), i.e.,

$$\rho_c^* := \inf \{ J_0(f) \mid J_i(f) \leq b_i \quad \forall i = 1, \dots, n; f \in \mathbb{F} \},$$

and similarly for $\rho_c^\#$. As in (4.6), in the present case we also have (by (2.3))

$$(5.3) \quad \rho_c^\# \leq \rho_c^*.$$

THEOREM 5.2. Under Assumption 5.1, we have the following.

- (a) The constrained problem (5.2) is solvable; i.e., there exists φ^* in Φ such that

$$(5.4) \quad J_0(\varphi^*) = \rho_c^\#$$

and

$$(5.5) \quad J_i(\varphi^*) \leq b_i \quad \forall i = 1, \dots, n.$$

Let $\Delta^* := \{ \varphi \in \Phi \mid \varphi \text{ satisfies (5.4), (5.5) } \}$ be the set of optimal solutions for problem (5.2).

- (b) If in addition A is a locally compact and separable metric space, then there exists $\varphi^* \in \Delta^*$ that is also in $\mathcal{R}_{n+1}^0$, the set of $(n+1)$ -randomizations in Theorem 2.6(a).
 (c) Suppose that in addition to Assumption 5.1 we have
 (c.1) The p.m. ν is nonatomic; and
 (c.2) A is a Polish space and $A(x) = A$ for all $x \in X$.

Then the (deterministic) constrained problem (5.1) is solvable, and $\rho_c^* = \rho_c^\#$.

Proof. (a) By Assumption 5.1(d), problem (5.2) is feasible. Let $\{\varphi_k\} \subset \Phi$ be a minimizing sequence for (5.2); that is, each φ_k satisfies the constraints

$$(5.6) \quad J_i(\varphi_k) \leq b_i \quad \forall i = 1, \dots, n$$

and

$$(5.7) \quad J_0(\varphi_k) \downarrow \rho_c^\# \text{ as } k \rightarrow \infty.$$

By Assumption 5.1(c), at least one of the nonnegative functions $c_0, \dots, c_n$ is inf-compact, which, together with (5.6), (5.7) and Lemma 4.4, yields that the sequence $\{\mu_k := \nu \cdot \varphi_k, k \geq N\}$ is relatively compact for some N sufficiently large. Observe that the latter statement is analogous to (4.19); then the remainder of the proof follows exactly the same arguments used in (4.20)–(4.23).

(b) Let $\Delta \subset \Phi$ be as in Theorem 2.6. As the cost functions $c_1, \dots, c_n$ are nonnegative (Assumption 5.1(b)), Δ coincides with the set of randomized strategies $\varphi \in \Phi$ that satisfy (2.6), which is the same as the set of φ 's that satisfy the constraints $J_i(\varphi) \leq b_i$ in (5.2). Let $\Delta^* \subset \Delta$ be as in part (a) and consider the set $\nu \cdot \Delta^*$; see

Definition 4.7. Then (as in Theorem 4.12) it is easily seen that $\nu \cdot \Delta^*$ is a weakly compact face of $\nu \cdot \Delta$, and so (as in the proof of Theorem 4.12(e)) $\nu \cdot \Delta^*$ contains an extreme point $\nu \cdot \varphi$, say, of $\nu \cdot \Delta$. Hence, by Lemma 4.11(a), $\bar{\varphi}$ is an extreme point of $\bar{\Delta}$. Finally, by the hypothesis on A and Remark 2.5(c) (or Remark 3.4(b)), the desired conclusion follows from Lemma 4.11(c).

(c) Under Assumptions 5.1, (c.1) and (c.2), the hypotheses of Lemma 4.9 are satisfied. Consequently if $\varphi^* \in \Phi$ is as in (5.4) and (5.5), then there exists $f^* \in \mathbb{F}$ such that

$$\rho_c^\# \geq \int_X c_0(x, f^*(x)) \nu(dx) = J_0(f^*) \geq \rho_c^*$$

and, similarly,

$$b_i \geq J_i(\varphi^*) \geq J_i(f^*) \quad \forall i = 1, \dots, n.$$

These inequalities and (5.3) yield (c). $\square$

6. Constrained MCPs. In this section we consider constrained MCPs with *discounted* cost criteria, but it will be clear from the following discussion that similar results can be obtained for other cost criteria, e.g., average cost, total cost, etc. On the other hand, as constrained MCPs are quite standard, we shall introduce only the minimum necessary material required to state the corresponding control problem and its connection with the results in sections 2 and 5. (For further details, if necessary, the reader may consult, e.g., [3, 14, 16, 17, 27, 28, 32, 33].)

In a constrained MCP we are given, first, a *control model*

$$(6.1) \quad (X, A, \mathbb{K}, Q, c_0, \dots, c_n, b_1, \dots, b_n)$$

with X , A , $\mathbb{K}$, and c_i ($i = 0, \dots, n$) as in Assumption 5.1(a), (b), (c), and $b_i \geq 0$ ($i = 1, \dots, n$). Thus with respect to the previous sections the only new component in (6.1) is the so-called system's *transition law* Q , which is a stochastic kernel from $\mathbb{K}$ to X , i.e.,

- $Q(B|\cdot)$ is a measurable function on $\mathbb{K}$ for each fixed $B \in \mathcal{B}(X)$, and
- $Q(\cdot|x, a)$ is a p.m. on X for each fixed pair $(x, a) \in \mathbb{K}$.

Denoting by x_t and a_t ($t = 0, 1, \dots$) the state and control variables, the transition law Q is interpreted as the transition probability

$$Q(B|x, a) = \text{Prob}(x_{t+1} \in B \mid x_t = x, a_t = a) \quad \forall B \in \mathcal{B}(X), (x, a) \in \mathbb{K}.$$

Control policies. A control policy is a sequence $\pi = \{\pi_t\}$ of stochastic kernels π_t that satisfy the constraints

$$\pi_t(A(x)|x_0, a_0, \dots, x_{t-1}, a_{t-1}, x_t = x) = 1$$

for all $t = 0, 1, \dots$, with (x_i, a_i) in $\mathbb{K}$ for all $i = 0, \dots, t-1$. Let Π be the set of all control policies.

Let Φ and $\mathbb{F}$ be as in Definitions 2.1 and 2.2; recall (2.3). A policy $\pi = \{\pi_t\}$ is said to be *randomized stationary* if there exists $\varphi \in \Phi$ such that

$$(6.2) \quad \pi_t(\cdot|x_0, a_0, \dots, x_{t-1}, a_{t-1}, x_t = x) = \varphi(\cdot|x) \quad \forall t = 0, 1, \dots, x \in X,$$

and *deterministic stationary* if there exists $f \in \mathbb{F}$ for which (6.2) holds with $\varphi(\cdot|x) = \delta_{f(x)}(\cdot)$. We shall identify Φ and $\mathbb{F}$ with the families of randomized stationary policies and deterministic stationary policies, respectively.

Performance criteria. Let $0 < \alpha < 1$ be a given “discount factor,” and let γ be a given p.m. (the “initial distribution”) on X . For each $i = 0, 1, \dots, n$ and each policy $\pi \in \Pi$ let

$$V_i(\pi, \gamma) := (1 - \alpha) E_\gamma^\pi \left[\sum_{t=0}^{\infty} \alpha^t c_i(x_t, a_t) \right]$$

be the α -discounted cost when using the control policy π , given the initial distribution γ . We may then express the *constrained MCP* we are concerned with as

$$(6.3) \quad \text{CP : minimize } V_0(\pi, \gamma)$$

$$(6.4) \quad \text{s.t. } V_i(\pi, \gamma) \leq b_i \quad \forall i = 1, \dots, n, \pi \in \Pi.$$

To prove, among other things, that CP is *solvable* (i.e., there exists $\pi^* \in \Pi$ that satisfies (6.4) and it attains the minimum in (6.3)—in which case π^* is said to be an *optimal policy* for CP) we will use the fact that the CP is *equivalent* to a constrained optimization problem similar to (5.2); see (6.6), (6.7), below. To do this we impose the following conditions.

ASSUMPTION 6.1. *Parts (a), (b), and (c) are the same as in Assumption 5.1. Moreover,*

(d) *the transition law is weakly continuous, i.e., for each $u \in C_b(X)$, the function*

$$(6.5) \quad (x, a) \mapsto \int_X u(y) Q(dy|x, a) \text{ is in } C_b(\mathbb{K});$$

(e) *CP is feasible; i.e., there exists $\hat{\pi} \in \Pi$ such that $V_0(\hat{\pi}, \gamma) < \infty$, and $\pi = \hat{\pi}$ satisfies (6.4).*

Under Assumption 6.1, there are several ways to prove that CP is solvable; see, for instance, Corollary 5.1 in [13] or Theorem 3.2 in [16]. In particular, a key step in the proof in the latter reference, [16], is to show that CP is *equivalent* to the problem

$$\text{CP}_0 : \text{ minimize } \int_X c_0(x, \varphi(x)) \hat{\mu}(dx)$$

$$(6.6) \quad \text{s.t. } \int_X c_i(x, \varphi(x)) \hat{\mu}(dx) \leq b_i \quad \forall i = 1, \dots, n,$$

$$(6.7) \quad \hat{\mu}(B) = (1 - \alpha) \gamma(B) + \alpha \int_X Q(B|x, \varphi(x)) \hat{\mu}(dx) \quad \forall B \in \mathcal{B}(X),$$

$$\mu = \hat{\mu} \cdot \varphi \in \mathbb{P}_0(\mathbb{K}), \quad \varphi \in \Phi,$$

where $\mathbb{P}_0(\mathbb{K})$ is the subset of measures μ in $\mathbb{P}(\mathbb{K})$ for which $\int c_0 d\mu < \infty$ (see Lemma 3.3 in [16]). The equivalence between CP and CP_0 is in the sense that one of these problems is solvable if and only if so is the other, and both problems have the same value. Moreover, if ρ is the optimal value of CP_0 and $\mu_0 = \hat{\mu}_0 \cdot \varphi_0 \in \mathbb{P}_0(\mathbb{K})$ is an optimal solution, i.e.,

$$(6.8) \quad \rho = \int c_0 d\mu_0 = \inf \left\{ \int c_0 d\mu \mid \mu \in M \right\},$$

where M is the (convex) set of *feasible solutions* for CP_0 , then the randomized stationary policy (or randomized strategy) $\varphi_0 \in \Phi$ is an optimal policy for CP .

To extend this result to an analogue of Theorem 5.2(b) we shall require, among other things, the transition law Q to be *nonatomic*, which means that $Q(\cdot|x, a)$ is nonatomic for each pair (x, a) in $\mathbb{K}$.

THEOREM 6.2. *Under Assumption 6.1 we have the following.*

- (a) CP_0 (equivalently, CP) is solvable.
- (b) Suppose that, in addition, A is a locally compact separable metric space and γ and Q are both nonatomic. Furthermore, let $\mu_0 = \hat{\mu}_0 \cdot \varphi_0$ be an optimal solution for CP_0 (see (6.8)). Then there is an optimal stationary strategy $\varphi^* \in \Phi$ for CP that is in $\mathcal{R}_{n+1}^0$ (the set of $(n+1)$ -randomizations in Theorem 2.6).

Proof. As was already noted above, part (a) has been established in (e.g.) [13, 16]. To prove (b), let $\mu_0 = \hat{\mu}_0 \cdot \varphi_0$ be the given optimal solution for CP_0 and consider the convex set Δ_0 of randomized strategies $\varphi \in \Phi$ such that $\hat{\mu}_0 \cdot \varphi$ is in M , i.e., φ satisfies

$$(6.9) \quad \int_X c_0(x, \varphi(x)) \hat{\mu}_0(dx) < \infty,$$

$$(6.10) \quad \int_X c_i(x, \varphi(x)) \hat{\mu}_0(dx) \leq b_i \quad \forall i = 1, \dots, n,$$

$$(6.11) \quad \hat{\mu}_0(\cdot) = (1 - \alpha)\gamma(\cdot) + \alpha \int_X Q(\cdot|x, \varphi(x)) \hat{\mu}_0(dx).$$

Observe that $\rho_0 := \inf \{ \int c_0 d\mu | \mu = \hat{\mu}_0 \cdot \varphi \in \hat{\mu}_0 \cdot \Delta_0 \}$ coincides with ρ , i.e.,

$$(6.12) \quad \rho_0 = \rho.$$

Indeed, by (6.8) it is clear that $\rho \leq \rho_0$, and, on the other hand, the reverse inequality is obvious because $\mu_0 = \hat{\mu}_0 \cdot \varphi_0$ is in $\hat{\mu}_0 \cdot \Delta_0$. Moreover, as in Theorem 4.12, it can be seen that $\hat{\mu}_0 \cdot \Delta_0$ is a convex compact subset of M . Hence, there is an extreme point $\mu^* = \hat{\mu}_0 \cdot \varphi^*$ of $\hat{\mu}_0 \cdot \Delta_0$, which, by (6.12), is an optimal solution for CP_0 , i.e.,

$$(6.13) \quad \int c_0 d\mu^* = \rho_0 = \rho.$$

On the other hand, from Lemma 4.11(a) with $\hat{\mu}_0$ and Δ_0 in lieu of ν and Δ , respectively, i.e.,

$$\overline{\Delta}_0 := \{ \overline{\varphi} | \varphi \in \Delta_0 \} \text{ with } \overline{\varphi} := \{ \varphi' \in \Delta_0 | \hat{\mu}_0 \cdot \varphi' = \hat{\mu}_0 \cdot \varphi \},$$

it follows that $\overline{\varphi}^*$ is an extreme point of $\overline{\Delta}_0$. This in turn gives that, as in Lemma 4.11(c), $\overline{\varphi}^*$ has an extreme point, which we shall denote φ^* again, of Δ_0 . To complete the proof of the theorem we will next show that such a $\varphi^* \in ex(\Delta_0)$ is precisely as required in part (b).

To this end, we will use Theorem 2.6 with $\nu = \hat{\mu}_0$ as above and Δ being the set of randomized strategies $\varphi \in \Phi$ that satisfy (6.10). Note that $\hat{\mu}_0 \cdot \Delta$ contains $\hat{\mu}_0 \cdot \Delta_0$, and so φ^* is in Δ . Therefore we may use (2.7) to conclude (b) if we can show that

$$(6.14) \quad \varphi^* \in ex(\Delta).$$

To prove (6.14), let us suppose that φ^* is *not* an extreme point of Δ . Then, as in the proof of (3.3), there exist $z \in X$ and $C \in \mathcal{B}(A)$ such that $0 < \varphi^*(C|z) < 1$. Moreover, let $\beta := \varphi^*(C|z)$, and for $i = 1, 2$ define $\varphi_i \in \Delta$ as in (3.4), (3.5) with φ^* in lieu of φ , so that

$$(6.15) \quad \varphi^* = \beta\varphi_1 + (1 - \beta)\varphi_2.$$

Now, to prove (6.14), we will next show that φ_1, φ_2 are in fact in Δ_0 , which, together with (6.15), contradicts the fact that φ^* is in $ex(\Delta_0)$. To prove this, note first that φ_1 and φ_2 satisfy (6.9) because so does φ^* . On the other hand, by (6.11) and the hypothesis that γ and Q are *nonatomic*, so is $\hat{\mu}_0$; in particular $\hat{\mu}_0(\{z\}) = 0$. Hence, by the definition of φ_1 and φ_2 , they both satisfy (6.11); hence they belong to Δ_0 . $\square$

REMARK 6.3. (a) *A key step in the previous proof was to reduce CP_0 to a problem similar to (5.2) with $\nu = \hat{\mu}_0$; see (6.10). However, because of the additional constraint (6.11), to this date (as far as we can tell) there are no general conditions ensuring the existence of optimal deterministic strategies for CP_0 , that is, for the constrained MCP (6.3), (6.4). In other words, for constrained control problems there is no analogue of, say, Theorem 5.2(c).*

(b) *A MCP is said to be unconstrained if in (6.4) we have $c_i(x, a) \equiv b_i$ for all $i = 1, \dots, n$ and (x, a) in $\mathbb{K}$. Theorem 6.2 can be simplified in the obvious manner to the unconstrained case. The results will be, of course, similar to those in section 4, but incorporating the constraint (6.7); see, e.g., [18, Chapter 6] for details.*

7. Concluding remarks and open problems. In the previous sections we have presented a fairly complete, reasonably self-contained analysis of some classes of constrained (and unconstrained) optimization and control problems that use randomized strategies—or, in other words, problems in which the decisions or control actions may or need to be “randomized”. For these problems, our results ensure the *existence* of optimal solutions, as well as the *characterization* of these solutions as extreme points of certain convex sets of randomized strategies. A key feature of our paper is that such a characterization is obtained via the *general* results in Theorem 2.6 (and Lemma 3.3).

But, on the other hand, there is an important question conspicuously missing in our results, namely, how to *compute* optimal solutions. Concerning this question there are several possibilities that might be worth exploring. First, one can try to apply some approximation procedure for general optimization problems—see, e.g., [4] and the references therein.

Second, one may note that our problems (see (4.4), (5.2) or (6.6), (6.7)) are in fact infinite-dimensional *linear programs*; therefore, one could try to adapt some of the existing approximation schemes, as in [19], for instance.

Third, it is well known that (at least some) constrained optimization problems can be seen as *multiobjective* problems in which one is looking for some particular *Pareto* (or efficient or nondominated) solutions—see, e.g., [14, 22, 27]. For example, consider the constrained problem (5.2) and let $J(\varphi)$ be the cost vector

$$J(\varphi) := (J_0(\varphi), J_1(\varphi), \dots, J_n(\varphi)) \in \mathbb{R}^{n+1},$$

and $\Gamma := \{J(\varphi) \mid \varphi \in \Phi\} \subset \mathbb{R}^{n+1}$ the corresponding achievable (or performance) set. As Γ is a *convex* set, one can easily characterize the set P^* of Pareto solutions of the multiobjective problem (MP), i.e., the set of randomized strategies $\varphi \in \Phi$ for which the cost vector $J(\varphi)$ is in the Pareto set of Γ . Moreover, under mild assumptions

(see the above references), *there is a Pareto solution $\varphi^* \in P^*$ to MP that is also a solution to the constrained problem (5.2)*. The key fact is that for such a φ^* the constraints (2.6) hold with *equality*. Therefore, by part (b) of Theorem 2.6, to solve (5.2) we would essentially be looking for Pareto solutions in the set of extreme points $ex(\Delta) = \mathcal{R}_{n+1}^0$. In other words, to find a solution for (5.2) it suffices to look for randomized strategies at the intersection $P^* \cap \mathcal{R}_{n+1}^0$. This should certainly simplify the search of optimal randomized strategies.

To conclude, note that the “Pareto approach” briefly described above is referred to as the “geometric approach” by some authors (e.g., [11]).

REFERENCES

- [1] W. ADAMSKI, *Complete spaces and zero-one measures*, Manuscripta Math., 18 (1976), pp. 343–352.
- [2] C. D. ALIPRANTIS AND K. C. BORDER, *Infinite Dimensional Analysis*, 2nd ed., Springer-Verlag, Berlin, 1999.
- [3] E. ALTMAN, *Constrained Markov Decision Processes*, Chapman & Hall/CRC, Boca Raton, FL, 1999.
- [4] J. ALVAREZ-MENA AND O. HERNÁNDEZ-LERMA, *Convergence and approximation of optimization problems*, SIAM J. Optim., 15 (2005), pp. 527–539.
- [5] J.-P. AUBIN, *Optima and Equilibria*, Springer-Verlag, Berlin, 1993.
- [6] R. J. AUMANN AND M. PERLES, *A variational problem arising in economics*, J. Math. Anal. Appl., 11 (1965), pp. 488–503.
- [7] E. J. BALDER, *Elimination of randomization in statistical decision theory revisited*, J. Multivariate Anal., 16 (1985), pp. 260–264.
- [8] E. J. BALDER, *New existence results for optimal controls in the absence of convexity: The importance of extremality*, SIAM J. Control Optim., 32 (1994), pp. 890–916.
- [9] E. J. BALDER, *Lectures on Young Measures*, Cahier de Mathématiques de la Décision, CEREMADE, Université Paris IX-Dauphine, Paris, 1995.
- [10] A. DVORETZKY, A. WALD, AND J. WOLFOWITZ, *Elimination of randomization in certain statistical decision procedures and zero-sum two-person games*, Ann. Math. Statist., 22 (1951), pp. 1–21.
- [11] YU. ERMOLIEV, A. GAIVORONSKI, AND C. NEDEVA, *Stochastic optimization problems with incomplete information on distribution functions*, SIAM J. Control Optim., 23 (1985), pp. 697–716.
- [12] E. A. FEINBERG AND A. B. PIUNOVSKIY, *Multiple objective nonatomic Markov decision processes with total reward criteria*, J. Math. Anal. Appl., 247 (2000), pp. 45–66.
- [13] E. A. FEINBERG AND A. B. PIUNOVSKIY, *Nonatomic total rewards Markov decision processes with multiple criteria*, J. Math. Anal. Appl., 273 (2002), pp. 93–111.
- [14] E. A. FEINBERG AND A. SHWARTZ, *Constrained discounted dynamic programming*, Math. Oper. Res., 21 (1996), pp. 922–945.
- [15] J. GONZÁLEZ-HERNÁNDEZ AND R. PÉREZ-HERNÁNDEZ, *An example of a bounded-in-probability, but non-tight Markov chain*, Morfismos, 2 (1998), pp. 75–79.
- [16] O. HERNÁNDEZ-LERMA AND J. GONZÁLEZ-HERNÁNDEZ, *Constrained Markov control processes in Borel spaces: The discounted case*, Math. Methods Oper. Res., 52 (2000), pp. 271–285.
- [17] O. HERNÁNDEZ-LERMA, J. GONZÁLEZ-HERNÁNDEZ, AND R. R. LÓPEZ-MARTÍNEZ, *Constrained average cost Markov control processes in Borel spaces*, SIAM J. Control Optim., 42 (2003), pp. 442–468.
- [18] O. HERNÁNDEZ-LERMA AND J. B. LASSERRE, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer-Verlag, New York, 1996.
- [19] O. HERNÁNDEZ-LERMA AND J. B. LASSERRE, *Approximation schemes for infinite linear programs*, SIAM J. Optim., 8 (1998), pp. 973–988.
- [20] O. HERNÁNDEZ-LERMA AND J. B. LASSERRE, *Further Topics on Discrete-Time Markov Control Processes*, Springer-Verlag, New York, 1999.
- [21] O. HERNÁNDEZ-LERMA AND J. B. LASSERRE, *Zero-sum stochastic games in Borel spaces: Average payoff criteria*, SIAM J. Control Optim., 39 (2001), pp. 1520–1539.
- [22] O. HERNÁNDEZ-LERMA AND R. ROMERA, *The scalarization approach to multiobjective Markov control problems: Why does it work?*, Appl. Math. Optim., 50 (2004), pp. 279–293.
- [23] K. HINDERER, *Foundations of Non-Stationary Dynamic Programming with Discrete-Time Pa-*

- rameter, Lecture Notes Oper. Res. Math. Syst. 33, Springer-Verlag, Berlin, 1970.
- [24] A. F. KARR, *Extreme points of certain sets of probability measures with applications*, Math. Oper. Res., 8 (1983), pp. 74–85.
 - [25] D. G. KENDALL, *Simplexes and vector lattices*, J. London Math. Soc., 37 (1962), pp. 365–371.
 - [26] M. G. KREIN AND A. A. NUDEL'MAN, *The Markov Moment Problem and Extremal Problems*, Transl. Math. Monogr. 50, AMS, Providence, RI, 1977.
 - [27] R. R. LÓPEZ-MARTÍNEZ AND O. HERNÁNDEZ-LERMA, *The Lagrange approach to constrained Markov control processes: A survey and extension of results*, Morfismos, 7 (2003), pp. 1–26.
 - [28] X. MAO AND A. PIUNOVSKIY, *Strategic measures in optimal control problems for stochastic sequences*, Stochastic Anal. Appl., 18 (2000), pp. 755–776.
 - [29] J. NEVEU, *Mathematical Foundations of the Calculus of Probability*, Holden-Day, San Francisco, CA, 1965.
 - [30] P. PEDREGAL, *Optimization, relaxation and Young measures*, Bull. Amer. Math. Soc. (N.S.), 36 (1999), pp. 27–58.
 - [31] R. R. PHELPS, *Lectures on Choquet's Theorem*, Van Nostrand, New York, 1966.
 - [32] A. B. PIUNOVSKIY, *Control of random sequences in problems with constraints*, Theory Probab. Appl., 38 (1994), pp. 751–762.
 - [33] A. B. PIUNOVSKIY, *Optimal Control of Random Sequences in Problems with Constraints*, Kluwer, Dordrecht, 1997.
 - [34] L. SCHWARTZ, *Radon Measures on Arbitrary Topological Spaces and Cylindrical Measures*, Oxford University Press, London, 1973.
 - [35] R. SERFOZO, *Convergence of Lebesgue integrals with varying measures*, Sankhyā Ser. A, 44 (1982), pp. 380–402.
 - [36] J. E. SMITH, *Generalized Chebyshev inequalities: Theory and applications in decision analysis*, Oper. Res., 43 (1995), pp. 807–825.
 - [37] F. TOPSOE, *Topology and Measure*, Lecture Notes in Math. 133, Springer-Verlag, Berlin, 1975.
 - [38] N. N. VAKHANIA, V. I. TARIELADZE, AND S. A. CHOBANYAN, *Probability Distributions in Banach Spaces*, D. Reidel, Dordrecht, 1987.
 - [39] A. WALD, *Statistical Decision Functions*, Wiley, New York, 1950.
 - [40] G. WINKLER, *Extreme points of moment sets*, Math. Oper. Res., 13 (1988), pp. 581–587.
 - [41] M. E. YAARI, *On the existence of an optimal plan in continuous-time allocation processes*, Econometrica, 32 (1964), pp. 576–590.
 - [42] L. C. YOUNG, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, PA, 1969.
 - [43] A. A. YUSHKEVICH, *The compactness of a policy space in dynamic programming via an extension theorem for Carathéodory functions*, Math. Oper. Res., 22 (1997), pp. 458–467.

A PREDICTOR-CORRECTOR ALGORITHM FOR LINEAR OPTIMIZATION BASED ON A SPECIFIC SELF-REGULAR PROXIMITY FUNCTION*

JIMING PENG[†], TAMÁS TERLAKY[†], AND YUNBIN ZHAO[‡]

Abstract. It is well known that the so-called first-order predictor-corrector methods working in a large neighborhood of the central path are among the most efficient interior-point methods (IPMs) for linear optimization (LO) problems. However, the best known iteration complexity of this type of method is $O(n \log \frac{(x^0)^T s^0}{\varepsilon})$. It is of interest to investigate whether the complexity of first-order predictor-corrector type methods can be further improved. In this paper, based on a specific self-regular proximity function, we define a new large neighborhood of the central path. In particular, we show that the new neighborhood matches the standard large neighborhood that is defined by the infinity norm and widely used in the IPM literature. A new first-order predictor-corrector method for LO that uses a search direction induced by self-regularity in corrector steps is proposed. We prove that our predictor-corrector algorithm, working in a large neighborhood, has an $O(\sqrt{n} \log n \log \frac{(x^0)^T s^0}{\varepsilon})$ iteration bound. Local superlinear convergence of the algorithm is also established.

Key words. linear optimization, interior-point methods, predictor-corrector method, self-regular proximity function, large neighborhoods, polynomial complexity, superlinear convergence

AMS subject classification. 90C05

DOI. 10.1137/040603991

1. Introduction. To begin, we first clarify the notation used in this paper. Let $\mathbb{R}^n$ denote the n -dimensional Euclidean space and e denote the vector with all components equal to 1. Further, $x \geq 0$ ($x > 0$) means that all components of x are nonnegative (positive). For any $x > 0$, $s > 0$, and real number r , the symbols xs , $\sqrt{x}$, and x^r denote vectors whose components are $x_i s_i$, $\sqrt{x_i}$, and x_i^r ($i = 1, \dots, n$), respectively. In particular, we have $x^{-1} = (\frac{1}{x_1}, \dots, \frac{1}{x_n})^T$.

Since Karmarkar's seminal paper [8], a large body of research has been done in the field of IPMs for LO problems. Several recent books on this subject [19, 26, 27] contain substantial materials about various aspects of interior-point methods (IPMs) for linear optimization (LO). In the present paper, we consider the standard LO problem that takes the following form:

$$\min\{c^T x : Ax = b, x \geq 0\}.$$

*Received by the editors February 9, 2004; accepted for publication (in revised form) November 22, 2004; published electronically July 26, 2005. This research was supported by the MITACS project "New Interior Point Methods and Software for Convex Conic-Linear Optimization and Their Application to Solve VLSI Circuit Layout Problems" and an FPP grant from IBM Watson Research Laboratory.

<http://www.siam.org/journals/siopt/15-4/60399.html>

[†]Department of Computing and Software, McMaster University, Hamilton, Ontario, Canada, L8S 4L7 (pengj@mcmaster.ca, terlaky@mcmaster.ca). The research of the first author was partially supported by grant RPG 249635-02 of the National Sciences and Engineering Research Council of Canada (NSERC) and a PREA award. The research of the second author was supported by NSERC grant RPG 0048923 and the Canada Research Chair program.

[‡]Institute of Applied Mathematics, AMSS, Chinese Academy of Science, Beijing, China (ybzha@amss.ac.cn). Part of the work was done when this author visited The Fields Institute for Research in Mathematical Sciences as a post doctor fellow. The research of this author was partially supported by grant 10201032 and grant 70221001 of the Natural Science Foundation of China.

Here $x \in \mathbb{R}^n$, $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, and A is an $m \times n$ matrix satisfying $\text{Rank}(A) = m$. Throughout the paper, we also assume [19] that there exists a strictly feasible primal-dual pair (x^0, y^0, s^0) such that

$$Ax^0 = b, \quad A^T y^0 + s^0 = c, \quad (x^0, s^0) > 0.$$

In the present paper, we focus on so-called first-order predictor-corrector IPMs that work in a large neighborhood of the central path.

We start with a brief review of predictor-corrector type methods. The idea of predictor-corrector type algorithms is very natural and elegant. These algorithms follow the central path by alternatively taking predictor steps and corrector steps. A predictor step aims at reducing the duality gap as much as possible, thus the resulting iterate might move close to the boundary of the feasible set, deviating from the central path, while a corrector step tries to bring the iterate back to a certain neighborhood of the central path. The best known representative of this class of methods is the Mizuno–Todd–Ye (MTY) algorithm for LO [11], which works in a small neighborhood of the central path. Quadratic convergence of the duality gap was proved in [29], and convergence of the iterates generated by the algorithm was proved in [4]. The MTY method was later extended to complementarity problems by Ye and Anstreicher [28] and to semidefinite optimization by Luo, Sturm, and Zhang [10]. According to numerical experiments, IPMs working in a large neighborhood of the central path perform much better than their counterparts in small neighborhoods [1]. Several authors have investigated IPMs that work in certain large neighborhoods of the central path [2, 3, 5, 9, 15, 16, 17, 24, 25, 31]. It should be noted that the so-far best known worst-case complexity for the first-order predictor-corrector algorithms with large neighborhoods is $O(n \log \frac{(x^0)^T s^0}{\epsilon})$, which is higher than that of those small neighborhood based algorithms. As remarked by Renegar [18], this becomes one of the ironies of the interior-point literature that algorithms which are more efficient in practice often have somewhat-worse complexity bounds.

Several strategies have already been proposed for improving the theoretical complexity of IPMs in large neighborhoods. For example, by using higher-order approximation of the central path, the iteration complexity of IPMs can become arbitrarily close to $O(\sqrt{n} \log \frac{(x^0)^T s^0}{n})$ [5, 7, 12, 17, 19, 20, 21, 30, 31]. In IPMs based on high-order approximation, we usually need to solve multiple linear systems and thus the computational cost per iteration increases as the order of the method increases. Recently, Peng, Roos, and Terlaky [13] proposed a class of IPMs in large neighborhoods and proved that some members of that family of IPMs have an $O(\sqrt{n} \log n \log \frac{(x^0)^T s^0}{\epsilon})$ iteration bound. The approach in [13] is based on the notion of self-regular functions. Different from high-order approximation-type methods, IPMs based on self-regular functions solve only one linear system at each step. Inspired by the results in [13], we consider in the present paper the issue whether the iteration bound of first-order predictor-corrector type IPMs within a large neighborhood can be improved. This is far from being an easy task. As pointed out in [17], MTY-type predictor-corrector methods are much more difficult to develop and analyze in a wide neighborhood of the central path. The difficulty for developing MTY-type methods in large neighborhoods comes from two points. First, theoretically, corrector steps based on the classical first-order centering direction (first-order correctors) are rather inefficient in a large neighborhood.

A remedy for this point is to employ a corrector search direction based on self-

regular proximity functions introduced in [13]. However, this gives rise to another issue, namely whether the neighborhood defined by a self-regular proximity function matches the large neighborhoods used in the context of predictor-corrector methods. Second, once we use a new search direction in the corrector step, it is not clear if we can still obtain superlinear or quadratic convergence for the new algorithm.

The purpose of this paper is to address the above-mentioned questions and present a new first-order predictor-corrector method that works in a large neighborhood and has an $O(\sqrt{n} \log n \log \frac{(x^0)^T s^0}{\epsilon})$ iteration bound. To this end, a proximity measure function, induced by a specific self-regular function introduced in [13], will be explored. We then use this proximity function to define a new neighborhood of the central path and show that this new neighborhood matches the neighborhood $\mathcal{N}_\infty^-$ that has been widely used in most practical implementations of IPMs. Correspondingly, we also use the self-regularity induced search direction in the corrector steps. It is worth mentioning that we use the usual primal-dual affine scaling direction in the predictor step.

We note that in [23], the authors proposed a family of so-called generic central region IPMs that work in a large neighborhood of the central path and proved that the proposed algorithms enjoy an $O(\sqrt{n} \log \frac{(x^0)^T s^0}{\epsilon})$ complexity. This approach was later extended to semidefinite optimization in [22], where a predictor-corrector version of the algorithm was also discussed. It should be noted that the predictor-corrector method of [22, 23] uses the second-order correction in the corrector step and thus it does not follow the same procedure as our algorithm.

The algorithm presented in this paper follows the prototype of MTY IPMs in a large neighborhood. In the predictor step, we take a step along the affine scaling search direction to reduce the duality gap, while in the corrector step we try to bring the iterate back to a certain neighborhood (might be a small neighborhood) of the central path.

This paper is organized as follows. In section 2, we introduce the new neighborhood of the central path induced by a self-regular proximity function and discuss some useful properties of the proximity function. In section 3, we describe our new algorithm and establish its $O(\sqrt{n} \log n \log \frac{(x^0)^T s^0}{\epsilon})$ complexity. Asymptotic superlinear convergence of the algorithm is proved in section 4. Finally, we close the paper by a few concluding remarks.

2. Proximity functions and neighborhoods. This section consists of two parts. In the first subsection, we discuss a specific self-regular proximity function and show that the neighborhood induced by this proximity function matches the large neighborhood defined by infinity norm and used in many practical implementations of IPMs. In the second subsection, we explore some properties of the proximity function.

2.1. Large neighborhoods and proximity functions. We note that under the interior-point assumption, for every given parameter $\mu > 0$ the system

$$(2.1) \quad \begin{aligned} Ax &= b, & x &> 0, \\ A^T y + s &= c, & s &> 0, \\ xs &= \mu e, \end{aligned}$$

has a unique solution denoted by $(x(\mu), y(\mu), s(\mu))$. The union of solutions $\{x(\mu) : \mu > 0\}$ forms the so-called primal central path, while $\{(y(\mu), s(\mu)) : \mu > 0\}$ is called

the central path of the dual problem in the dual space. The iterates, generated by IPMs, are confined to be in certain neighborhoods of the central path, which are defined by certain proximity functions.

To describe these proximity functions, we need to introduce some notation. For any strictly feasible primal-dual pair (x, s) and any positive number μ ,¹ we define

$$(2.2) \quad v := \sqrt{\frac{xs}{\mu}} \quad \text{and} \quad v^{-1} := \sqrt{\frac{\mu e}{xs}}$$

to be the vectors whose i th components are $\sqrt{\frac{x_i s_i}{\mu}}$ and $\sqrt{\frac{\mu}{x_i s_i}}$, respectively. Let us denote the parameter μ associated with the present duality gap by

$$\mu(x, s) = \frac{x^T s}{n}.$$

A neighborhood of the central path is usually defined by

$$(2.3) \quad \mathcal{N} := \{(x, s) > 0 : Ax = b, A^T y + s = c, \Phi(x, s, \mu(x, s)) \leq \eta(\tau, n)\},$$

where $\eta(\tau, n)$ is a function of a parameter τ and the dimension n of the underlying problem, and $\Phi(x, s, \mu(x, s))$ is some measurement to measure the distance from the current iterate to the targeted center.

There are many choices for the measurement function $\Phi(x, s, \mu)$ and the function $\eta(\tau, n)$ in (2.3). For instance, if we choose $\Phi(x, s, \mu) = \|\frac{xs}{\mu(x, s)} - e\|$ and $\eta(\tau, n) = \rho$ with $\rho \in (0, 1)$ (e.g., $\eta(\tau, n) = 0.25$), then we obtain a so-called small neighborhood. If we choose $\Phi(x, s, \mu) = \|\frac{xs}{\mu(x, s)} - e\|_\infty$ (or $\Phi(x, s, \mu) = \|(\frac{xs}{\mu(x, s)} - e)^-\|_\infty$ with $w^- = \min(w, 0)$) and $\eta(\tau, n) = \rho$ with $\rho \in (0, 1)$, then we get a so-called large neighborhood. Let us denote the neighborhoods generated via using the proximity functions $\|\frac{xs}{\mu(x, s)} - e\|_\infty$ and $\|(\frac{xs}{\mu(x, s)} - e)^-\|_\infty$ by $\mathcal{N}_\infty(\rho)$ and $\mathcal{N}_\infty^-(\rho)$, respectively. It is worth pointing out that the neighborhood $\mathcal{N}_\infty^-(\rho)$ is frequently used in practical implementations of IPMs, although the corresponding proximity function does not have a barrier property. On the other hand, as observed by many researchers, various IPMs are tightly associated with some proximity functions that have some barrier and coercive properties to prevent the iterate from moving to infinity and the boundary of the feasible set. For instance, Peng, Roos, and Terlaky [13] introduced the so-called self-regular proximity functions, which takes the form

$$(2.4) \quad \Phi(x, s, \mu) := \Psi(v) = \sum_{i=1}^n \psi(v_i),$$

where $\psi(t)$ is a so-called self-regular kernel function. For self-regular proximity functions, if we still choose $\eta(\tau, n)$ as a small constant, then the region defined by (2.3) becomes a small neighborhood. If $\eta(\tau, n) = \tau n = \mathcal{O}(n)$, then we obtain a large neighborhood.

However, as observed in [14], it is possible that the neighborhood $\mathcal{N}$ defined by (2.3) with some self-regular proximity function is too large and significantly differ-

¹We alert the reader that throughout this paper μ is cast as an independent parameter, while the duality gap is associated with $\mu(x, s)$.

ent from the neighborhood $\mathcal{N}_\infty^-(\rho)$. To select an appropriate kernel function such that the resulting neighborhood defined by (2.3) can really match $\mathcal{N}_\infty^-(\rho)$, we consider the self-regular kernel function

$$(2.5) \quad \psi(t) = \frac{t^2 - 1}{2} + \frac{t^{-\log n} - 1}{\log n},$$

which is a self-regular function with $\nu_1 = 1$, $\nu_2 = \log n + 1$, $p = 1$, and $q = 1 + \log n$. The corresponding proximity function becomes

$$(2.6) \quad \Phi(x, s, \mu) = \Psi(v) = \frac{1}{2} \|v\|^2 - \frac{n}{2} + \sum_{i=1}^n \frac{v_i^{-\log n} - 1}{\log n}.$$

One particular reason why we consider this specific proximity function is that, as shown in [13], among various large-update IPMs based on self-regular proximity functions, the algorithm with $q = O(\log n)$ has the best worst-case complexity result. Since $q = 1 + \log n$ and the case for small q has been well studied in [13], in this paper we focus only on the case where q is sufficiently large, e.g., $q \geq 3$, which implies $n \geq 8$. Therefore, throughout this paper we assume $n \geq 8$.

Because the proximity function $\Psi(v)$ has an extremely strong barrier property, we need to choose correspondingly an appropriate function $\eta(\tau, n)$ in (2.3) so that the neighborhood given by (2.3) can match the neighborhood $\mathcal{N}_\infty^-(\rho)$ with $\rho \in (0, 1)$. Let us define

$$(2.7) \quad \eta(\tau, n) := \frac{n^\tau - n}{\log n}.$$

Note that if $(x, s) \in \mathcal{N}_\infty^-(\rho)$ and $\mu = \mu(x, s)$, then we have $v_i^2 \geq 1 - \rho$, which implies $-\log v_i \leq -\frac{1}{2} \log(1 - \rho)$ for every $i \in \{1, \dots, n\}$. From (2.6) we immediately obtain

$$\Phi(x, s, \mu(x, s)) = \Psi(v) = \sum_{i=1}^n \frac{v_i^{-\log n} - 1}{\log n} = \sum_{i=1}^n \frac{n^{-\log v_i} - 1}{\log n} \leq \frac{n^{1-\frac{1}{2}\log(1-\rho)} - n}{\log n}.$$

Our above discussion indicates that if we choose $\tau = 1 - \frac{1}{2} \log(1 - \rho)$ in (2.7), then the neighborhood $\mathcal{N}$ defined by (2.3) contains the neighborhood $\mathcal{N}_\infty^-(\rho)$. It is worth mentioning that $\eta(\tau, n)$ is an increasing function of τ . In the rest of this paper, $\eta(\tau, n)$ is defined by (2.7) for some $\tau > 1$ and thus our algorithm will operate in a large neighborhood.

2.2. Properties of the proximity function. In this section we investigate various properties of the proximity function $\Phi(x, s, \mu)$. We first consider the case that $\mu = \mu(x, s)$. Recall that the value of $\Phi(x, s, \mu(x, s))$ is essentially determined by the sum $\sum_{i=1}^n v_i^{-\log n}$. To estimate this sum, we first observe that

$$u \geq 2 \implies \frac{|t|^u - 1}{u} \geq \frac{t^2 - 1}{2} \quad \forall t \in \mathbb{R}.$$

Since $\log n > 2$ if $n \geq 8$, we have

$$t^{-\log n} = (t^{-1})^{\log n} \geq \frac{\log n}{2} (t^{-2} - 1) + 1 \quad \forall t > 0, n \geq 8.$$

This relation and the fact that $\|v\|^2 = n$ when $\mu = \mu(x, s)$ imply²

$$(2.8) \quad \sum_{i=1}^n v_i^{-\log n} \geq \frac{\log n}{2} (\|v^{-1}\|^2 - n) + n = \frac{\log n}{2} \|v - v^{-1}\|^2 + n \geq n \quad \forall n \geq 8,$$

and equality holds if and only if $v = e$.

We proceed to explore more properties of the proximity function $\Phi(x, s, \mu)$. For fixed $(x, s) > 0$, there exists $\tau \geq 1$ such that

$$\Phi(x, s, \mu(x, s)) = \sum_{i=1}^n \frac{v_i^{-\log n} - 1}{\log n} = \eta(\tau, n),$$

which equals

$$\tau = \frac{\log(\sum_{i=1}^n v_i^{-\log n})}{\log n}.$$

The above relation indicates that τ provides another way to measure the proximity function. Let us define

$$f(\theta) := \Phi(x, s, \theta \mu(x, s)).$$

It is easy to see that $f(\theta)$ is convex for $\theta > 0$. Denote

$$\theta^* = \arg \min_{\theta > 0} f(\theta).$$

It follows from the optimality condition of $f(\theta)$ that

$$(2.9) \quad \frac{n}{2(\theta^*)^2} - \frac{(\theta^*)^{\frac{1}{2} \log n}}{2\theta^*} \sum_{i=1}^n v_i^{-\log n} = 0.$$

From this equality and (2.7) we obtain

$$n^{1-\tau} = (\theta^*)^{1+\frac{1}{2} \log n},$$

which yields

$$(1 - \tau) \log n = \left(1 + \frac{1}{2} \log n\right) \log \theta^*.$$

Thus we have

$$(2.10) \quad \theta^* = \exp^{-\frac{(2\tau-2) \log n}{2+\log n}}.$$

For a given primal-dual pair (x, s) , we next consider the behavior of the proximity function $\Phi(x, s, \mu)$ as a function of μ . First, we mention that it is trivial to

²One can easily derive that $\Phi(x, s, \mu(x, s)) \geq \frac{1}{2} \|v - v^{-1}\|^2$, and the right-hand side is a well-known proximity measure in the literature, which is also a self-regular proximity with $p = 1$ and $q = 3$; see, e.g., [13].

verify that Φ is convex in μ . From the choice (2.10) of θ^* , we immediately have the following lemma.

LEMMA 2.1. *Given a primal-dual pair (x, s) , suppose that $\Phi(x, s, \mu(x, s)) = \eta(\tau, n)$ for some $\tau > 1$ where $\eta(\tau, n)$ is given by (2.7), and θ^* is given by (2.10). Then the proximity function $\Phi(x, s, \mu)$ is convex with respect to μ and attains its global minimum at*

$$\mu^* = \theta^* \mu(x, s).$$

In particular, we have

$$\sum_{i=1}^n \left(\frac{x_i s_i}{\mu} \right)^{-\frac{1}{2} \log n} - \frac{1}{\mu} \sum_{i=1}^n x_i s_i \leq 0, \iff 0 < \mu \leq \mu^*,$$

and the equality holds if and only if $\mu = \mu^*$.

By using (2.10) one can easily show that

$$(2.11) \quad \exp^{2-2\tau} \leq \theta^*.$$

Further, we have

$$(2.12) \quad \theta^* \leq \exp^{1-\tau} \quad \forall n \geq 8.$$

The above two bounds indicate that if $n \geq 8$, θ^* is uniformly bounded above and below and these bounds depend only on τ . Let us cast θ^* as a function of τ and n , i.e.,

$$(2.13) \quad \theta^* := \theta(\tau, n) = \exp^{-\frac{(2\tau-2) \log n}{2+\log n}}.$$

We point out that we use $\mu = \theta^* \mu(x, s)$ in the update of the parameter μ in our algorithm. Lemma 2.1 indicates that if $\Phi(x, s, \mu(x, s)) = \eta(\tau, n)$ with $n \geq 8$ and we use $\mu = \theta(\tau, n) \mu(x, s)$ (which is indeed a large update) as our targeted μ in the algorithm, then the proximity function $\Phi(x, s, \mu)$ will not increase after the update. It is straightforward to verify that $\theta(\tau, n)$ is a decreasing function of τ .

For fixed $(x, s) > 0$ and $\theta(\tau, n)$ given by (2.13), let us define

$$(2.14) \quad \eta_1(\tau, n) := \left(\frac{n}{2} + \frac{n}{\log n} \right) \left(\frac{1}{\theta(\tau, n)} - 1 \right).$$

If $\Phi(x, s, \mu(x, s)) = \eta(\tau, n)$, then from (2.9) we can conclude

$$\Phi(x, s, \theta(\tau, n) \mu(x, s)) = \eta_1(\tau, n).$$

It is trivial to verify that $\eta_1(\tau, n)$ is strictly increasing with respect to τ . This gives the following corollary.

COROLLARY 2.2. *Let $(x, s) > 0$, $n \geq 8$, and $\tau \geq 2$, and suppose that the proximity function $\Phi(x, s, \mu)$ attains its minimum at μ^* . Then $\Phi(x, s, \mu(x, s)) \leq \eta(\tau, n)$ if and only if $\Phi(x, s, \mu^*) \leq \eta_1(\tau, n)$.*

We also point out here that when $n \geq 8$, $\frac{\eta_1(n, \tau)}{n}$ is uniformly bounded below and above by some constants independent of n .

It should be mentioned that in the above discussion, we sometimes implicitly assume that the parameter τ is dependent on the present iterate (x, s) . However, in most IPMs, we usually fix the parameter in the definition of the neighborhood (such as ρ in $\mathcal{N}_\infty^-(\rho)$) and only require the iterate to be in a certain neighborhood. Therefore, in the remaining part of this paper, we only assume that $\tau \geq 2$ is fixed. If the present iterate is in a certain neighborhood (say $\Phi(x, s, \mu(x, s)) \leq \eta(\tau, n)$), then we update the parameter μ and compute a search direction w.r.t. the targeted μ . In this case, we need to investigate the change of the proximity function. Our following result provides an upper bound for $\Phi(x, s, \theta(\tau, n)\mu(x, s))$ whenever $\Phi(x, s, \mu(x, s)) < \eta(\tau, n)$.

LEMMA 2.3. *If $\Phi(x, s, \mu(x, s)) \leq \eta(\tau, n)$, then one has*

$$\Phi(x, s, \theta(\tau, n)\mu(x, s)) \leq \eta_1(\tau, n).$$

Proof. From the assumption in the lemma, without loss of generality, we can assume

$$\Phi(x, s, \mu(x, s)) = \eta(\tau_0, n) \leq \eta(\tau, n)$$

for some $\tau_0 \leq \tau$. Since $\theta(\tau, n)$ is decreasing in τ , we therefore have $\theta(\tau_0, n) \geq \theta(\tau, n)$. Now recall the definition of $\theta(\tau_0, n)$. From the optimality condition (2.9) we have

$$\frac{x^T s}{\theta(\tau_0, n)\mu(x, s)} = \sum_{i=1}^n \left(\frac{x_i s_i}{\theta(\tau_0, n)\mu(x, s)} \right)^{-\frac{1}{2} \log n}.$$

It follows that

$$\begin{aligned} \sum_{i=1}^n \left(\frac{x_i s_i}{\theta(\tau, n)\mu(x, s)} \right)^{-\frac{1}{2} \log n} &= \left(\frac{\theta(\tau, n)}{\theta(\tau_0, n)} \right)^{\frac{1}{2} \log n} \sum_{i=1}^n \left(\frac{x_i s_i}{\theta(\tau_0, n)\mu(x, s)} \right)^{-\frac{1}{2} \log n} \\ (2.15) \quad &\leq \sum_{i=1}^n \left(\frac{x_i s_i}{\theta(\tau_0, n)\mu(x, s)} \right)^{-\frac{1}{2} \log n} = \frac{x^T s}{\theta(\tau_0, n)\mu(x, s)} \\ &\leq \frac{x^T s}{\theta(\tau, n)\mu(x, s)} = \frac{n}{\theta(\tau, n)}. \end{aligned}$$

Applying this inequality to the proximity function $\Phi(x, s, \mu)$, we have

$$\begin{aligned} \Phi(x, s, \theta(\tau, n)\mu(x, s)) &= \frac{n}{2} \left(\frac{1}{\theta(\tau, n)} - 1 \right) + \frac{1}{\log n} \left(\sum_{i=1}^n \left(\frac{x_i s_i}{\theta(\tau, n)\mu(x, s)} \right)^{-\frac{1}{2} \log n} - n \right) \\ &\leq \frac{n}{2} \left(\frac{1}{\theta(\tau, n)} - 1 \right) + \frac{1}{\log n} \left(\frac{n}{\theta(\tau, n)} - n \right) = \eta_1(\tau, n), \end{aligned}$$

which finishes the proof of the lemma. $\square$

We present another lemma that characterizes the behavior of the proximity function $\Phi(x, s, \mu)$.

LEMMA 2.4. *Let $(x, s) > 0$, $n \geq 8$, and $\tau \geq 2$. Then the relation $\Phi(x, s, \mu(x, s)) \leq \eta(\tau, n)$ holds if and only if there exists a parameter $\mu > 0$ such that $\Phi(x, s, \mu) \leq \eta_1(\tau, n)$.*

Proof. The necessary part follows directly from Lemma 2.3, thus it remains to consider the sufficient part. Suppose that the proximity function $\Phi(x, s, \mu)$ has a global minimum at μ^* . The assumption in the lemma implies

$$\Phi(x, s, \mu^*) \leq \eta_1(\tau, n).$$

It follows immediately from Corollary 2.2

$$\Phi(x, s, \mu(x, s)) \leq \eta(\tau, n).$$

This completes the proof of the lemma. $\square$

We close this section with the following corollary, which gives an estimation of μ^* whenever $\Phi(x, s, \mu(x, s)) \leq \eta(\tau, n)$. The corollary follows from Lemma 2.4 and the fact that $\theta(\tau, n)$ is decreasing with respect to τ .

COROLLARY 2.5. *Let $(x, s) > 0$, $n \geq 8$, and $\tau \geq 2$, and suppose that the proximity function $\Phi(x, s, \mu)$ attains its minimum at μ^* . Then $\Phi(x, s, \mu(x, s)) \leq \eta(\tau, n)$ if and only if*

$$\mu^* \geq \theta(\tau, n)\mu(x, s).$$

3. A pseudo predictor-corrector method. In this section, we describe our algorithm and establish its complexity result. The section is divided into three parts. In the first subsection, we describe our new algorithm. The second subsection is devoted to investigating the behavior of the proximity function in the corrector step. In the last part, we estimate the step size used in the predictor step and summarize the complexity result.

3.1. The algorithm. We start with a brief description of standard IPMs. Suppose that the present iterate (x, s) is strictly feasible. For any given duality gap parameter μ , we can apply Newton method to system (2.1) and obtain a search direction via solving the following linear system of equations:

$$\begin{aligned} (3.1) \quad & A\Delta x = 0, \\ & A^T \Delta y + \Delta s = 0, \\ & x\Delta s + s\Delta x = \mu e - xs. \end{aligned}$$

Let us consider the case that v is defined by (2.2) with $\mu = \theta(\tau, n)\mu(x, s)$. Therefore the relation $\Phi(x, s, \theta(\tau, n)\mu(x, s)) = \Psi(v)$ holds. We denote the search direction in the scaled v -space as

$$(3.2) \quad d_x := \frac{v\Delta x}{x} \quad \text{and} \quad d_s := \frac{v\Delta s}{s}.$$

Using this notation and (2.2), we can write system (3.1) in the scaled v -space as

$$\begin{aligned} (3.3) \quad & \bar{A}d_x = 0, \\ & \bar{A}^T \Delta y + d_s = 0, \\ & d_x + d_s = v^{-1} - v, \end{aligned}$$

where $\bar{A} = \frac{1}{\mu}AV^{-1}X$, $V = \text{diag}(v)$, and $X = \text{diag}(x)$. Recall that the so-called primal-dual affine scaling search direction is the solution of the system

$$(3.4) \quad \begin{aligned} \bar{A}d_x &= 0, \\ \bar{A}^T \Delta y + d_s &= 0, \\ d_x + d_s &= -v, \end{aligned}$$

while the search direction induced by the self-regular proximity function $\Psi(v)$ can be computed by solving [13]

$$(3.5) \quad \begin{aligned} \bar{A}d_x &= 0, \\ \bar{A}^T \Delta y + d_s &= 0, \\ d_x + d_s &= v^{-1-\log n} - v = -\nabla \Psi(v). \end{aligned}$$

The following technical result shows that if the present iterate is in the neighborhood $\mathcal{N}$ defined by (2.3), then the duality gap will not increase along the search direction obtained by solving (3.5).

PROPOSITION 3.1. *Let (x, s) be a strictly feasible pair and $\Psi(v) \leq \eta_1(\tau, n)$, where $\mu = \theta(\tau, n)\mu(x, s)$ and $\eta_1(\tau, n)$ is defined by (2.14). Then the solution (d_x, d_s) of system (3.5) satisfies*

$$v^T(d_x + d_s) \leq 0,$$

and the equality holds if and only if $\Psi(v) = \eta_1(\tau, n)$.

Proof. Since $\Psi(v) \leq \eta_1(\tau, n)$, from Lemma 2.4 we obtain

$$\Phi(x, s, \mu(x, s)) \leq \eta(n, \tau).$$

Let μ^* be the global minimizer of $\Phi(x, s, \mu)$ with respect to μ . From Corollary 2.5 we conclude

$$\mu^* \geq \theta(\tau, n)\mu(x, s).$$

The proposition follows immediately from the last conclusion of Lemma 2.1 and (3.5). $\square$

The above proposition indicates that if we take a step along the search direction obtained from (3.5) with $\mu = \mu^*$, then the duality gap remains constant and it decreases if and only if the target value $\mu \leq \mu^*$.

Now we are ready to describe our algorithm. To start the algorithm, we assume that a strictly feasible starting point is available and the starting point is in the neighborhood $\mathcal{N}$ where $\eta(\tau, n)$ is defined by (2.7). Then we use a corrector-type step to reduce the value of the proximity function and thus bring the iterate into a smaller neighborhood of the central path. We would like to alert the reader to the distinction between our corrector step and the standard corrector step in the IPM literature. In a standard corrector step, the search direction is always computed by solving (3.1) with $\mu = \mu(x, s)$ and thus the duality gap remains invariant. However, as stated in Proposition 3.1, the duality gap in a corrector step of Algorithm 3.1 will never increase but might decrease. Each corrector step is followed by a predictor step that aims at reducing the duality gap while keeping the iterate in the large neighborhood $\mathcal{N}$. This process is repeated until the duality gap is small enough.

The procedure of our algorithm is outlined as follows. In the procedure, we use the notation $x(\alpha) = x + \alpha\Delta x$, $y(\alpha) = y + \alpha\Delta y$, and $s(\alpha) = s + \alpha\Delta s$.

Algorithm 3.1

Input:

Proximity parameters $\tau \geq 2$, $\theta(\tau, n)$ given by (2.13),
 and $\eta(\tau, n)$ given by (2.7);
 an accuracy parameter $\varepsilon > 0$;
 $(x, s) = (x^0, s^0)$ such that $\Phi(x, s, \mu(x, s)) \leq \eta(\tau, n)$.

begin

while $x^T s \geq \varepsilon$ **do**

begin**corrector step**

Compute $\mu(x, s)$ and update μ to $\mu := \theta(\tau, n)\mu(x, s)$;
 Solve the system (3.5) for $(\Delta x, \Delta y, \Delta s)$;
 Find a feasible step size α that reduces
 the proximity function sufficiently;^a
 Update the iterate by $x = x(\alpha)$, $y = y(\alpha)$, $s = s(\alpha)$;

Predictor step

Solve the system (3.4) for $(\Delta x, \Delta y, \Delta s)$,
 Find a sufficiently large step size $\alpha \in (0, 1]$
 such that^b $\Phi(x(\alpha), s(\alpha), \mu(x, s)(\alpha)) \leq \eta(\tau, n)$;
 Update the iterate by $x = x(\alpha)$, $y = y(\alpha)$, $s = s(\alpha)$;

end**end**

^aIn Theorem 3.5, we shall specify a step size α that can reduce the proximity function sufficiently.

^bIn Theorem 3.7, we shall specify a step size α that keeps the iterate in the neighborhood.

3.2. The corrector step. In this subsection we estimate the change of the proximity function in the corrector step. For this we first discuss how to choose a suitable step size to reduce the proximity function Φ in the corrector step. Suppose that the current point (x, s) is in the neighborhood $\mathcal{N}$ (given by (2.3)) where $\eta(\tau, n)$ is given by (2.7), and let

$$\sigma = \|\nabla \Psi(v)\|.$$

Note that when $\log n \geq 1$, we have

$$(3.6) \quad \sigma \geq \|v - v^{-1}\|.$$

In what follows we estimate the magnitudes of σ and $v_{\min}$. In particular, we will derive a lower bound of $v_{\min}$, from which we can further get an estimate of the maximal feasible step size.

LEMMA 3.2. *Suppose that the present iterate (x, s) is in the neighborhood $\mathcal{N}$ defined by (2.3) and v is defined by (2.2) with $\mu = \theta(\tau, n)\mu(x, s)$. If $n \geq 8$ and $\tau \geq 2$, then one has*

$$(3.7) \quad v_{\min} \geq \exp^{-\frac{(2\tau-2)\log n}{2+\log n}} \geq \exp^{-\tau},$$

$$(3.8) \quad \sigma \geq \frac{5}{6} v_{\min}^{-1-\log n}.$$

Proof. Using the notation v , we can write $\Phi(x, s, \theta(\tau, n)\mu(x, s)) = \Psi(v)$. When $(x, s) \in \mathcal{N}$, from Lemma 2.3 we obtain $\Psi(v) \leq \eta_1(\tau, n)$, where $\eta_1(\tau, n)$ is given by (2.14). It follows from (2.15) that

$$v_{\min}^{-\log n} < \sum_{i=1}^n v_i^{-\log n} \leq \sum_{i=1}^n (v_i)^2 = \frac{x^T s}{\theta(\tau, n)\mu(x, s)} = n \exp^{\frac{(2\tau-2)\log n}{2+\log n}}.$$

Therefore, one has

$$\log v_{\min}^{-1} < 1 + \frac{2\tau-2}{2+\log n} \leq \tau,$$

which further yields (3.7).

To prove (3.8), it suffices to consider the case where $v_{\min} \leq 1$. Observe that if $n \geq 8$ and $\tau \geq 2$, then from (3.6) and (2.12) we obtain

$$\sigma^2 \geq \|v - v^{-1}\|^2 > \|v\|^2 - 2n = \left(\frac{1}{\theta(\tau, n)} - 2 \right) n \geq (\exp^{\tau-1} - 2)n \geq 6 \quad \forall n \geq 8, \tau \geq 2.$$

This gives (3.8) when

$$v_{\min}^{1+\log n} > \frac{1}{6}.$$

If

$$v_{\min}^{1+\log n} \leq \frac{1}{6},$$

then we have

$$\begin{aligned} v_{\min}^{1+\log n} \sigma &\geq v_{\min}^{1+\log n} (v_{\min}^{-1-\log n} - v_{\min}) \\ &= 1 - v_{\min}^{2+\log n} \geq 1 - v_{\min}^{1+\log n} \\ &\geq \frac{5}{6}. \end{aligned}$$

The proof of the lemma is finished. $\square$

LEMMA 3.3. *Given $t, w \in [0, 1)$. If $-\log(1-t) \geq w$, then one has $t \geq \frac{w}{1+w}$.*

Proof. Since $t \in [0, 1)$, one has

$$w \leq -\log(1-t) = \log \left(1 + \frac{t}{1-t} \right) \leq \frac{t}{1-t},$$

which further gives the desired relation in the lemma. $\square$

The following technical lemma from [13] will be used in our later analysis. For self-completeness, we include it here without proof.

LEMMA 3.4. *Suppose that $h(t)$ is a twice differentiable convex function with $h(0) = 0$, $h'(0) < 0$ and that $h(t)$ attains its global minimum at its stationary point $t^* > 0$. If $h''(t)$ is increasing with respect to t , then for any $t \in [0, t^*]$ we have*

$$h(t) \leq \frac{h'(0)t}{2}.$$

We proceed to specify a step size that reduces the proximity function sufficiently in the corrector step of Algorithm 3.1. We mention that by using Theorem 1 in [13],

we can get a step size α that reduces the proximity function value. However, the estimation in [13] focuses on the relation of the reduction and the proximity function itself, while our following theorem tries to estimate the reduction of the proximity function by using the information of the new updated iterate. For self-completeness, we also present a detailed proof here.

THEOREM 3.5. *Suppose that the present iterate (x, s) is in the neighborhood $\mathcal{N}$ defined by (2.3) and v is defined by (2.2) with $\mu = \theta(\tau, n)\mu(x, s)$. Let (d_x, d_s) be the solution of system (3.5) with $n \geq 8$ and $\tau \geq 2$. Then the step size $\alpha = \frac{v_{\min}}{3\sigma \log n}$ is strictly feasible. Moreover, for this step size, we have*

$$\Phi(x(\alpha), s(\alpha), \mu) \leq \Phi(x(0), s(0), \mu) - \frac{5}{66 \log n} \max\{v_{\min}\sigma, v_{\min}^{-\log n}(\alpha)\}.$$

Proof. To prove the theorem, we need to estimate the maximal feasible step size $\alpha_{\max}$. By using the orthogonality of d_x and d_s , we have

$$(3.9) \quad \max\{\|d_x\|^2, \|d_s\|^2\} \leq \|d_x\|^2 + \|d_s\|^2 = \|d_x + d_s\|^2 = \sigma^2.$$

From (3.2) we conclude that the strict feasibility of $(x + \alpha\Delta x, s + \alpha\Delta s)$ can be retained if and only if $(v + \alpha d_x, v + \alpha d_s)$ is strictly positive. This is certainly true if $v_{\min} - \alpha\sigma > 0$. This gives

$$\alpha_{\max} \geq v_{\min}\sigma^{-1} \geq \sigma^{-1} \exp^{-\tau}.$$

We progress to investigate the behavior of the function $\Phi(x(\alpha), s(\alpha), \theta(\tau, n)\mu(x, s))$ along the search direction $(\Delta x, \Delta s)$ obtained by solving (3.5). Let us define

$$v(\alpha) = \sqrt{(v + \alpha d_x)(v + \alpha d_s)}.$$

Using this, we can write the proximity function $\Phi(x(\alpha), s(\alpha), \theta(\tau, n)\mu(x, s))$ in the scaled v -space as

$$\Phi(x(\alpha), s(\alpha), \theta(\tau, n)\mu(x, s)) = \Psi(v(\alpha)).$$

Note that from the choice of $\Psi(v(\alpha))$, we have

$$\begin{aligned} \Psi(v(\alpha)) &= \frac{(v + \alpha d_x)^T(v + \alpha d_s)}{2} - \frac{n}{2} + \frac{1}{\log n} \sum_{i=1}^n (v_i(\alpha)^{-\log n} - 1) \\ &\leq \frac{1}{2} \|v\|^2 + \alpha v^T(d_x + d_s) - \frac{n}{2} \\ &\quad + \frac{1}{2 \log n} \sum_{i=1}^n ((v + \alpha d_x)_i^{-\log n} + (v + \alpha d_s)_i^{-\log n} - 2), \end{aligned}$$

where the inequality follows from the orthogonality of d_x and d_s as well as the self-regularity of the function $\psi(t)$. Let us cast the difference between the proximity functions before and after a step as a function of the step size α ; then from (3.5) we obtain

$$\begin{aligned} g(\alpha) &:= \Psi(v(\alpha)) - \Psi(v(0)) \\ &\leq -\alpha \sum_{i=1}^n (v_i^2 - v_i^{-\log n}) \\ &\quad + \frac{1}{2 \log n} \sum_{i=1}^n ((v + \alpha d_x)_i^{-\log n} + (v + \alpha d_s)_i^{-\log n} - 2v_i^{-\log n}). \end{aligned}$$

One can easily verify that $g(\alpha)$ is convex. Furthermore, by simple calculus one gets

$$g(0) = 0 \quad \text{and} \quad g'(0) = -\sigma^2.$$

For any $\alpha \leq \alpha_{\max}$, we have

$$\begin{aligned} g''(\alpha) &= (1 + \log n) \sum_{i=1}^n ((d_x^i)^2 (v + \alpha d_x)_i^{-2-\log n} + (d_s^i)^2 (v + \alpha d_s)_i^{-2-\log n}) \\ &\leq (1 + \log n) \sum_{i=1}^n ((d_x^i)^2 (v_i - \alpha\sigma)^{-2-\log n} + (d_s^i)^2 (v_i - \alpha\sigma)^{-2-\log n}) \\ &\leq (1 + \log n) \sigma^2 (v_{\min} - \alpha\sigma)^{-2-\log n}, \end{aligned}$$

where the inequalities follow from the fact $\sigma = \|d_x + d_s\|$ and

$$\min_{i=1,\dots,n} \{v_i + \alpha d_x^i, v_i + \alpha d_s^i\} \geq \min_{i=1,\dots,n} v_i - \alpha\sigma \geq v_{\min} - \alpha\sigma.$$

Let

$$g_1(\alpha) = -\alpha\sigma^2 + \int_0^\alpha \int_0^\zeta (1 + \log n) \sigma^2 (v_{\min} - t\sigma)^{-2-\log n} dt d\zeta.$$

It is easy to see that $g_1(\alpha)$ is also convex in the feasible region with

$$g_1(0) = 0 \quad \text{and} \quad g_1'(0) = -\sigma^2.$$

Moreover, since $g_1''(\alpha) \geq g''(\alpha)$, one has

$$g(\alpha) \leq g_1(\alpha).$$

In what follows we estimate the step size α^* at which $g_1(\alpha)$ has a global minimum. From the convexity of $g_1(\alpha)$ it follows that α^* must satisfy $g_1'(\alpha) = 0$, which gives

$$(3.10) \quad -\sigma^2 + \sigma((v_{\min} - \alpha^*\sigma)^{-1-\log n} - v_{\min}^{-1-\log n}) = 0,$$

or equivalently

$$-\sigma v_{\min}^{1+\log n} + ((1 - \alpha^* v_{\min}^{-1} \sigma)^{-1-\log n} - 1) = 0.$$

Using (3.8), we derive

$$(3.11) \quad \frac{11}{6} \leq (1 - \alpha^* v_{\min}^{-1} \sigma)^{-1-\log n},$$

which is equal to

$$-(1 + \log n) \log(1 - \alpha^* v_{\min}^{-1} \sigma) \geq \log \frac{11}{6} > \frac{3}{5}.$$

Since $\log n \geq 2$, the above relation gives

$$-\log(1 - \alpha^* v_{\min}^{-1} \sigma) \geq \frac{3}{5 + 5 \log n} \geq \frac{2}{5 \log n}.$$

Using Lemma 3.3 and the fact that $\log n \geq 2$, we derive

$$\alpha^* v_{\min}^{-1} \sigma \geq \frac{2}{5 \log n + 2} \geq \frac{1}{3 \log n}.$$

Now let us focus on the case where the step size

$$(3.12) \quad \alpha = \frac{v_{\min}}{3 \sigma \log n}.$$

By using Lemma 3.4, we have

$$(3.13) \quad \Psi(v(\alpha)) \leq \Psi(v) + g_1(\alpha) \leq \Psi(v) - \frac{\alpha \sigma^2}{2} \leq \Psi(v) - \frac{v_{\min} \sigma}{6 \log n}.$$

On the other hand, because $\alpha \leq \alpha^*$, it must hold $g'_1(\alpha) \leq 0$, i.e.,

$$(3.14) \quad -\sigma^2 + \sigma((v_{\min} - \alpha \sigma)^{-1 - \log n} - v_{\min}^{-1 - \log n}) \leq 0.$$

It follows that

$$(v_{\min} - \alpha \sigma)^{-1 - \log n} \leq \sigma + v_{\min}^{-1 - \log n} \leq \frac{11}{5} \sigma,$$

where the last inequality is given by (3.8). From (3.13) we immediately obtain

$$\Psi(v(\alpha)) - \Psi(v) \leq -\frac{v_{\min} \sigma}{6 \log n} \leq -\frac{5 v_{\min}}{66 \log n} (v_{\min} - \alpha \sigma)^{-1 - \log n} \leq -\frac{5 (v_{\min} - \alpha \sigma)^{-\log n}}{66 \log n}.$$

Notice that for all feasible step size α , we have

$$v_{\min}(\alpha) = \min_{1 \leq i \leq n} \sqrt{(v + \alpha d_x)(v + \alpha d_s)} \geq v_{\min} - \alpha \sigma.$$

It follows from the above two relations that

$$\Psi(v(\alpha)) \leq \Psi(v) - \frac{5 v_{\min}^{-\log n}(\alpha)}{66 \log n}.$$

Combining the above inequality with (3.13), we obtain the desired inequality in the theorem. $\square$

3.3. The predictor step and complexity of the algorithm. In this subsection we estimate the step size used in the predictor step that keeps the new iterate in the neighborhood $\mathcal{N}$ defined by (2.3). To guarantee the resulting iterate in a certain neighborhood, it suffices to show that the value of the proximity function for a strictly feasible step size is bounded above by $\eta(\tau, n)$ or $\eta_1(\tau, n)$, depending on the duality parameter μ used in the definition of the proximity function. In order to distinguish the iterates before and after a corrector step, let (x^+, s^+) denote the primal-dual pair after the corrector step. It should be mentioned that in the predictor step, we have the freedom to choose the reference duality gap parameter μ because in the predictor step, the resulting search directions Δx and Δs in the original x - and s -spaces are completely independent of the parameter μ . Let μ^+ be the reference duality gap parameter used in the predictor step. Correspondingly, the scaled vector v^+ can be defined as

$$(3.15) \quad v^+ = \sqrt{\frac{x^+ s^+}{\mu^+}} \quad \text{and} \quad (v^+)^{-1} = \sqrt{\frac{\mu^+ e}{x^+ s^+}}.$$

In what follows we will discuss how to choose an appropriate duality gap parameter μ^+ such that the resulting scaled vector v^+ satisfies the following inequalities:

$$(3.16) \quad \|v^+\|^2 \leq \frac{n}{\theta(\tau, n)},$$

$$(3.17) \quad \sum_{i=1}^n (v_i^+)^{-\log n} - \|v^+\|^2 \leq 0,$$

$$(3.18) \quad \Psi(v^+) \leq \Phi(x, s, \theta(\tau, n)\mu(x, s)) - \frac{5}{66 \log n} (v_{\min}^+)^{-\log n}.$$

The following lemma proves the existence of such a parameter μ^+ .

LEMMA 3.6. *Let $n \geq 8$, $\tau \geq 2$, and (x^+, s^+) be the iterate after the corrector step and $\mu(x, s)^+ = (x^+)^T s^+ / n$. Then there exists a duality gap parameter μ^+ such that the vector v^+ given by (3.15) satisfies inequalities (3.16)–(3.18).*

Proof. To prove the lemma, we first observe that if

$$(3.19) \quad \sum_{i=1}^n \left(\frac{x_i^+ s_i^+}{\theta(\tau, n)\mu(x, s)} \right)^{-\frac{1}{2} \log n} - \frac{(x^+)^T s^+}{\theta(\tau, n)\mu(x, s)} \leq 0,$$

then we can set $\mu^+ = \theta(\tau, n)\mu(x, s)$, the same as in the corrector step. In this case, (3.19) gives (3.17) and (3.18) follows from Theorem 3.5. Moreover, as indicated by Proposition 3.1, the duality gap will not increase in the corrector step; this implies $\mu^+ \geq \theta(\tau, n)\mu(x, s)^+$, and thus inequality (3.16) holds as well.

Thus it remains to consider the case where

$$(3.20) \quad \sum_{i=1}^n \left(\frac{x_i^+ s_i^+}{\theta(\tau, n)\mu(x, s)} \right)^{-\frac{1}{2} \log n} - \frac{(x^+)^T s^+}{\theta(\tau, n)\mu(x, s)} > 0$$

after the corrector step. Now let us cast $\Phi(x^+, s^+, \mu)$ as a function of μ . Since $\Phi(x^+, s^+, \mu)$ is a convex function of μ , it must attain its global minimum at some point μ^* satisfying

$$\sum_{i=1}^n \left(\frac{x_i^+ s_i^+}{\mu^*} \right)^{-\frac{1}{2} \log n} - \frac{(x^+)^T s^+}{\mu^*} = 0.$$

Because $\log n \geq 2$, (3.20) implies

$$\mu^* < \theta(\tau, n)\mu(x, s).$$

On the other hand, if we choose $\mu^+ = \mu^*$ and define v^+ by (3.15), then from Theorem 3.5 we obtain

$$\begin{aligned} \Psi(v^+) &\leq \Phi(x^+, s^+, \theta(\tau, n)\mu(x, s)) \\ &\leq \Phi(x, s, \theta(\tau, n)\mu(x, s)) - \frac{5}{66 \log n} \max_{i=1, \dots, n} \left\{ \left(\sqrt{\frac{x_i^+ s_i^+}{\theta(\tau, n)\mu(x, s)}} \right)^{-\log n} \right\} \\ &\leq \Phi(x, s, \theta(\tau, n)\mu(x, s)) - \frac{5}{66 \log n} \max_{i=1, \dots, n} \left\{ \left(\sqrt{\frac{x_i^+ s_i^+}{\mu^*}} \right)^{-\log n} \right\} \\ &= \Phi(x, s, \theta(\tau, n)\mu(x, s)) - \frac{5}{66 \log n} (v_{\min}^+)^{-\log n}. \end{aligned}$$

It remains to verify that (3.16) holds for this special choice. Note that because

$$\Phi(x^+, s^+, \mu^+) < \eta_1(\tau, n),$$

by Lemma 2.4, $\Phi(x^+, s^+, (x^+)^T s^+ / n) = \eta(\tau_0, n) < \eta(\tau, n)$ for some $\tau_0 < \tau$. By the definition of $\theta(\cdot, \cdot)$, we have $\mu^+ = \theta(\tau_0, n)(x^+)^T s^+ / n$. Since $\theta(\cdot, n)$ is a decreasing function of τ , we have $\theta(\tau_0, n) > \theta(\tau, n)$ and $\mu^+ \geq \theta(\tau, n)(x^+)^T s^+ / n$. Moreover, one has

$$\frac{(x^+)^T s^+}{\mu^+} = \frac{n}{\theta(\tau_0, n)} < \frac{n}{\theta(\tau, n)},$$

which implies (3.16). This completes the proof of the lemma. $\square$

Now we can progress to estimate the increase of the proximity function in the predictor step. To release the notational load, we drop the subscript $+$ used in our previous discussion and simply denote the present iterate by (x, s) and correspondingly the scaled vector by v . We emphasize here again that the scaled vector v satisfies all the inequalities (3.16), (3.17), and (3.18). We start by estimating $v_{\min}$ for the scaled vector v . Since v satisfies all the inequalities from (3.16)–(3.18), by following an analogue chain of reasoning as in the proof of the inequality (3.7) in Lemma 3.2, we obtain

$$(3.21) \quad v_{\min} \geq \exp^{-\tau}.$$

Now we can state the main result of this section.

THEOREM 3.7.

Let (d_x, d_s) be the solution of system (3.4) with the vector v satisfying properties (3.16), (3.17), and (3.18). If $n \geq 8$ and $\tau \geq 2$, then the step size

$$\alpha = \frac{\exp^{1-2\tau}}{15\sqrt{n} \log n}$$

is strictly feasible. Moreover, the new iterate updated with this step size is still in the neighborhood defined by (2.3).

Proof. We start our discussion on the value of the function $\Psi(v(\alpha))$ for a strictly feasible step size α . Note that in the predictor step, all the inequalities (3.16), (3.17), and (3.18) hold. Therefore, we have

$$(3.22) \quad \max\{\|d_x\|, \|d_s\|\} \leq \|d_x + d_s\| = \|v\| \leq \sqrt{\frac{n}{\theta(\tau, n)}} \leq \exp^{\tau-1} \sqrt{n},$$

where the last inequality follows from (2.11). It follows immediately that

$$\alpha_{\max} \geq v_{\min} \sqrt{\frac{\theta(\tau, n)}{n}} \geq \frac{\exp^{1-\tau} v_{\min}}{\sqrt{n}}.$$

By following a similar reasoning chain as in our discussion in the corrector step, we can show that

$$\begin{aligned} & \Psi(v(\alpha)) - \Psi(v(0)) \\ & \leq -\alpha \|v\|^2 + \frac{1}{2 \log n} \sum_{i=1}^n ((v + \alpha d_x)_i^{-\log n} + (v + \alpha d_s)_i^{-\log n} - 2v_i^{-\log n}). \end{aligned}$$

Let

$$h(\alpha) := -\alpha \|v\|^2 + \frac{1}{2 \log n} \sum_{i=1}^n ((v + \alpha d_x)_i^{-\log n} + (v + \alpha d_s)_i^{-\log n} - 2v_i^{-\log n}).$$

It is easy to verify that $h(\alpha)$ is a convex function of α . Furthermore, by simple calculus one gets

$$h(0) = 0 \quad \text{and} \quad h'(0) = \sum_{i=1}^n v_i^{-\log n} - \|v\|^2.$$

From (3.17), we can claim that $h'(0) \leq 0$. Moreover, for any $0 < \alpha \leq \alpha_{\max}$, one has

$$\begin{aligned} h''(\alpha) &= (1 + \log n) \sum_{i=1}^n ((d_x^i)^2 (v + \alpha d_x)_i^{-2-\log n} + (d_s^i)^2 (v + \alpha d_s)_i^{-2-\log n}) \\ &\leq (1 + \log n) \sum_{i=1}^n (v_i - \alpha \|v\|)^{-2-\log n} ((d_x^i)^2 + (d_s^i)^2) \\ &\leq (1 + \log n) \|v\|^2 (v_{\min} - \alpha \|v\|)^{-2-\log n}. \end{aligned}$$

It follows that for any $\alpha \in (0, \alpha_{\max}]$,

$$\begin{aligned} \Psi(v(\alpha)) &\leq \Psi(v) + \int_0^\alpha \int_0^\zeta (1 + \log n) \|v\|^2 (v_{\min} - t \|v\|)^{-2-\log n} dt d\zeta \\ &< \Psi(v) + \int_0^\alpha \|v\| ((v_{\min} - \zeta \|v\|)^{-1-\log n}) d\zeta \\ &= \Psi(v) + \frac{(v_{\min} - \alpha \|v\|)^{-\log n} - v_{\min}^{-\log n}}{\log n}. \end{aligned}$$

By using (3.18) we obtain

$$\Psi(v) \leq \eta_1(\tau, n) - \frac{5}{66 \log n} v_{\min}^{-\log n}.$$

Therefore, for any step size α satisfying

$$(3.23) \quad (v_{\min} - \alpha \|v\|)^{-\log n} \leq \frac{5v_{\min}^{-\log n}}{66} + v_{\min}^{-\log n},$$

one has

$$\Psi(v(\alpha)) \leq \eta_1(\tau, n),$$

which, by Lemma 2.4, further implies that

$$\Phi(x(\alpha), s(\alpha), \mu(x, s)(\alpha)) \leq \eta(\tau, n).$$

Thus for any step size α satisfying (3.23), the new iterate $(x(\alpha), s(\alpha))$ is still in the neighborhood defined by (2.3). Now we are going to estimate the maximal step size α^* that satisfies (3.23), which is the unique solution of the following equation:

$$(3.24) \quad (1 - \alpha v_{\min}^{-1} \|v\|)^{-\log n} = \frac{71}{66}.$$

The above relation implies that

$$(3.25) \quad \begin{aligned} -\log(1 - \alpha^* v_{\min}^{-1} \|v\|) &= \frac{\log \frac{71}{66}}{\log n} \\ &\geq \frac{1}{14 \log n}. \end{aligned}$$

Using Lemma 3.3 and the fact that $\log n \geq 1$, we have

$$\alpha^* v_{\min}^{-1} \|v\| \geq \frac{1}{1 + 14 \log n} > \frac{1}{15 \log n}.$$

It follows immediately that

$$\begin{aligned} \alpha^* &\geq \frac{v_{\min}}{15 \|v\| \log n} \geq \frac{\exp^{1-\tau} v_{\min}}{15 \sqrt{n} \log n} \\ &\geq \frac{\exp^{1-2\tau}}{15 \sqrt{n} \log n}, \end{aligned}$$

where the second inequality follows from (3.22) and the last inequality from (3.21). The proof of the theorem is finished. $\square$

We close this section by giving an upper bound for the number of iterations of Algorithm 3.1. As indicated in Proposition 3.1, the duality gap will not increase in the corrector step. Theorem 3.7 shows that the duality gap will decrease at least at a rate of

$$1 - \frac{\exp^{1-2\tau}}{15 \sqrt{n} \log n}$$

in each predictor step. As a direct consequence of these two results, we have the following theorem.

THEOREM 3.8. *If $n \geq 8$ and $\tau \geq 2$, then after at most*

$$\left\lceil 15 \exp^{2\tau-1} \sqrt{n} \log n \log \frac{(x^0)^T s^0}{\epsilon} \right\rceil$$

iterations, Algorithm 3.1 will find an approximate solution satisfying $x^T s \leq \epsilon$.

We note that, although Algorithm 3.1 has an $O(\sqrt{n} \log n \log \frac{(x^0)^T s^0}{\epsilon})$ complexity, the upper bound of the number of total iterations given by Theorem 3.8 is exponential in τ .

4. Local superlinear convergence. In this section we prove the local superlinear convergence of the algorithm. Let (x_-, s_-) denote the primal-dual pair before the corrector step and (x, s) the new pair after a corrector step. Recall that for the predictor direction, Ye and Anstreicher [28] have proved the following relation for monotone complementarity problems:

$$(4.1) \quad |\Delta x_i \Delta s_i| = O((\mu(x, s))^2), \quad i = 1, \dots, n,$$

when the product $x^T s$ (or $\mu(x, s)$) is sufficiently small. Since an LO problem can be cast as a special class of monotone linear complementarity problems, the above relation remains valid for LO if $\mu(x, s)$ is small enough. To simplify the analysis, we

ignore the hidden constant in the big-O relation. Now let us recall the choice of the parameter μ in the predictor step in our algorithm. We then have

$$(4.2) \quad |d_x^i d_s^i| = \frac{|\Delta x_i \Delta s_i|}{\mu} \leq \frac{|\Delta x_i \Delta s_i|}{\theta(\tau, n) \mu(x, s)} = O(\mu(x, s)),$$

where the inequality follows from Lemma 3.6. We next give a technical result about the step size used in the predictor step.

LEMMA 4.1. *Let $n \geq 8$, $\tau \geq 2$, $r \in (0, 1)$, and (x, s) be an iterate in Algorithm 3.1. If the present duality gap $x^T s$ is sufficiently small so that the relation (4.1) holds, then the step size α used in the predictor step satisfies $\alpha \geq 1 - O((\mu(x, s))^r)$.*

Proof. We start with an estimation of the maximal feasible step size. As we mentioned in section 3, the new iterate is strictly feasible if and only if both $v + \alpha d_x$ and $v + \alpha d_s$ are strictly feasible. Further, the maximal feasible step size $\alpha_{\max}$ satisfies the condition

$$(v + \alpha d_x)(v + \alpha d_s) = v^2 - \alpha v^2 + \alpha^2 d_x d_s \geq 0, \quad \alpha \in [0, \alpha_{\max}],$$

or

$$e - \alpha + \alpha^2 v^{-2} d_x d_s \geq 0 \quad \forall \alpha \in [0, \alpha_{\max}].$$

Since $v_{\min} \geq \exp^{-\tau}$ and $|d_x d_s| = O(\mu(x, s))$, we can conclude that $\alpha_{\max} \geq 1 - O(\mu(x, s))$.

Now we are going to show that for any fixed $0 < r < 1$, the step size α used in the predictor step also satisfies the relation

$$(4.3) \quad \alpha \geq 1 - O((\mu(x, s))^r).$$

To prove (4.3), we notice that for any strictly feasible step size $\alpha \in (0, 1]$, from the choice of μ and the corresponding scaled vector v we obtain

$$\begin{aligned} & \Phi(x(\alpha), s(\alpha), (1 - \alpha)\mu) - \Phi(x(0), s(0), \mu) \\ & \leq \frac{1}{\log n} \sum_{i=1}^n \left(\left(\frac{v_i(\alpha)}{\sqrt{1 - \alpha}} \right)^{-\log n} - v_i^{-\log n} \right) \\ & = \sum_{i=1}^n \frac{v_i^{-\log n}}{\log n} \left(\left(\frac{v_i(\alpha)}{\sqrt{1 - \alpha} v_i} \right)^{-\log n} - 1 \right) \\ & = \sum_{i=1}^n \frac{v_i^{-\log n}}{\log n} \left(\left(\frac{v_i^2(\alpha)}{(1 - \alpha) v_i^2} \right)^{-\frac{1}{2} \log n} - 1 \right) \\ & = \sum_{i=1}^n \frac{v_i^{-\log n}}{\log n} \left(\left(1 + \frac{\alpha^2}{1 - \alpha} v_i^{-2} d_x^i d_s^i \right)^{-\frac{1}{2} \log n} - 1 \right). \end{aligned}$$

Let $\mathcal{I}_-$ be the index set

$$\mathcal{I}_- = \{i \in \{1, \dots, n\} : d_x^i d_s^i < 0\}.$$

It follows immediately that

$$\begin{aligned}
 & \Phi(x(\alpha), s(\alpha), (1 - \alpha)\mu) - \Phi(x(0), s(0), \mu) \\
 & \leq \sum_{i \in \mathcal{I}_-} \frac{v_i^{-\log n}}{\log n} \left(\left(1 + \frac{\alpha^2}{1 - \alpha} v_i^{-2} d_x^i d_s^i \right)^{-\frac{1}{2} \log n} - 1 \right) \\
 & \leq \frac{v_{\min}^{-\log n}}{\log n} \sum_{i \in \mathcal{I}_-} \left(\left(1 + \frac{\alpha^2}{1 - \alpha} v_i^{-2} d_x^i d_s^i \right)^{-\frac{1}{2} \log n} - 1 \right) \\
 & \leq \frac{v_{\min}^{-\log n}}{\log n} \left(\left(1 + \sum_{i \in \mathcal{I}_-} \frac{\alpha^2}{1 - \alpha} v_i^{-2} d_x^i d_s^i \right)^{-\frac{1}{2} \log n} - 1 \right) \\
 & \leq \frac{v_{\min}^{-\log n}}{\log n} \left(\left(1 - \frac{\alpha^2}{1 - \alpha} O(\mu(x, s)) \right)^{-\frac{1}{2} \log n} - 1 \right).
 \end{aligned}$$

This inequality indicates that, for any fixed $r \in (0, 1)$ and sufficiently small $\mu(x, s)$, there exists a step size $\alpha \geq 1 - O((\mu(x, s))^r)$ such that

$$\Phi(x(\alpha), s(\alpha), (1 - \alpha)\mu) - \Phi(x(0), s(0), \mu) \leq \frac{5v_{\min}^{-\log n}}{66 \log n},$$

as stated in Theorem 3.5. This completes the proof of the lemma. $\square$

It is straightforward to verify that if the step size used in the predictor step satisfies $\alpha \geq 1 - O((\mu(x, s))^r)$, then after the predictor step we have

$$\mu_{gap}^+ = (1 - \alpha)\mu(x, s) = O((\mu(x, s))^{1+r}).$$

Now we can state the main result of this section.

THEOREM 4.2. *Let $n \geq 8$, $\tau \geq 2$, and (x^k, s^k) be generated by Algorithm 3.1. The algorithm is superlinearly convergent in the sense that $\mu_{gap}^{k+1} = O((\mu_{gap}^k)^{1+r})$ for any fixed $r \in (0, 1)$ and every accumulation point of the sequence (x^k, s^k) is a strictly complementary solution of the problem.*

Proof. The superlinear convergence of the algorithm follows from Lemma 4.1. The convergence properties of every accumulation point of the iterates can be proved by using the properties of the central path, which converges to a strictly complementarity solution of the underlying problem. The details of the proof are omitted here and we refer the readers to [27] for analogous discussions. $\square$

5. Conclusions. A new predictor-corrector algorithm working in a large neighborhood of the central path is proposed for linear optimization. This algorithm maintains the simple structure of the MTY method, that is, each iteration consists of one predictor step and only one corrector step, and uses only first-order information of the problem. We prove that the algorithm retains local superlinear convergence and has an $O(\sqrt{n} \log n \log \frac{(x^0)^T s^0}{\epsilon})$ iteration bound. This complexity result improves the so-far best known complexity $O(n \log \frac{(x^0)^T s^0}{\epsilon})$ for first-order predictor-corrector algorithms working in large neighborhoods [15].

We mention that it is possible to extend the results in this paper to other cases, e.g., predictor-corrector methods for complementarity problems. The details for such an extension are left to the interested reader.

Acknowledgments. The authors would like to thank two anonymous referees for their useful comments. They are grateful for Dr. C. Roos for his invaluable suggestions that led to substantial improvement in the presentation of the paper.

REFERENCES

- [1] E. D. ANDERSEN, J. GONDZIO, Cs. MÉSZÁROS, AND X. XU, *Implementation of interior point methods for large scale linear programs*, in Interior Point Methods of Mathematical Programming, T. Terlaky, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996, pp. 189–252.
- [2] K. M. ANSTREICHER AND R. A. BOSCH, *A new infinity-norm path following algorithm for linear programming*, SIAM J. Optim., 5 (1995), pp. 236–246.
- [3] C. C. GONZAGA, *Complexity of predictor-corrector algorithms for LCP based on a large neighborhood of the central path*, SIAM J. Optim., 10 (1999), pp. 183–194.
- [4] C. C. GONZAGA AND R. A. TAPIA, *On the convergence of the Mizuno–Todd–Ye algorithm to the analytical center of the solution set*, SIAM J. Optim., 7 (1997), pp. 47–65.
- [5] P. F. HUANG AND Y. YE, *An asymptotical $O(\sqrt{n}L)$ -iteration path-following linear programming algorithm that uses wide neighborhoods*, SIAM J. Optim., 6 (1996), pp. 570–586.
- [6] B. JANSEN, C. ROOS, T. TERLAKY, AND J. P. VIAL, *Primal-dual path-following algorithms for linear programming*, Ann. Oper. Res., 21 (1996), pp. 197–232.
- [7] B. JANSEN, C. ROOS, T. TERLAKY, AND Y. YE, *Improved complexity using higher-order corrections for primal-dual Dikin affine scaling*, Math. Programming, Series B, 76 (1997), pp. 117–130.
- [8] N. K. KARMARKAR, *A new polynomial-time algorithm for linear programming*, Combinatorica, 4 (1984), pp. 373–395.
- [9] M. KOJIMA, N. MEGIDDO, AND S. MIZUNO, *Theoretical convergence of large-step primal-dual interior point algorithms for linear programming*, Math. Programming, Series A, 59 (1993), pp. 1–21.
- [10] Z.-Q. LUO, J. F. STURM, AND S. ZHANG, *Superlinear convergence of a symmetric primal-dual path-following algorithm for semidefinite programming*, SIAM J. Optim., 8 (1998), pp. 59–81.
- [11] S. MIZUNO, M. J. TODD, AND Y. YE, *On adaptive-step primal-dual interior-point algorithms for linear programming*, Math. Oper. Res., 18 (1993), pp. 964–981.
- [12] R. C. MONTERIO, I. ADLER, AND M. G. C. RESENDE, *A polynomial-time primal-dual affine scaling algorithm for linear and convex quadratic programming and its power series extension*, Math. Oper. Res., 15 (1990), pp. 191–214.
- [13] J. M. PENG, C. ROOS, AND T. TERLAKY, *Self-regular proximities and new search directions for linear and semidefinite optimization*, Math. Programming, Series A, 93 (2002), pp. 129–171.
- [14] J. M. PENG AND T. TERLAKY, *A dynamic large-update primal-dual interior-point method for linear optimization*, Optim. Methods Softw., 17 (2002), pp. 1077–1104.
- [15] F. A. POTRA, *The Mizuno–Todd–Ye algorithm in a larger neighborhood of the central path*, European J. Oper. Res., 143 (2002), pp. 257–267.
- [16] F. A. POTRA AND R. SHENG, *A large-step infeasible-interior-point method for the P_* -matrix LCP*, SIAM J. Optim., 7 (1997), pp. 318–335.
- [17] F. A. POTRA, *A Superlinear Convergent Predictor-Corrector Method for Degenerate LCP in a Wide Neighborhood of the Central Path with $O(\sqrt{n}L)$ -Iteration Complexity*, Technical report, Department of Mathematics and Statistics, University of Maryland, College Park, MD, 2001.
- [18] J. RENEGAR, *A Mathematical View of Interior-Point Methods in Convex Optimization*, MPS/SIAM Ser. Optim. 3, SIAM, Philadelphia, 2001.
- [19] C. ROOS, T. TERLAKY, AND J. PH. VIAL, *Theory and Algorithms for Linear Optimization: An Interior Approach*, John Wiley & Sons, Chichester, UK, 1997.
- [20] J. STOER AND M. WECHS, *The complexity of high-order predictor-corrector methods for solving sufficient linear complementarity problems*, Optim. Methods Softw., 10 (1998), pp. 393–417.
- [21] J. STOER, W. WECHS, AND S. MIZUNO, *High order infeasible-interior-point methods for solving sufficient linear complementarity problems*, Math. Oper. Res., 23 (1998), pp. 832–862.
- [22] J. STURM, *Theory and algorithms of semidefinite programming*, in High Performance Optimization, H. Frenk, C. Roos, T. Terlaky, and S. Zhang, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000, pp. 1–194.
- [23] J. STURM AND S. ZHANG, *On a wide region of centers and primal-dual interior-point algorithms*

- for linear programming, *Math. Oper. Res.*, 22 (1997), pp. 408–431.
- [24] J. SUN AND G. ZHAO, *Global linear and local quadratic convergence of a long-step adaptive-mode interior point method for some monotone variational inequality problems*, *SIAM J. Optim.*, 8 (1998), pp. 123–139.
- [25] L. TÜNCEL, *On the convergence of primal-dual interior-point methods with wide neighborhoods*, *Comput. Optim. Appl.*, 4 (1995), pp. 139–158.
- [26] S. J. WRIGHT, *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia, 1997.
- [27] Y. YE, *Interior-Point Algorithms, Theory and Analysis*, John Wiley and Sons, Chichester, UK, 1997.
- [28] Y. YE AND K. ANSTREICHER, *On quadratic and $O(\sqrt{n}L)$ convergence of a predictor-corrector algorithm for LCP*, *Math. Programming, Series A*, 62 (1993), pp. 537–551.
- [29] Y. YE, O. GÜLER, R. A. TAPIA, AND Y. ZHANG, *A quadratically convergent $O(\sqrt{n}L)$ -iteration algorithm for linear programming*, *Math. Programming, Series A*, 59 (1993), pp. 151–162.
- [30] Y. ZHANG AND D. ZHANG, *On polynomiality of the Mehrotra-type predictor-corrector interior point algorithms*, *Math. Programming, Series A*, 68 (1995), pp. 303–318.
- [31] G. ZHAO, *Interior-point algorithms for linear complementarity problems based on large neighborhoods of the central path*, *SIAM J. Optim.*, 8 (1998), pp. 397–413.

ON SADDLE POINTS OF AUGMENTED LAGRANGIANS FOR CONSTRAINED NONCONVEX OPTIMIZATION*

X. L. SUN[†], D. LI[‡], AND K.I.M. MCKINNON[§]

Abstract. We present in this paper new results on the existence of saddle points of augmented Lagrangian functions for constrained nonconvex optimization. Four classes of augmented Lagrangian functions are considered: the essentially quadratic augmented Lagrangian, the exponential-type augmented Lagrangian, the modified barrier augmented Lagrangian, and the penalized exponential-type augmented Lagrangian. We first show that under second-order sufficiency conditions, all these augmented Lagrangian functions possess local saddle points. We then prove that global saddle points of these augmented Lagrangian functions exist under certain mild additional conditions. The results obtained in this paper provide a theoretical foundation for the use of augmented Lagrangians in constrained global optimization. Our findings also give new insights to the role played by augmented Lagrangians in local duality theory of constrained nonconvex optimization.

Key words. nonconvex optimization, constrained global optimization, saddle point and duality, augmented Lagrangian functions

AMS subject classifications. 90C26, 90C46

DOI. 10.1137/030602770

1. Introduction. Consider the following inequality-constrained optimization problem:

$$\begin{aligned} \text{(P)} \quad & \min f(x) \\ & \text{subject to } g_i(x) \leq 0, \quad i = 1, \dots, m, \\ & x \in X, \end{aligned}$$

where f and g_i are twice continuously differentiable functions, and X is a nonempty closed set in $\mathbb{R}^n$.

By associating with each constraint in (P) a nonnegative λ_i , the Lagrangian function of (P) is defined as

$$L(x, \lambda) = f(x) + \sum_{i=1}^m \lambda_i g_i(x).$$

The dual function of (P) is given by

$$(1.1) \quad d(\lambda) = \min_{x \in X} L(x, \lambda).$$

*Received by the editors December 29, 2003; accepted for publication (in revised form) November 16, 2004; published electronically July 26, 2005. This research was partially supported by the Research Committee of the Chinese University of Hong Kong, grant 2050252, by the Research Grants Council of Hong Kong, grant CUHK4214/01E, and by the National Natural Science Foundation of China under grants 79970107 and 10271073.

<http://www.siam.org/journals/siopt/15-4/60277.html>

[†]Department of Mathematics, Shanghai University, Baoshan 200444, Shanghai, China (xlsun@staff.shu.edu.cn).

[‡]Corresponding author. Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N. T., Hong Kong (dli@se.cuhk.edu.hk).

[§]School of Mathematics, The University of Edinburgh, King's Buildings, Edinburgh EH9 3JZ, UK (ken@maths.ed.ac.uk).

It is well known that for any feasible x and $\lambda \geq 0$, the weak duality relation $d(\lambda) \leq f(x)$ always holds. The Lagrangian dual problem of (P) is then to search a multiplier vector $\lambda^* \geq 0$ which maximizes $d(\lambda)$ for all $\lambda \geq 0$:

$$(D) \quad \max_{\lambda \geq 0} d(\lambda).$$

A pair (x^*, λ^*) is said to be a global saddle point of $L(x, \lambda)$ if

$$(1.2) \quad L(x^*, \lambda) \leq L(x^*, \lambda^*) \leq L(x, \lambda^*)$$

for all $x \in X$ and $\lambda \geq 0$. If there exists a $\delta > 0$ such that (1.2) holds for all $x \in X \cap N(x^*, \delta)$ and $\lambda \geq 0$, where $N(x^*, \delta) = \{x \in \mathbb{R}^n \mid \|x - x^*\| \leq \delta\}$, then (x^*, λ^*) is said to be a local saddle point of $L(x, \lambda)$. Here $\|\cdot\|$ denotes the 2-norm of vectors.

A key issue in the Lagrangian dual methods for solving (P) is whether or not a nonzero duality gap exists between (P) and (D) (see [13]). It is well known that a zero duality gap is equivalent to the existence of a saddle point of the Lagrangian function (see [14]). The existence of a saddle point thus plays a critical role in a successful adoption of Lagrangian dual methods for solving (P). The existence of a saddle point, however, cannot be ensured when the objective function or constraint functions, or both in problem (P), are nonconvex. A remedy to duality gaps in classical Lagrangian methods is to introduce nonlinear Lagrangian functions or augmented Lagrangian functions.

The first augmented Lagrangian method was independently proposed by Hestenes [3] and Powell [18] for equality-constrained problems by incorporating a quadratic penalty term in the classical Lagrangian function. The approach for dealing with inequality constraints was introduced by Rockafellar [19]. A more general class of augmented Lagrangians for (P) is proposed by Bertsekas [2]. Let $\phi : \mathbb{R} \rightarrow \mathbb{R}$ be a twice continuously differentiable and strictly convex function satisfying $\phi(0) = 0$ and $\phi'(0) = 0$. Define

$$(1.3) \quad P(s, t) = \begin{cases} ts + \phi(s) & \text{if } t + \phi'(s) \geq 0, \\ \min_{\tau \in \mathbb{R}} [t\tau + \phi(\tau)] & \text{otherwise.} \end{cases}$$

The *essentially quadratic* augmented Lagrangian function for (P) is defined as

$$(1.4) \quad L_1(x, \lambda, c) = f(x) + \frac{1}{c} \sum_{i=1}^m P(cg_i(x), \lambda_i),$$

where $c > 0$, $x \in X$, and $\lambda = (\lambda_1, \dots, \lambda_m)^T \geq 0$. It is easy to see that the augmented Lagrangian of Rockafellar [19] is a special case of $L_1(x, \lambda, c)$ when setting $\phi(s) = (1/2)s^2$ in (1.3).

The essentially quadratic augmented Lagrangian function L_1 is continuously differentiable with respect to x but not twice differentiable, thus preventing the use of Newton-type methods from solving the corresponding unconstrained Lagrangian relaxation problems. A prominent example of twice differentiable nonlinear Lagrangian functions for (P) is the exponential penalty function proposed by Kort and Bertsekas [8]. The exponential-type augmented Lagrangian can be derived by applying classical Lagrangian formulation to a reformulation of (P). Let $\psi : \mathbb{R} \rightarrow \mathbb{R}$ be a twice differentiable and strictly increasing function satisfying $\psi(0) = 0$. Rewrite (P) as

$$(1.5) \quad \begin{aligned} &\min f(x) \\ &\text{subject to } (1/c)\psi(cg_i(x)) \leq 0, \quad i = 1, \dots, m, \\ &\quad x \in X, \end{aligned}$$

where $c > 0$ is a parameter. Since ψ is a one-to-one mapping between $\mathbb{R}$ and $\mathbb{R}$, problem (1.5) is equivalent to (P). The Lagrangian function of (1.5) takes the following form:

$$(1.6) \quad L_2(x, \lambda, c) = f(x) + \frac{1}{c} \sum_{i=1}^m \lambda_i \psi(cg_i(x)), \quad x \in X, \lambda \geq 0.$$

It is clear that $L_2(x, \lambda, c)$ is twice differentiable with respect to x . The function L_2 includes the exponential penalty function as a special case by setting $\psi(t) = e^t - 1$ (see [2, 28]).

To avoid ill-conditional numerical behavior of the exponential-type augmented Lagrangians, modified barrier functions were proposed by Polyak [16]. Good numerical behavior of modified barrier functions was shown in several papers (e.g., [1, 17]). We now consider a general class of modified barrier functions. Let $\varphi: (-\infty, 1) \rightarrow \mathbb{R}$ be a twice differentiable and strictly increasing function satisfying $\varphi(0) = 0$. Similar to (1.5), we can reformulate (P) by using φ . Let $\Omega_c = \{x \in X \mid cg_i(x) < 1, i = 1, \dots, m\}$. Define the modified barrier augmented Lagrangian as follows:

$$(1.7) \quad L_3(x, \lambda, c) = \begin{cases} f(x) + \frac{1}{c} \sum_{i=1}^m \lambda_i \varphi(cg_i(x)), & x \in \Omega_c, \\ \infty, & x \in X \setminus \Omega_c, \end{cases}$$

where $c > 0$ and $\lambda \geq 0$. Taking $\varphi(t) = -\ln(1-t)$ or $\varphi(t) = 1 - 1/(1-t)$ in (1.7) gives rise to the modified Frish function or the modified Carroll function introduced in [16], respectively.

Another useful class of twice differentiable augmented Lagrangian functions for (P) can be derived by introducing a penalty term in the objective function of (1.5). Let $\xi: \mathbb{R} \rightarrow \mathbb{R}$ be a twice differentiable function satisfying $\xi(t) = 0$ for all $t \leq 0$ and $\xi(t) > 0$ for all $t > 0$. Consider the following problem:

$$(1.8) \quad \begin{aligned} \min \quad & f(x) + (1/c) \sum_{i=1}^m \xi(cg_i(x)) \\ \text{subject to} \quad & (1/c)\psi(cg_i(x)) \leq 0, \quad i = 1, \dots, m, \\ & x \in X. \end{aligned}$$

Again, problem (1.8) is equivalent to (P) as the two problems have the same feasible region and the objective function of (1.8) takes the same value as $f(x)$ in the feasible region. The Lagrangian function of (1.8) is

$$(1.9) \quad L_4(x, \lambda, c) = f(x) + \frac{1}{c} \sum_{i=1}^m [\lambda_i \psi(cg_i(x)) + \xi(cg_i(x))], \quad x \in X, \lambda \geq 0.$$

As we shall see later, one of the advantages of introducing the penalty function ξ in L_4 is that the sufficient conditions for the global saddle point are more natural and weaker than the ones for L_2 . This advantage becomes clear when the difficulties arise in the convergence analysis of the exponential Lagrangian method in convex programming (see [2, 15, 28]).

Same as the case of the classical Lagrangian, the saddle point condition of augmented Lagrangian L_i ($i = 1, 2, 3, 4$) is a sufficient condition for the optimality of (P).

THEOREM 1.1. *For $i = 1, 2, 3, 4$, if (x^*, λ^*) is a global (local) saddle point of $L_i(x, \lambda, c)$ for some $c > 0$, then x^* is a global (local) optimal solution to (P).*

Proof. Since L_i ($i = 2, 3, 4$) can be derived by applying the classical Lagrangian formulation to the reformulations of (P) (see (1.5) and (1.8)), the existence of saddle points of L_i ($i = 2, 3, 4$) as a sufficient condition for optimality of (P) follows from the classical saddle point theory. Let (x^*, λ^*) be a global (local) saddle point of $L_1(x, \lambda, c)$. By (1.2), the saddle point inequalities become

$$(1.10) \quad L_1(x^*, \lambda, c) \leq L_1(x^*, \lambda^*, c) \leq L_1(x, \lambda^*, c).$$

We claim that x^* is a feasible solution to (P). Suppose on the contrary $g_i(x^*) > 0$ for some i . Choose sufficiently large $\lambda_i > 0$ such that $\lambda_i + \phi'(cg_i(x^*)) \geq 0$. Then, by (1.3), we have

$$P(cg_i(x^*), \lambda_i) = c\lambda_i g_i(x^*) + \phi(cg_i(x^*)) \rightarrow \infty, \quad \lambda_i \rightarrow \infty,$$

which contradicts the first inequality of (1.10). Thus $g_i(x^*) \leq 0$ for all $i = 1, \dots, m$.

For any feasible x and $\lambda_i \geq 0$, since $g_i(x) \leq 0$, we have

$$(1.11) \quad P(cg_i(x), \lambda_i) = \min_{\tau \geq cg_i(x)} [\lambda_i \tau + \phi(\tau)] \leq \lambda_i \cdot 0 + \phi(0) = 0.$$

Also,

$$(1.12) \quad P(cg_i(x^*), 0) = \min_{\tau \geq cg_i(x^*)} \phi(\tau) = \phi(0) = 0.$$

On the other hand, setting $\lambda = 0$ in the first inequality of (1.10) and using (1.12) yield

$$f(x^*) = L_1(x^*, 0, c) \leq L_1(x^*, \lambda^*, c) = f(x^*) + \frac{1}{c} \sum_{i=1}^m P(cg_i(x^*), \lambda_i^*).$$

Thus,

$$(1.13) \quad \sum_{i=1}^m P(cg_i(x^*), \lambda_i^*) \geq 0.$$

Combining (1.11) with (1.13) yields $P(cg_i(x^*), \lambda_i^*) = 0$ for $i = 1, \dots, m$ and hence $L_1(x^*, \lambda^*, c) = f(x^*)$.

Let $x \in X$ (or $X \cap N(x^*, \delta)$) be a feasible solution of (P). By (1.11), $P(cg_i(x), \lambda_i^*) \leq 0$ for each i . The second inequality of (1.10) then gives rise to

$$f(x^*) = L_1(x^*, \lambda^*, c) \leq L_1(x, \lambda^*, c) = f(x) + \frac{1}{c} \sum_{i=1}^m P(cg_i(x), \lambda_i^*) \leq f(x).$$

Therefore, x^* is a global (local) optimum of (P). $\square$

A crucial issue in applying dual search methods in constrained optimization is the existence of a saddle point of the adopted Lagrangian associated with problem (P). The existence of a saddle point of the conventional Lagrangian was proven by Karlin [7] for convex situations. Various types of nonlinear Lagrangian formulations [9, 10, 11, 12, 6, 23, 24] have been developed in recent years to eliminate the duality gap in nonconvex situations, thus ensuring a success of the dual search in finding an optimal solution of the primal problem. The existence of a global saddle point of p th power nonlinear Lagrangians was investigated in [9, 10, 11, 12] for nonconvex

constrained optimization problems. Despite the popularity of the four types of the augmented Lagrangian functions mentioned previously, it appears that no systematic investigation has been done for the existence of saddle points of these augmented Lagrangian functions. The purpose of this paper is twofold. First, we establish new local duality properties of the augmented Lagrangian functions L_i ($i = 1, 2, 3, 4$). We prove that under the second-order sufficiency conditions, local saddle points of L_i ($i = 1, 2, 3, 4$) exist. Our results for L_1 and L_3 can be regarded as generalizations of the existence of a local saddle point for Rockafellar's augmented Lagrangian function (see [20, 21]) and the modified barrier functions (see [16]). Second, we show that global saddle points of L_1 , L_3 , and L_4 exist under the conditions that the global solution to (P) is unique and X is compact. The existence of a global saddle point of L_2 can also be ensured under an additional separation condition. The need to study the global saddle point in constrained global optimization is obvious. On one hand, unconstrained global optimization has been extensively studied and various deterministic and stochastic methods have been proposed (see [4, 5]). On the other hand, implementable methods for constrained global optimization has been only developed for some special problems such as concave minimization (see [22]) and monotone optimization (see [26, 29]). It is desirable to reduce the constrained global optimization problem into a sequence of unconstrained global optimization problems so that the methods developed for unconstrained global optimization can be used. This methodology was adopted in constructing auxiliary functions for constrained global optimization (see [25, 27]). The global duality results established in this paper provide a theoretical foundation for the use of augmented Lagrangian methods in constrained global optimization.

This paper is organized as follows. In section 2, we establish local duality properties of augmented Lagrangian functions L_i ($i = 1, 2, 3, 4$). In particular, we derive sufficient conditions for the existence of a local saddle point of L_i ($i = 1, 2, 3, 4$). Global saddle points of L_i ($i = 1, 2, 3, 4$) are investigated in section 3. Finally, we give some concluding remarks in section 4.

2. Existence of local saddle point. Let $x^* \in X$ be a local solution to (P). Since the abstract set X usually represents $\mathbb{R}^n$ or simple constraints of lower and upper bounds of the variables, we can assume, without loss of generality, that $x^* \in \text{int}(X)$. This does not impose any restriction on x^* as we can always enlarge X if necessary. Let $I(x^*) = \{i \mid g_i(x^*) = 0\}$. We need the following assumption about x^* .

Assumption 2.1 (second-order sufficiency conditions). Let $x^* \in \text{int}(X)$ be a local solution to (P). Assume that x^* is a regular point of (P) and satisfies the second-order sufficiency conditions; that is, $\nabla g_i(x^*)$, $i \in I(x^*)$, are linearly independent and there exists $\lambda^* \geq 0$ such that

$$(2.1) \quad \nabla f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla g_i(x^*) = 0,$$

$$(2.2) \quad \lambda_i^* g_i(x^*) = 0, \quad i = 1, \dots, m,$$

and the Hessian matrix

$$\nabla_{xx}^2 L(x^*, \lambda^*) = \nabla^2 f(x^*) + \sum_{i \in J(x^*)} \lambda_i^* \nabla^2 g_i(x^*)$$

is positive definite on the cone

$$M(x^*) = \{d \in \mathbb{R}^n, d \neq 0 \mid d^T \nabla g_i(x^*) = 0 \text{ for any } i \in J(x^*)\},$$

where $J(x^*) = \{i \mid \lambda_i^* > 0, i = 1, \dots, m\}$.

2.1. Essentially augmented Lagrangian function. Consider the essentially augmented Lagrangian function L_1 defined in (1.4). Suppose that the function ϕ in (1.3) satisfies the following conditions:

- (A1) ϕ is twice continuously differentiable and strictly convex on $\mathbb{R}$;
- (A2) $\phi(0) = 0, \phi'(0) = 0$.

We first observe that the conditions (A1) and (A2) imply that $\phi(s) \geq 0$ and $\phi''(s) > 0$ for all $s \in \mathbb{R}$ and that $s = 0$ is the unique minimizer of $\phi(s)$ on $\mathbb{R}$. Furthermore, ϕ is strictly decreasing on $(-\infty, 0]$ and strictly increasing on $[0, \infty)$. It is clear that $P(\cdot, t)$ defined in (1.3) is a convex function for any fixed $t \geq 0$. Since for any $t \geq 0$, $t\tau + \phi(\tau)$ is a strictly convex function of τ , $P(s, t)$ can be rewritten as

$$(2.3) \quad P(s, t) = \min_{\tau \geq s} [t\tau + \phi(\tau)].$$

THEOREM 2.1. *Let x^* be a local optimal solution to (P). Suppose that Assumption 2.1 is satisfied at x^* . Then there exist $c_0 > 0$ and $\delta > 0$ such that, for all $c \geq c_0$,*

$$(2.4) \quad L_1(x^*, \lambda, c) \leq L_1(x^*, \lambda^*, c) \leq L_1(x, \lambda^*, c)$$

for any $x \in N(x^*, \delta)$ and $\lambda \geq 0$, where L_1 is defined in (1.4) with conditions (A1) and (A2) for ϕ being satisfied.

Proof. First, we show that $L_1(x^*, \lambda^*, c) = f(x^*)$ which implies the first inequality of (2.4) by noting that $P(cg_i(x^*), \lambda_i) \leq 0$ for $\lambda_i \geq 0$ (cf. (1.11)). For $i \in J(x^*)$, we have $\lambda_i^* > 0$ and $g_i(x^*) = 0$ by (2.2). Thus $\lambda_i^* + \phi'(cg_i(x^*)) = \lambda_i^* > 0$. By (2.3), we get

$$P(cg_i(x^*), \lambda_i^*) = \min_{\tau \geq cg_i(x^*)} [\lambda_i^* \tau + \phi(\tau)] = c\lambda_i^* g_i(x^*) + \phi(cg_i(x^*)) = 0.$$

If $i \in \{1, \dots, m\} \setminus J(x^*)$, then $\lambda_i^* = 0$. By (1.12), we have $P(cg_i(x^*), \lambda_i^*) = 0$. Therefore $L_1(x^*, \lambda^*, c) = f(x^*)$.

Next, we observe that for any fixed $s \in \mathbb{R}$ and $t \geq 0$, $(1/c)P(cs, t)$ is a nondecreasing function of $c > 0$ and hence $L_1(x, \lambda, c)$ is a nondecreasing function of $c > 0$. Indeed, for any $0 < c_1 \leq c_2$, by the convexity of $P(\cdot, t)$, we have

$$\begin{aligned} P(c_1 s, t) &= P\left(\frac{c_1}{c_2} c_2 s + \left(1 - \frac{c_1}{c_2}\right) \cdot 0, t\right) \\ &\leq \frac{c_1}{c_2} P(c_2 s, t) + \left(1 - \frac{c_1}{c_2}\right) P(0, t) \\ &= \frac{c_1}{c_2} P(c_2 s, t). \end{aligned}$$

Hence, $(1/c_1)P(c_1 s, t) \leq (1/c_2)P(c_2 s, t)$.

Now we prove the second inequality of (2.4). If $J(x^*) = \emptyset$, then $\lambda_i^* = 0$ for $i = 1, \dots, m$ and Assumption 2.1 implies that x^* is a (strict) local minimizer of $f(x)$. The second inequality of (2.4) becomes

$$f(x^*) \leq f(x) + (1/c) \sum_{i=1}^m P(cg_i(x), 0), \quad x \in N(x^*, \delta).$$

This inequality then holds trivially by noting that $P(s, 0) = \min_{\tau \geq s} \phi(\tau) \geq 0$ for any s . We assume in what follows that $J(x^*) \neq \emptyset$. Since $L_1(x^*, \lambda^*, c) = f(x^*)$ and $L_1(x, \lambda^*, c)$ is a nondecreasing function of $c > 0$, it suffices to prove that there exist $c_0 > 0$ and $\delta > 0$ such that the second inequality of (2.4) holds at $c = c_0$ for all $x \in X \cap N(x^*, \delta)$. Suppose on the contrary such c_0 and δ do not exist. Then, for each positive integer k , there must exist a sequence $\{y_k^j\} \subset X$ with $y_k^j \rightarrow x^*$ ($j \rightarrow \infty$) such that

$$(2.5) \quad L_1(x^*, \lambda^*, k) > L_1(y_k^j, \lambda^*, k) \quad \text{for any } j.$$

Notice that for each $i \in J(x^*)$, we have $\lambda_i^* > 0$ and $g_i(x^*) = 0$. Thus, for each k , by the continuity of g_i , there exists $\delta_k > 0$ such that $|g_i(x)| < 1/k^2$ for any $i \in J(x^*)$ when $\|x - x^*\| \leq \delta_k$. Since $y_k^j \rightarrow x^*$ ($j \rightarrow \infty$), we can always choose a j such that $x_k = y_k^j$ satisfies $\|x_k - x^*\| \leq \delta_k$, $x_k \rightarrow x^*$ ($k \rightarrow \infty$) and $x_k \neq x^*$. Thus

$$(2.6) \quad |g_i(x)| < \frac{1}{k^2}$$

for any x satisfying $\|x - x^*\| \leq \|x_k - x^*\|$ and $i \in J(x^*)$. Notice that (2.6) implies $kg_i(x_k) \rightarrow 0$ ($k \rightarrow \infty$). The continuous differentiability of ϕ and $\phi'(0) = 0$ then imply that there exists k_0 such that

$$(2.7) \quad \phi'(kg_i(x_k)) > -\lambda_i^*, \quad k \geq k_0, \quad \text{for any } i \in J(x^*).$$

From (2.5) and the fact that $L_1(x^*, \lambda^*, k) = f(x^*)$, we have

$$f(x^*) > L_1(x_k, \lambda^*, k) \quad \text{for any } k,$$

which is equivalent to

$$(2.8) \quad 0 > f(x_k) - f(x^*) + \frac{1}{k} \sum_{i=1}^m P(kg_i(x_k), \lambda_i^*) \quad \text{for any } k.$$

For $i \in J(x^*)$, it follows from (1.3) and (2.7) that

$$(2.9) \quad \frac{1}{k} P(kg_i(x_k), \lambda_i^*) = \lambda_i^* g_i(x_k) + \frac{1}{k} \phi(kg_i(x_k)), \quad k \geq k_0, \quad \text{for any } i \in J(x^*).$$

For $i \notin J(x^*)$, it holds $\lambda_i^* = 0$ by definition. Since $\phi(s) \geq 0$ for any s , we have

$$(2.10) \quad P(kg_i(x_k), \lambda_i^*) \geq 0 \quad \text{for any } i \notin J(x^*).$$

Combining (2.8) with (2.9) and (2.10), we obtain for $k \geq k_0$,

$$\begin{aligned} (2.11) \quad 0 &> f(x_k) - f(x^*) + \frac{1}{k} \sum_{i=1}^m P(kg_i(x_k), \lambda_i^*) \\ &\geq f(x_k) - f(x^*) + \frac{1}{k} \sum_{i \in J(x^*)} P(kg_i(x_k), \lambda_i^*) \\ &= f(x_k) - f(x^*) + \sum_{i \in J(x^*)} \left[\lambda_i^* g_i(x_k) + \frac{1}{k} \phi(kg_i(x_k)) \right] \\ &= L(x_k, \lambda^*) - L(x^*, \lambda^*) + \frac{1}{k} \sum_{i \in J(x^*)} \phi(kg_i(x_k)). \end{aligned}$$

Since $\nabla_x L(x^*, \lambda^*) = 0$ by (2.1), we have

$$L(x_k, \lambda^*) - L(x^*, \lambda^*) = \frac{1}{2}(x_k - x^*)^T \nabla_{xx}^2 L(x^*, \lambda^*)(x_k - x^*) + o(\|x_k - x^*\|^2). \quad (2.12)$$

On the other hand, since ϕ and g_i are twice continuously differentiable, we have

$$\begin{aligned} & \frac{1}{k} \sum_{i \in J(x^*)} \phi(kg_i(x_k)) \\ &= \frac{1}{k} \sum_{i \in J(x^*)} \phi(kg_i(x^*)) + \sum_{i \in J(x^*)} \phi'(kg_i(x^*)) \nabla^T g_i(x^*)(x_k - x^*) \\ &+ \frac{1}{2} \sum_{i \in J(x^*)} (x_k - x^*)^T [\phi'(kg_i(\xi_k)) \nabla^2 g_i(\xi_k) + k\phi''(kg_i(\xi_k)) \nabla g_i(\xi_k) \nabla^T g_i(\xi_k)](x_k - x^*), \end{aligned} \quad (2.13)$$

where $\xi_k = x^* + \theta_k(x_k - x^*)$, $\theta_k \in [0, 1]$. Let $d_k = (x_k - x^*)/\|x_k - x^*\|$. Notice that $\phi(kg_i(x^*)) = \phi(0) = 0$ and $\phi'(kg_i(x^*)) = \phi'(0) = 0$ for $i \in J(x^*)$. It then follows from (2.11), (2.12), and (2.13) that

$$\begin{aligned} 0 &> \frac{1}{2} d_k^T \nabla_{xx}^2 L(x^*, \lambda^*) d_k \\ &+ \frac{1}{2} \sum_{i \in J(x^*)} \{ \phi'(kg_i(\xi_k)) d_k^T \nabla^2 g_i(\xi_k) d_k + k\phi''(kg_i(\xi_k)) [d_k^T \nabla g_i(\xi_k)]^2 \} + \frac{o(\|x_k - x^*\|^2)}{\|x_k - x^*\|^2}. \end{aligned} \quad (2.14)$$

Since $\|d_k\| = 1$, we can assume without loss of generality that $d_k \rightarrow d$ ($k \rightarrow \infty$) with $\|d\| = 1$. Notice also that $\|\xi_k - x^*\| \leq \|x_k - x^*\|$. Thus, $\xi_k \rightarrow x^*$ ($k \rightarrow \infty$). Furthermore, by (2.6), we have $kg_i(\xi_k) \rightarrow 0$ ($k \rightarrow \infty$) for $i \in J(x^*)$. This in turn implies

$$\begin{aligned} \phi'(kg_i(\xi_k)) &\rightarrow \phi'(0) = 0, \quad k \rightarrow \infty, \quad i \in J(x^*), \\ \phi''(kg_i(\xi_k)) &\rightarrow \phi''(0) > 0, \quad k \rightarrow \infty, \quad i \in J(x^*). \end{aligned}$$

Taking limits in (2.14), we must have

$$(2.15) \quad d^T \nabla g_i(x^*) = 0 \quad \text{for any } i \in J(x^*),$$

$$(2.16) \quad 0 \geq \frac{1}{2} d^T \nabla_{xx}^2 L(x^*, \lambda^*) d, \quad \|d\| = 1.$$

Equation (2.15) implies that $d \in M(x^*)$ and thus, by Assumption 2.1,

$$d^T \nabla_{xx}^2 L(x^*, \lambda^*) d > 0,$$

which contradicts (2.16). $\square$

2.2. Exponential-type and modified barrier augmented Lagrangians.

Let the function ψ in (1.6) satisfy the following conditions:

- (B1) ψ is a twice differentiable and strictly convex function on $\mathbb{R}$;
- (B2) $\psi(0) = 0$, $\psi'(0) = 1$, $\lim_{t \rightarrow -\infty} \psi'(t) = 0$.

Similarly, let φ in (1.7) satisfy the following:

(B1)' φ is a twice differentiable and strictly convex function on $(-\infty, 1)$;

(B2)' $\varphi(0) = 0$, $\varphi'(0) = 1$, $\lim_{t \rightarrow -\infty} \varphi'(t) = 0$.

We note that the exponential function $\psi(t) = e^t - 1$ satisfies (B1)–(B2) and functions $\varphi(t) = -\ln(1-t)$ and $\varphi(t) = 1 - 1/(1-t)$ satisfy (B1)'–(B2)' and we need the following technical lemma [2, Lemma 1.25].

LEMMA 2.2. *Let A be a symmetric $n \times n$ matrix and B be a positive semidefinite symmetric $n \times n$ matrix. Assume that $y^T A y > 0$ for any $y \neq 0$ satisfying $By = 0$. Then there exists $\tau_0 > 0$ such that $A + \tau B$ is positive definite when $\tau \geq \tau_0$.*

THEOREM 2.3. *Let x^* be a local optimal solution to (P). Suppose that Assumption 2.1 is satisfied at x^* . Then there exist $c_0 > 0$ and $\delta > 0$ such that for all $c \geq c_0$,*

$$(2.17) \quad L_i(x^*, \lambda, c) \leq L_i(x^*, \lambda^*, c) \leq L_i(x, \lambda^*, c), \quad i = 2, 3,$$

for any $x \in N(x^*, \delta)$ and $\lambda \geq 0$, where L_2 is defined in (1.6) with conditions (B1)–(B2) for ψ being satisfied, and L_3 is defined in (1.7) with conditions (B1)'–(B2)' for φ being satisfied.

Proof. By the definitions of L_i ($i = 2, 3$) and conditions (B1)–(B2) and (B1)'–(B2)', we only need to prove the theorem for L_2 as similar arguments apply to L_3 . From condition (B1), we have $\psi''(t) > 0$ for any t , which together with $\lim_{t \rightarrow -\infty} \psi'(t) = 0$ implies that $\psi'(t) > 0$ for any t . Thus ψ is a strictly increasing function on $\mathbb{R}$. The property $\psi(0) = 0$ and (2.2) then imply $\psi(cg_i(x^*)) \leq 0$ and $\lambda_i^* \psi(cg_i(x^*)) = 0$ for all i . Hence, $L_2(x^*, \lambda^*, c) = f(x^*)$ for any $c > 0$ and the first inequality of (2.17) holds. To prove the second inequality of (2.17), we first observe that $(1/c)\psi(ct)$ is a nondecreasing function of $c > 0$ for any t by the convexity of ψ , similar to the proof for $(1/c)P(cs, t)$ in Theorem 2.1. It suffices to show that the second inequality of (2.17) holds for some c_0 . It follows from (1.6), (2.1), and the condition (B2) that

$$\begin{aligned} \nabla_x L_2(x^*, \lambda^*, c) &= \nabla f(x^*) + \sum_{i=1}^m \lambda_i^* \psi'(cg_i(x^*)) \nabla g_i(x^*) \\ &= \nabla f(x^*) + \sum_{i \in J(x^*)} \lambda_i^* \psi'(0) \nabla g_i(x^*) \\ &= \nabla_x L(x^*, \lambda^*) \\ &= 0. \end{aligned} \quad (2.18)$$

Moreover, we have

$$\begin{aligned} \nabla_{xx}^2 L_2(x^*, \lambda^*, c) &= \nabla^2 f(x^*) + \sum_{i=1}^m \lambda_i^* [\psi'(cg_i(x^*)) \nabla^2 g_i(x^*) \\ &\quad + c\psi''(cg_i(x^*)) \nabla g_i(x^*) \nabla^T g_i(x^*)] \\ &= \nabla^2 f(x^*) + \sum_{i \in J(x^*)} \lambda_i^* [\psi'(0) \nabla^2 g_i(x^*) + c\psi''(0) \nabla g_i(x^*) \nabla^T g_i(x^*)] \\ &= \nabla_{xx}^2 L(x^*, \lambda^*) + c\psi''(0) \sum_{i \in J(x^*)} \lambda_i^* \nabla g_i(x^*) \nabla^T g_i(x^*). \end{aligned} \quad (2.19)$$

Notice here that $\psi''(0) > 0$. Thus, by Assumption 2.1 and Lemma 2.2, we infer from (2.19) that there exists $c_0 > 0$ such that $\nabla_{xx}^2 L_2(x^*, \lambda^*, c_0)$ is positive definite. Combined with (2.18), this implies that x^* is a (strict) local minimizer of $L_2(x, \lambda^*, c_0)$ and the second inequality of (2.17) holds. $\square$

2.3. Penalized exponential-type augmented Lagrangian. Suppose that the penalty function ξ in (1.9) satisfies the following conditions:

- (C1) ξ is a twice differentiable and convex function on $\mathbb{R}$ with $\xi''(0) = 0$;
 (C2) $\xi(t) = 0$ for $t \leq 0$, $\xi(t) > 0$ for $t > 0$.

By the convexity of ξ and (C2), we have $\xi(0) \geq \xi(t) + \xi'(t)(0-t)$, so $\xi'(t) \geq \xi(t)/t > 0$ for any $t > 0$. Therefore, $\xi(t)$ is a strictly increasing function for $t > 0$. A typical form of function ξ satisfying (C1)–(C2) is $\xi(t) = (1/\rho)[\max(0, t)]^\rho$, $\rho \geq 3$.

THEOREM 2.4. *Let x^* be a local optimal solution to (P). Suppose that Assumption 2.1 is satisfied at x^* . Then there exist $c_0 > 0$ and $\delta > 0$ such that, for all $c \geq c_0$,*

$$(2.20) \quad L_4(x^*, \lambda, c) \leq L_4(x^*, \lambda^*, c) \leq L_4(x, \lambda^*, c)$$

for any $x \in N(x^*, \delta)$ and $\lambda \geq 0$, where L_4 is defined in (1.9) with conditions (B1)–(B2) for ψ and conditions (C1) and (C2) for ξ being satisfied.

Proof. We first note from conditions (C1) and (C2) that $\xi'(t) = 0$ for $t \leq 0$. By (1.9), (2.1), and condition (B2), we have

$$\begin{aligned} \nabla_x L_4(x^*, \lambda^*, c) &= \nabla f(x^*) + \sum_{i=1}^m [\lambda_i^* \psi'(cg_i(x^*)) + \xi'(cg_i(x^*))] \nabla g_i(x^*) \\ &= \nabla f(x^*) + \sum_{i \in J(x^*)} [\lambda_i^* \psi'(0) + \xi'(0)] \nabla g_i(x^*) \\ &= \nabla_x L(x^*, \lambda^*) \\ (2.21) \quad &= 0. \end{aligned}$$

We also have

$$\begin{aligned} \nabla_{xx}^2 L_4(x^*, \lambda^*, c) &= \nabla^2 f(x^*) + \sum_{i=1}^m [\lambda_i^* \psi'(cg_i(x^*)) + \xi'(cg_i(x^*))] \nabla^2 g_i(x^*) \\ &\quad + c \sum_{i=1}^m [\lambda_i^* \psi''(cg_i(x^*)) + \xi''(cg_i(x^*))] \nabla g_i(x^*) \nabla^T g_i(x^*) \\ &= \nabla^2 f(x^*) + \sum_{i \in J(x^*)} [\lambda_i^* \psi'(0) + \xi'(0)] \nabla^2 g_i(x^*) \\ &\quad + c \sum_{i \in J(x^*)} [\lambda_i^* \psi''(0) + \xi''(0)] \nabla g_i(x^*) \nabla^T g_i(x^*) \\ (2.22) \quad &= \nabla_{xx}^2 L(x^*, \lambda^*) + c \psi''(0) \sum_{i \in J(x^*)} \lambda_i^* \nabla g_i(x^*) \nabla^T g_i(x^*). \end{aligned}$$

The last equality in (2.22) is due to the fact that $\psi'(0) = 1$, $\xi'(0) = \xi''(0) = 0$. The reminder of the proof is similar to that of Theorem 2.3. $\square$

Theorems 2.1, 2.3, and 2.4 show that the second-order sufficiency conditions can ensure the existence of local saddle points of L_i ($i = 1, 2, 3, 4$). From a duality point of view, this leads to a zero duality gap in a local sense when the minimization in the Lagrangian relaxation problem (1.1) is considered in a neighborhood of x^* . It is worth pointing out that Theorem 2.1 can also be viewed as a generalization of the existence theorem for Rockafellar's augmented Lagrangian (see [19, 20]).

Together with Theorem 1.1, Theorems 2.1, 2.3, and 2.4 imply that the local saddle point condition for L_i ($i = 1, 2, 3, 4$) is both a sufficient and necessary optimality condition for a local optimum to (P) under Assumption 2.1.

3. Existence of global saddle point. In this section we will establish results for the existence of global saddle points for L_i ($i = 1, 2, 3, 4$). Since the existence of a global saddle point in constrained nonconvex optimization corresponds to a strong duality relation between the primal and dual problems, the results obtained here will provide theoretical guarantee for the success of augmented Lagrangian methods in constrained global optimization (see [4, 25]).

First, we discuss the essentially augmented Lagrangian L_1 . In addition to (A1) and (A2), let ϕ in (1.4) satisfy the following coercivity condition:

$$(A3) \quad \phi(s)/s \rightarrow \infty, \quad (|s| \rightarrow \infty).$$

By the convexity of ϕ , condition (A3) implies that $\phi'(s) \rightarrow \infty$ ($s \rightarrow \infty$). Obviously, if ϕ is strongly convex, then (A3) is satisfied. The Rockafellar's augmented Lagrangian function $\phi(s) = (1/2)s^2$ satisfies (A3). Examples of other forms of ϕ satisfying condition (A3) in the class of essentially quadratic augmented Lagrangian functions include

$$\begin{aligned} \phi(s) &= \rho|s|^\rho, \quad \rho > 2; \\ \phi(s) &= \rho|s|^\rho + \frac{1}{2}s^2, \quad \rho > 2. \end{aligned}$$

THEOREM 3.1. *Let x^* be a unique global optimal solution of (P). Suppose that X in (P) is compact and Assumption 2.1 is satisfied at x^* . Then there exists $c_1 \geq c_0$ such that for all $c \geq c_1$, (x^*, λ^*) is a global saddle point of $L_1(x, \lambda, c)$:*

$$(3.1) \quad L_1(x^*, \lambda, c) \leq L_1(x^*, \lambda^*, c) \leq L_1(x, \lambda^*, c)$$

for any $x \in X$ and $\lambda \geq 0$, where L_1 is defined in (1.4) with conditions (A1)–(A3) for ϕ being satisfied.

Proof. By Theorem 2.1, for all $c \geq c_0$, it holds that

$$(3.2) \quad L_1(x^*, \lambda, c) \leq L_1(x^*, \lambda^*, c) \leq L_1(x, \lambda^*, c)$$

for all $x \in N(x^*, \delta)$ and $\lambda \geq 0$. Let c_0 be defined in Theorem 2.1. In what follows we shall show that there exists a $c_1 \geq c_0$ such that the second inequality in (3.2) holds for all $x \in X$ when $c \geq c_1$. For $\epsilon \geq 0$, define

$$(3.3) \quad U_\epsilon = \{x \in X \mid f(x) \leq f(x^*) + \epsilon\},$$

$$(3.4) \quad F_\epsilon = \{x \in X \mid g_i(x) \leq \epsilon, \quad i = 1, \dots, m\}.$$

Then the uniqueness of x^* implies $U_0 \cap F_0 = \{x^*\}$ and therefore

$$(3.5) \quad [U_0 \setminus N(x^*, \delta)] \cap [F_0 \setminus N(x^*, \delta)] = \emptyset.$$

By the compactness of X and the continuity of f and g_i , we deduce from (3.5) that there exists an $\epsilon_1 > 0$ such that

$$(3.6) \quad [U_{\epsilon_1} \setminus N(x^*, \delta)] \cap [F_{\epsilon_1} \setminus N(x^*, \delta)] = \emptyset.$$

Now, for any $x \in X \setminus N(x^*, \delta)$, we consider the following two cases.

Case (a): $x \in U_{\epsilon_1} \setminus N(x^*, \delta)$. Then by (3.6), $x \notin F_{\epsilon_1}$. There exists an i_0 such that $g_{i_0}(x) > \epsilon_1$, and thus

$$\lambda_{i_0}^* + \phi'(cg_{i_0}(x)) \geq \phi'(cg_{i_0}(x)) > 0.$$

From (1.3) and the fact that ϕ is strictly increasing on $[0, \infty)$, we have

$$(3.7) \quad \begin{aligned} P(cg_{i_0}(x), \lambda_{i_0}^*) &= c\lambda_{i_0}^* g_{i_0}(x) + \phi(cg_{i_0}(x)) \\ &> c\lambda_{i_0}^* \epsilon_1 + \phi(c\epsilon_1). \end{aligned}$$

Let $w_i = (\phi')^{-1}(-\lambda_i^*)$, where $(\phi')^{-1}$ denotes the inverse of the strictly increasing function ϕ' . Then $\tau = w_i$ is the unique minimizer of function $\lambda_i^* \tau + \phi(\tau)$. We therefore have

$$(3.8) \quad \begin{aligned} P(cg_i(x), \lambda_i^*) &= \min_{\tau \geq cg_i(x)} [\lambda_i^* \tau + \phi(\tau)] \\ &\geq \min_{\tau \in \mathbb{R}} [\lambda_i^* \tau + \phi(\tau)] \\ &= \lambda_i^* w_i + \phi(w_i). \end{aligned}$$

Let $\underline{f} = \min_{x \in X} f(x)$. By the compactness of X , we have $\underline{f} > -\infty$. By using (3.7), (3.8), and the coercivity of ϕ , we obtain

$$\begin{aligned} L_1(x, \lambda^*, c) &= f(x) + \frac{1}{c} \sum_{i=1}^m P(cg_i(x), \lambda_i^*) \\ &= f(x) + \lambda_{i_0}^* g_{i_0}(x) + \frac{1}{c} \phi(cg_{i_0}(x)) + \frac{1}{c} \sum_{i \neq i_0} P(cg_i(x), \lambda_i^*) \\ &> \underline{f} + \lambda_{i_0}^* \epsilon_1 + \frac{1}{c} \phi(c\epsilon_1) + \frac{1}{c} \sum_{i \neq i_0} [\lambda_i^* w_i + \phi(w_i)] \\ &\rightarrow \infty, \quad (c \rightarrow \infty). \end{aligned}$$

Hence, there exists a $c_2 \geq c_0$ independent of x such that $L_1(x, \lambda^*, c) \geq L(x^*, \lambda^*)$ for any $x \in U_{\epsilon_1} \setminus N(x^*, \delta)$ when $c \geq c_2$.

Case (b): $x \in X \setminus [U_{\epsilon_1} \cup N(x^*, \delta)]$. Then $f(x) > f(x^*) + \epsilon_1$. By (3.8), we have

$$(3.9) \quad \begin{aligned} L_1(x, \lambda^*, c) &= f(x) + \frac{1}{c} \sum_{i=1}^m P(cg_i(x), \lambda_i^*) \\ &> f(x^*) + \epsilon_1 + \frac{1}{c} \sum_{i=1}^m [\lambda_i^* w_i + \phi(w_i)]. \end{aligned}$$

Let

$$c_3 \geq \max \left\{ c_0, -\frac{1}{\epsilon_1} \sum_{i=1}^m [\lambda_i^* w_i + \phi(w_i)] \right\}.$$

It then follows from (3.9) that $L_1(x, \lambda^*, c) > f(x^*) = L_1(x^*, \lambda^*, c)$ when $c \geq c_3$.

Let $c_1 = \max(c_2, c_3)$. The above discussion indicates that the second inequality of (3.2) holds for all $x \in X$ when $c \geq c_1$. $\square$

Next, we discuss the global saddle point property of L_2 . In addition to (B1) and (B2), we assume that ψ satisfies the following condition:

(B3) $\lim_{t \rightarrow -\infty} \psi(t) > -\infty$, $\lim_{t \rightarrow \infty} \psi(t)/t = \infty$.

We note that the exponential function $\psi(t) = e^t - 1$ satisfies (B3). Let

$$(3.10) \quad S_\epsilon = \{x \in X \mid g_i(x) \leq \epsilon, i \in J(x^*)\}.$$

In order to ensure the existence of a global saddle point, we need the following additional assumption about x^* :

$$(3.11) \quad U_0 \cap S_0 = \{x^*\}.$$

We observe that the separation condition (3.11) implies the uniqueness of a global optimal solution of (P) since $F_0 = \{x \in X \mid g_i(x) \leq 0, i = 1, \dots, m\} \subset S_0$ and thus by (3.11), $U_0 \cap F_0 = \{x^*\}$. It is easy to see that this assumption is automatically satisfied in the strictly convex case of (P).

THEOREM 3.2. *Let x^* be a unique global optimal solution of (P). Suppose that X is compact and Assumption 2.1 is satisfied at x^* . Moreover, assume that the separation condition (3.11) is satisfied. Then there exists $c_1 \geq c_0$ such that for all $c \geq c_1$, (x^*, λ^*) is a global saddle point of L_2 :*

$$(3.12) \quad L_2(x^*, \lambda, c) \leq L_2(x^*, \lambda^*, c) \leq L_2(x, \lambda^*, c)$$

for any $x \in X$ and $\lambda \geq 0$, where L_2 is defined in (1.6) with conditions (B1)–(B3) for ψ being satisfied.

Proof. By Theorem 2.3, there exist $\delta > 0$ and c_0 such that for all $c \geq c_0$, (3.12) holds for all $x \in N(x^*, \delta)$ and $\lambda \geq 0$. In what follows we shall show that there exists a $c_1 \geq c_0$ such that the second inequality in (3.12) holds for all $x \in X$ when $c \geq c_1$. Since S_0 and U_0 are compact, by (3.11),

$$(3.13) \quad [U_0 \setminus N(x^*, \delta)] \cap [S_0 \setminus N(x^*, \delta)] = \emptyset.$$

By the compactness of X and the continuity of f and g_i , we deduce from (3.13) that there exists an $\epsilon_1 > 0$ such that

$$(3.14) \quad [U_{\epsilon_1} \setminus N(x^*, \delta)] \cap [S_{\epsilon_1} \setminus N(x^*, \delta)] = \emptyset.$$

Now, for any $x \in X \setminus N(x^*, \delta)$, we consider the following two cases.

Case (a): $x \in U_{\epsilon_1} \setminus N(x^*, \delta)$. Then by (3.14), $x \notin S_{\epsilon_1}$. There exists an $i_0 \in J(x^*)$ such that $g_{i_0}(x) > \epsilon_1$. Since ψ is a strictly increasing function, we have $\psi(cg_{i_0}(x)) > \psi(c\epsilon_1)$. Condition (B3) implies that $\underline{\psi} = \min_{t \in \mathbb{R}} \psi(t) > -\infty$. Moreover, by the compactness of X , $\underline{f} = \min_{x \in X} f(x) > -\infty$. Therefore

$$(3.15) \quad \begin{aligned} L_2(x, \lambda^*, c) &= f(x) + (1/c) \sum_{i=1}^m \lambda_i^* \psi(cg_i(x)) \\ &> \underline{f} + (1/c) \lambda_{i_0}^* \psi(c\epsilon_1) + \underline{\psi} \sum_{i \neq i_0} \lambda_i^*. \end{aligned}$$

Note that $\lambda_{i_0}^* > 0$ and $(1/c)\psi(c\epsilon_1) \rightarrow \infty$ ($c \rightarrow \infty$) by condition (B3). We deduce from (3.15) that there exists $c_2 \geq c_0$ independent of x such that when $c \geq c_2$, $L_2(x, \lambda^*, c) > f(x^*) = L_2(x^*, \lambda^*, c)$ holds for all $x \in U_{\epsilon_1} \setminus N(x^*, \delta)$.

Case (b): $x \in X \setminus [U_{\epsilon_1} \cup N(x^*, \delta)]$. Then $f(x) > f(x^*) + \epsilon_1$. We have

$$(3.16) \quad \begin{aligned} L_2(x, \lambda^*, c) &= f(x) + \frac{1}{c} \sum_{i=1}^m \lambda_i^* \psi(cg_i(x)) \\ &> f(x^*) + \epsilon_1 + \frac{1}{c} \sum_{i=1}^m \underline{\psi} \lambda_i^*. \end{aligned}$$

Since the last term in (3.16) tends to 0 as c tends to infinity, we can find a c_3 independent of x such that for $c \geq c_3$, $L_2(x, \lambda^*, c) > f(x^*) = L_2(x^*, \lambda^*, c)$ for all $x \in X \setminus [U_{\epsilon_1} \cup N(x^*, \delta)]$.

Let $c_1 = \max(c_2, c_3)$. Then, the second inequality of (3.12) holds for all $x \in X$ when $c \geq c_1$. $\square$

We turn now to the modified barrier augmented Lagrangian L_3 . In addition to (B1)' and (B2)', the following growth condition is assumed for φ :

$$(B3)' \lim_{t \rightarrow -\infty} \varphi(t)/t = 0.$$

It is easy to verify that $\varphi(t) = -\ln(1-t)$ and $\varphi(t) = 1 - 1/(1-t)$ satisfy (B3)'.

THEOREM 3.3. *Let x^* be a unique global optimal solution of (P). Suppose that X is compact and Assumption 2.1 is satisfied at x^* . Then there exists $c_1 \geq c_0$ such that for all $c \geq c_1$, (x^*, λ^*) is a global saddle point of L_3 :*

$$(3.17) \quad L_3(x^*, \lambda, c) \leq L_3(x^*, \lambda^*, c) \leq L_3(x, \lambda^*, c)$$

for any $x \in X$ and $\lambda \geq 0$, where L_3 is defined in (1.7) with conditions (B1)'–(B3)' for φ being satisfied.

Proof. By Theorem 2.3, it suffices to show that there exists a $c_1 \geq c_0$ such that the second inequality in (3.17) holds for all $x \in X \setminus N(x^*, \delta)$ when $c \geq c_1$. Similar to the proof of Theorem 3.1, there exists $\epsilon_1 > 0$ such that (3.6) holds. By the definition of L_3 , $L_3(x, \lambda^*, c) = \infty$ for any $x \in X \setminus \Omega_c$. Thus, we only need to show that the second inequality of (3.17) holds for any $x \in (X \cap \Omega_c) \setminus N(x^*, \delta)$. By the definitions of Ω_c and F_{ϵ_1} , $\Omega_c \subset F_{\epsilon_1}$ for sufficiently large $c > 0$. Hence, from (3.6), we have

$$(3.18) \quad [\Omega_c \setminus N(x^*, \delta)] \cap [U_{\epsilon_1} \setminus N(x^*, \delta)] = \emptyset$$

for sufficiently large $c > 0$. Now, suppose $x \in (X \cap \Omega_c) \setminus N(x^*, \delta)$. It follows from (3.18) that $x \notin U_{\epsilon_1}$, i.e.,

$$(3.19) \quad f(x) > f(x^*) + \epsilon_1.$$

Let $I_-(x) = \{i \mid g_i(x) < 0\}$. Then, by conditions (B1)' and (B2)', we have

$$(3.20) \quad \varphi(cg_i(x)) > 0 \quad \text{for any } i \notin I_-(x).$$

On the other hand, condition (B3)' and the compactness of X imply that

$$(1/c)\varphi(cg_i(x)) \rightarrow 0$$

uniformly for any $x \in X \cap \Omega_c$ and $i \in I_-(x)$, as $c \rightarrow \infty$. Thus, there exists $c_1 \geq c_0$ such that

$$(3.21) \quad \frac{1}{c} \sum_{i \in I_-(x)} \lambda_i^* \varphi(cg_i(x)) \geq -\frac{1}{2}\epsilon_1 \quad \text{for any } c \geq c_1.$$

Combining (3.19), (3.20), and (3.21), for any $c \geq c_1$, we obtain

$$\begin{aligned} L_3(x, \lambda^*, c) &= f(x) + \frac{1}{c} \sum_{i=1}^m \lambda_i^* \varphi(cg_i(x)) \\ &> f(x^*) + \epsilon_1 + \frac{1}{c} \sum_{i \in I_-(x)} \lambda_i^* \varphi(cg_i(x)) \\ &\geq f(x^*) + \frac{1}{2}\epsilon_1 \\ &> f(x^*) = L_3(x^*, \lambda^*, c). \end{aligned}$$

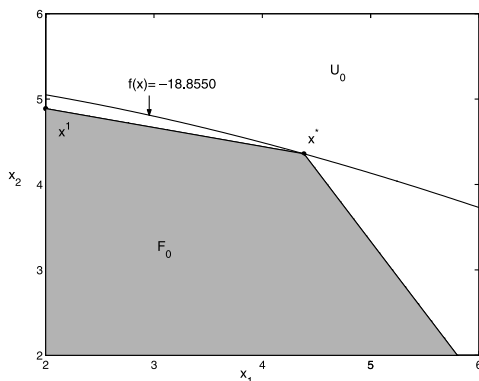
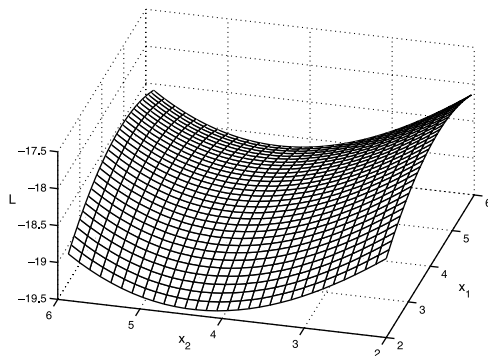


FIG. 1. The feasible region of Example 3.1.

FIG. 2. The classical Lagrangian $L(x, \lambda^*)$ of Example 3.1.

This completes the proof. $\square$

Finally, we derive the conditions for the existence of a global saddle point of L_4 defined in (1.9). We need an additional condition for function ξ :

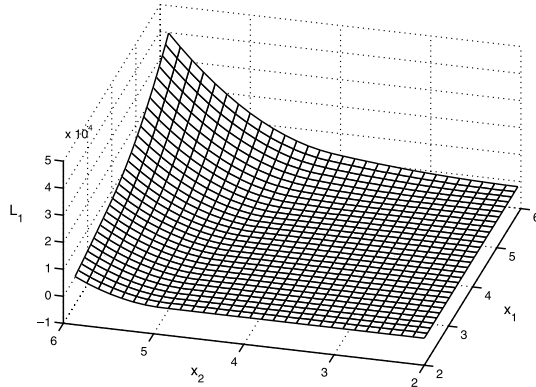
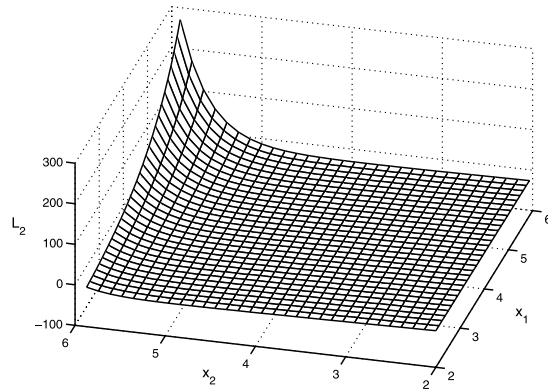
$$(C3) \lim_{t \rightarrow \infty} \xi(t)/t = \infty.$$

THEOREM 3.4. *Let x^* be a unique global optimal solution of (P). Suppose that X is compact and Assumption 2.1 is satisfied at x^* . Then there exists $c_1 \geq c_0$ such that for all $c \geq c_1$, (x^*, λ^*) is a global saddle point of L_4 :*

$$(3.22) \quad L_4(x^*, \lambda, c) \leq L_4(x^*, \lambda^*, c) \leq L_4(x, \lambda^*, c)$$

for any $x \in X$ and $\lambda \geq 0$, where L_4 is defined in (1.9) with conditions (B1)–(B3) for ψ and conditions (C1)–(C3) for ξ being satisfied.

Proof. By the uniqueness of x^* , we can show that (3.6) holds. The reminder of the proof is similar to that of Theorem 3.2 by replacing S_ϵ with F_ϵ . We mention here the major differences in case (a) when $x \in U_{\epsilon_1} \setminus N(x^*, \delta)$. By (3.6), it holds that $x \notin F_{\epsilon_1}$. We have $g_{i_0}(x) > \epsilon_1$ for some $i_0 \in \{1, \dots, m\}$. Since $\xi(t)$ is strictly increasing for $t > 0$, by using condition (C3), we have $(1/c)\xi(cg_{i_0}(x)) > (1/c)\xi(c\epsilon_1) \rightarrow \infty$ ($c \rightarrow \infty$).

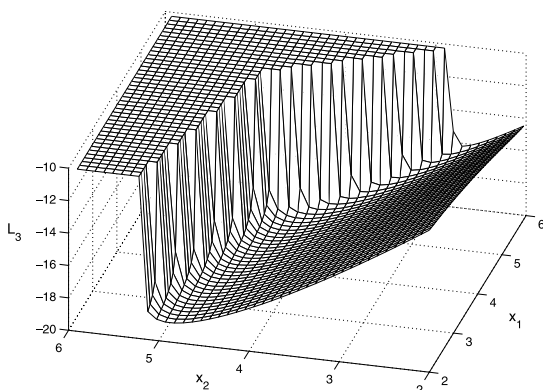
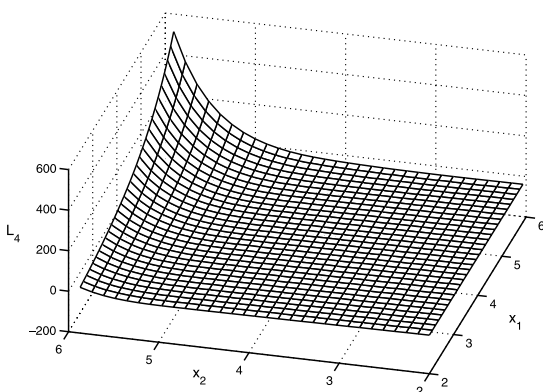
FIG. 3. $L_1(x, \lambda^*, c)$ of Example 3.1.FIG. 4. $L_2(x, \lambda^*, c)$ of Example 3.1.

Thus,

$$\begin{aligned}
 L_4(x, \lambda^*, c) &= f(x) + (1/c) \sum_{i=1}^m [\lambda_i^* \psi(cg_i(x)) + \xi(cg_i(x))] \\
 &> \underline{f} + (1/c)\xi(c\epsilon_1) + \underline{\psi} \sum_{i=1}^m \lambda_i^* \\
 &\rightarrow \infty, \quad (c \rightarrow \infty). \quad \square
 \end{aligned}$$

A prominent property of L_3 and L_4 is that by introducing the barrier property in L_3 and incorporating the penalty term ξ into L_2 , respectively, the compactness of X and the uniqueness of x^* are sufficient to guarantee the existence of a global saddle point, without appealing to the separation condition (3.11).

The above theorems on global saddle points guarantee a zero duality gap of the four augmented Lagrangian functions $L_i (i = 1, 2, 3, 4)$ in constrained global optimization. A conceptual primal-dual augmented Lagrangian method can then be devised for searching a global saddle point. At k -iteration, the primal problem is a global

FIG. 5. $L_3(x, \lambda^*, c)$ of Example 3.1.FIG. 6. $L_4(x, \lambda^*, c)$ of Example 3.1.

unconstrained problem,

$$(3.23) \quad \min_{x \in X} L_i(x, \lambda^k, c_k),$$

where c_k can either take a sufficient large constant or be updated at each iteration. Let x^{k+1} be the global solution of (3.23). The new multiplier vector λ^{k+1} is updated by a certain solution scheme for solving the dual problem corresponding to L_i . It is expected that under the similar conditions assumed in the theorems of this section, $\{x^k\}$ and $\{\lambda^k\}$ converge to x^* and λ^* , respectively, where (x^*, λ^*) is a global saddle point pair of L_i .

Let us consider an example to illustrate the above theorems.

Example 3.1.

$$\begin{aligned} \min \quad & f(x) = -(1/10)x_1^2 + (1/60)x_2^3 - (1/5)x_1 - 4x_2 \\ \text{subject to} \quad & g_1(x) = 2x_1 + 9x_2 \leq 48, \\ & g_2(x) = 5x_1 + 3x_2 \leq 35, \\ & x \in X = \{x \mid 2 \leq x_i \leq 6, \ i = 1, 2\}. \end{aligned}$$

The objective function of this example is nonconvex. The problem has two local optimal solutions: $x^1 = (2, 4.8889)^T$ with $f(x^1) = -18.4081$ and $x^2 = (4.3846, 4.3590)^T$ with $f(x^2) = -18.8550$. The optimal multipliers corresponding to the global solution $x^* = x^2$ is $\lambda^* = (0.3082, 0.09211)^T$. The feasible region of the example is shown in Figure 1. We have $\nabla g_1(x^*) = (2, 9)^T$, $\nabla g_2(x^*) = (5, 3)^T$, and $\nabla_{xx}^2 L(x^*, \lambda^*) = \text{diag}(-0.2, 0.4359)$. It is thus clear that there does not exist a saddle point for the conventional Lagrangian function. Figure 2 illustrates the picture of the classical Lagrangian $L(x, \lambda^*)$.

Since for any $y \in \mathbb{R}^2$, $[\nabla g_1(x^*), \nabla g_2(x^*)]y = 0$ implies $y = 0$, the second-order sufficient conditions hold trivially. Moreover, the global optimal solution is unique and X is compact. Thus the conditions in Theorems 3.1–3.4 are fulfilled. Now, we choose the parameters and the functions ϕ , ψ , and ξ for L_i ($i = 1, 2, 3, 4$) as follows: (1) $L_1(x, \lambda, c)$: $\phi(s) = (1/2)s^2$ and $c = 200$; (2) $L_2(x, \lambda, c)$: $\psi(t) = e^t - 1$ and $c = 0.35$; (3) $L_3(x, \lambda, c)$: $\varphi(t) = -\ln(1-t)$ and $c = 0.35$; (4) $L_4(x, \lambda, c)$: $\psi(t) = e^t - 1$, $\xi(t) = (1/3)[\max(t, 0)]^3$, and $c = 0.35$. Note $\psi''(0) = \varphi''(0) = 1$. By (2.19) and (2.22), we can calculate

$$\nabla_{xx}^2 L_i(x^*, \lambda^*, c) = \nabla_{xx}^2 L(x^*, \lambda^*) + c \sum_{j=1}^2 \lambda_j^* \nabla g_j(x^*) \nabla g_j(x^*)^T = \begin{pmatrix} 1.0374 & 2.4252 \\ 2.4252 & 9.4653 \end{pmatrix}$$

for $i = 2, 3, 4$, which is a positive definite matrix. It can be also verified that with the above parameters for L_i ($i = 1, 2, 3, 4$), x^* solves problem $\min_{x \in X} L_i(x, \lambda^*, c)$ globally and therefore (x^*, λ^*) is a global saddle point of L_i ($i = 1, 2, 3, 4$). Figures 3–6 illustrate the pictures of $L_i(x, \lambda^*, c)$, $i = 1, 2, 3, 4$, respectively.

4. Concluding remarks. The saddle point of the classical Lagrangian function is only a sufficiency condition for optimality in nonconvex optimization. There often does not exist a saddle point of the classical Lagrangian function at a local or a global optimal solution of a nonconvex optimization problem. In this paper we have systematically investigated properties of the local and global duality of four types of augmented Lagrangian functions. The theorems in section 2 for local saddle points of the augmented Lagrangians extend and strengthen the state-of-the-art in the literature. The new results established in section 3 for global saddle points of the augmented Lagrangians offer fresh insights to better understanding the conditions to achieve zero duality properties in constrained global optimization. Most important, the results in this paper have revealed that the global saddle point condition associated with the four types of augmented Lagrangian functions is both a necessary and sufficient global optimality condition for nonconvex constrained optimization problems under certain conditions. Computational investigation of the use of augmented Lagrangians for constrained global optimization will be carried out in future research.

Acknowledgments. The authors are grateful to the three anonymous referees for their valuable comments and suggestions, which have contributed to an improvement of the paper. In particular, the authors would like to express their appreciation to one of the referees who has drawn the authors' attention to the modified penalty/barrier functions.

REFERENCES

- [1] A. BEN-TAL AND M. ZIBULEVSKY, *Penalty/barrier multiplier methods for convex programming problems*, SIAM J. Optim., 7 (1997), pp. 347–366.

- [2] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.
- [3] M. R. HESTENES, *Multiplier and gradient methods*, J. Optim. Theory Appl., 4 (1969), pp. 303–320.
- [4] R. HORST AND H. TUY, *Global Optimization: Deterministic Approaches*, 2nd ed., Springer-Verlag, Heidelberg, 1993.
- [5] R. HORST, P. M. PARDALOS, AND N. V. THOAI, *Introduction to Global Optimization*, 2nd ed., Nonconvex Optim. Appl., 48, Kluwer Academic Publishers, Dordrecht, 2000.
- [6] X. X. HUANG AND X. Q. YANG, *A unified augmented Lagrangian approach to duality and exact penalization*, Math. Oper. Res., 28 (2003), pp. 524–532.
- [7] S. KARLIN, *Mathematical Methods and Theory in Games, Programming and Economics*, Vol. 1, Addison-Wesley, Reading, MA, 1959.
- [8] B. W. KORT AND D. P. BERTSEKAS, *A new penalty method for constrained minimization*, in Proceedings of the IEEE Conference on Decision and Control, New Orleans, 1972, pp. 162–166.
- [9] D. LI, *Zero duality gap for a class of nonconvex optimization problems*, J. Optim. Theory Appl., 85 (1995), pp. 309–324.
- [10] D. LI, *Saddle-point generation in nonlinear nonconvex optimization*, Nonlinear Anal., 30 (1997), pp. 4339–4344.
- [11] D. LI AND X. L. SUN, *Existence of a saddle point in nonconvex constrained optimization*, J. Global Optim., 21 (2001), pp. 39–50.
- [12] D. LI AND X. L. SUN, *Convexification and existence of saddle point in a p th-power reformulation for nonconvex constrained optimization*, Nonlinear Anal., 47 (2001), pp. 5611–5622.
- [13] O. L. MANGASARIAN, *Nonlinear Programming*, Classics Appl. Math. 10, SIAM, Philadelphia, 1994.
- [14] M. MINOUX, *Mathematical Programming: Theory and Algorithms*, John Wiley and Sons, Chichester, UK, 1986.
- [15] V. H. NGUYEN AND J. J. STRODIOT, *On the convergence rate for a penalty function method of exponential type*, J. Optim. Theory Appl., 27 (1979), pp. 495–508.
- [16] R. POLYAK, *Modified barrier functions: Theory and methods*, Math. Programming, 54 (1992), pp. 177–222.
- [17] R. POLYAK AND I. GRIVA, *Primal-dual nonlinear rescaling method for convex optimization*, J. Optim. Theory Appl., 122 (2004), pp. 111–156.
- [18] M. J. D. POWELL, *A method for nonlinear constraints in minimization problems*, in Optimization, R. Fletcher, ed., Academic Press, London, 1969, pp. 283–298.
- [19] R. T. ROCKAFELLAR, *Augmented Lagrange multiplier functions and duality in nonconvex programming*, SIAM J. Control Optim., 12 (1974), pp. 268–285.
- [20] R. T. ROCKAFELLAR, *Lagrange multipliers and optimality*, SIAM Rev., 35 (1993), pp. 183–238.
- [21] R. T. ROCKAFELLAR AND R. J.-B. WETS, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
- [22] J. B. ROSEN AND P. M. PARDALOS, *Constrained Global Optimization: Algorithms and Applications*, Lecture Notes in Comput. Sci. 268, Springer-Verlag, Berlin, 1987.
- [23] A. M. RUBINOV, X. X. HUANG, AND X. Q. YANG, *The zero duality gap property and lower semicontinuity of the perturbation function*, Math. Oper. Res., 27 (2002), pp. 775–791.
- [24] A. M. RUBINOV, X. X. HUANG, AND X. Q. YANG, *Lagrange-Type Functions in Constrained Non-convex Optimization*, Kluwer Academic Publishers, Boston, 2003.
- [25] X. L. SUN AND D. LI, *Value-estimation function method for constrained global optimization*, J. Optim. Theory Appl., 102 (1999), pp. 385–409.
- [26] X. L. SUN, K. MCKINNON, AND D. LI, *A convexification method for a class of global optimization problems with applications to reliability optimization*, J. Global Optim., 21 (2001), pp. 185–199.
- [27] P. T. THATCH AND H. TUY, *The relief indicator method for constrained global optimization*, Naval Res. Logist., 37 (1990), pp. 473–497.
- [28] P. TSENG, *On the convergence of the exponential multiplier method for convex programming*, Math. Programming, 60 (1993), pp. 1–19.
- [29] H. TUY, *Monotonic optimization: Problems and solution approaches*, SIAM J. Optim., 11 (2000), pp. 464–494.

A NOTE ON THE LOCAL CONVERGENCE OF A PREDICTOR-CORRECTOR INTERIOR-POINT ALGORITHM FOR THE SEMIDEFINITE LINEAR COMPLEMENTARITY PROBLEM BASED ON THE ALIZADEH–HAEBERLY–OVERTON SEARCH DIRECTION*

ZHAOSONG LU[†] AND RENATO D. C. MONTEIRO[†]

Abstract. This note points out an error in the local quadratic convergence proof of the predictor-corrector interior-point algorithm for solving the semidefinite linear complementarity problem based on the Alizadeh–Haeberly–Overton search direction presented in [M. Kojima, M. Shida, and S. Shindoh, *SIAM J. Optim.*, 9 (1999), pp. 444–465]. Their algorithm is slightly modified and the local quadratic convergence of the resulting method is established.

Key words. semidefinite linear complementarity problem, semidefinite programming, interior-point algorithm, predictor-corrector algorithm, local quadratic convergence

AMS subject classifications. 90C22, 90C25, 90C30, 65K05

DOI. 10.1137/04060531X

1. Introduction. Let $\mathcal{S}$ denote the set of all $n \times n$ symmetric real matrices. Given matrices X and Y in $\mathbb{R}^{p \times q}$, the standard inner product is defined by $X \bullet Y \equiv \text{tr}(X^T Y)$, where $\text{tr}(\cdot)$ denotes the trace of a matrix. The Euclidean norm and its associated operator norm, i.e., the spectral norm, are both denoted by $\|\cdot\|$. The Frobenius norm of a $p \times q$ -matrix X is defined as $\|X\|_F \equiv (X \bullet X)^{1/2}$. If $X \in \mathcal{S}$ is positive semidefinite (resp., definite), we write $X \succeq 0$ (resp., $X \succ 0$). The cone of positive semidefinite (resp., definite) matrices is denoted by $\mathcal{S}_+$ (resp., $\mathcal{S}_{++}$). The identity matrix will be denoted by I .

Let $\mathcal{F}$ be a $n(n+1)/2$ -dimensional affine subspace of $\mathcal{S} \times \mathcal{S}$, and

$$\mathcal{F}_+ = \{(X, Y) \in \mathcal{F} : X \succeq 0, Y \succeq 0\}.$$

We are concerned with the semidefinite linear complementarity problem (SDLCP):

$$(1.1) \quad \text{find a } (X, Y) \in \mathcal{F}_+ \text{ such that } X \bullet Y = 0.$$

We call a $(X, Y) \in \mathcal{F}_+$ a feasible solution of the SDLCP (1.1). Throughout this note we assume the monotonicity of the affine subspace $\mathcal{F}$:

$$(U' - U) \bullet (V' - V) \geq 0 \text{ for every } (U', V'), (U, V) \in \mathcal{F}.$$

Kojima, Shida, and Shindoh [3] have proposed a globally convergent Mizuno–Todd–Ye-type predictor-corrector infeasible-interior-point algorithm (Algorithm 2.1 of [3]), with the use of the Alizadeh–Haeberly–Overton (AHO) search direction, for the monotone SDLCP (1.1), and demonstrated its local quadratic convergence under the strict complementarity condition.

*Received by the editors March 16, 2004; accepted for publication (in revised form) November 24, 2004; published electronically August 3, 2005.

<http://www.siam.org/journals/siopt/15-4/60531.html>

[†]School of ISyE, Georgia Tech, Atlanta, GA 30332 (zhaosong@isye.gatech.edu, monteiro@isye.gatech.edu). The second author was supported in part by NSF grants CCR-0203113 and INT-9910084 and ONR grant N00014-03-1-0401.

This note has two purposes. One is to point out an error in the proof of the local quadratic convergence of the algorithm presented in [3]. The other is to describe a modified variant of this method and establish its local quadratic convergence.

This note is organized as follows. In section 2, we describe the algorithm presented in [3] and point out an error made in [3] on the proof of its local quadratic convergence. In section 3, we describe a slight modification of this algorithm and establish the local quadratic convergence of the resulting method.

1.1. Notation. Given functions $f : \Omega \rightarrow E$ and $g : \Omega \rightarrow \mathfrak{R}_{++}$, where Ω is an arbitrary set and E is a normed vector space, and a subset $\tilde{\Omega} \subset \Omega$, we write $f(w) = \mathcal{O}(g(w))$ for all $w \in \tilde{\Omega}$ to mean that there exists a constant $M > 0$ such that $\|f(w)\| \leq M g(w)$ for all $w \in \tilde{\Omega}$; moreover, for a function $U : \Omega \rightarrow \mathcal{S}_{++}$, we write $U(w) = \Theta(g(w))$ for all $w \in \tilde{\Omega}$ if $U(w) = \mathcal{O}(g(w))$ and $U(w)^{-1} = \mathcal{O}(1/g(w))$ for all $w \in \tilde{\Omega}$. The latter condition is equivalent to the existence of a constant $M > 0$ such that

$$\frac{1}{M}I \preceq \frac{1}{g(w)}U(w) \preceq MI \quad \forall w \in \Omega.$$

2. A predictor-corrector interior-point algorithm. In this section, we describe the predictor-corrector infeasible-interior-point algorithm using AHO search direction (Algorithm 2.1 in [3]) for monotone SDLCP (1.1), and point out an error in the proof of its local quadratic convergence in Theorem 5.1 of [3].

Throughout this note we use the same notation as in [3],

$$\begin{aligned} \zeta &: \text{ a constant not less than } 1/\sqrt{n}, \\ \mathcal{F}_0 &= \{(U', V') - (U, V) : (U', V'), (U, V) \in \mathcal{F}\}, \\ \tilde{\mathcal{N}}(\gamma, \tau) &= \left\{ (X, Y) \in \mathcal{S}_+ \times \mathcal{S}_+ : \begin{array}{l} (XY + YX)/2 \succeq (1 - \gamma)\tau I, \\ X \bullet Y/n \leq (1 + \zeta\gamma)\tau \end{array} \right\}, \end{aligned}$$

for each $\gamma \in [0, 1]$ and each $\tau \geq 0$.

Before describing Algorithm 2.1 of [3], we recall Hypothesis 2.1 of [3].

Hypothesis 2.1 of [3]. Let $\omega^* \geq 1$. There exists a solution (X^*, Y^*) of SDLCP (1.1) such that

$$\omega^* X^0 \succeq X^* \text{ and } \omega^* Y^0 \succeq Y^*.$$

For notational convenience, we introduce one operator as follows:

$$H_I(M) = \frac{M + M^T}{2} \quad \forall M \in \mathfrak{R}^{n \times n}.$$

We are ready to describe Algorithm 2.1 of [3] as follows.

ALGORITHM 2.1 OF [3].

Step 0. Choose an accuracy parameter $\epsilon \geq 0$, a neighborhood parameter $\gamma \in (0, 1)$, and an initial point $(X^0, Y^0) = (\sqrt{\mu^0}I, \sqrt{\mu^0}I)$ with some $\mu^0 > 0$. Let $\theta^0 = 1$, $\sigma = 2\omega^*/(1 - \gamma) + 1$, $\gamma^0 = 0$, and $k = 0$.

Step 1. If the inequality

$$\theta^k(X^0 \bullet Y^k + X^k \bullet Y^0) \leq \sigma X^k \bullet Y^k$$

does not hold, then stop.

Step 2 (predictor step). Compute a solution (dX_p^k, dY_p^k) of the system of equations

$$\begin{aligned} H_I(X^k dY_p^k + dX_p^k Y^k) &= -H_I(X^k Y^k), \\ (X^k + dX_p^k, Y^k + dY_p^k) &\in \mathcal{F}. \end{aligned}$$

Let

$$\begin{aligned} \delta_p^k &= \frac{\|dX_p^k\|_F \|dY_p^k\|_F}{\theta^k \mu^0}, \\ \hat{\alpha}_p^k &= \frac{2}{\sqrt{1 + 4\delta_p^k/(\gamma - \gamma^k)} + 1}, \\ \check{\alpha}_p^k &= \max \left\{ \alpha' \in [0, 1] : \begin{array}{l} (X^k + \alpha dX_p^k, Y^k + \alpha dY_p^k) \\ \in \tilde{\mathcal{N}}(\gamma, (1 - \alpha)\theta^k \mu^0) \\ \text{for every } \alpha \in [0, \alpha'] \end{array} \right\}. \end{aligned}$$

Choose a step length $\alpha_p^k \in [\hat{\alpha}_p^k, \check{\alpha}_p^k]$. Let

$$(X_c^k, Y_c^k) = (X^k, Y^k) + \alpha_p^k (dX_p^k, dY_p^k) \text{ and } \theta^{k+1} = (1 - \alpha_p^k)\theta^k.$$

Step 3. If $\theta^{k+1} \leq \epsilon$, then stop. If the inequality

$$(2.1) \quad \theta^{k+1}(X^0 \bullet Y_c^k + X_c^k \bullet Y^0) \leq \sigma X_c^k \bullet Y_c^k$$

does not hold, then stop.

Step 4 (corrector step). Compute a solution (dX_c^k, dY_c^k) of the system of equations

$$(2.2) \quad \begin{cases} H_I(X_c^k dY_c^k + dX_c^k Y_c^k) = \theta^{k+1} \mu^0 I - H_I(X_c^k Y_c^k), \\ (dX_c^k, dY_c^k) \in \mathcal{F}_0. \end{cases}$$

Let

$$(2.3) \quad \begin{aligned} \delta_c^k &= \frac{\|dX_c^k\|_F \|dY_c^k\|_F}{\theta^{k+1} \mu^0}, \\ \hat{\alpha}_c^k &= \begin{cases} \gamma/(2\delta_c^k) & \text{if } \gamma \leq 2\delta_c^k, \\ 1 & \text{if } \gamma > 2\delta_c^k, \end{cases} \\ \check{\gamma}^{k+1} &= \begin{cases} \gamma(1 - \gamma/(4\delta_c^k)) & \text{if } \gamma \leq 2\delta_c^k, \\ \delta_c^k & \text{if } \gamma > 2\delta_c^k, \end{cases} \\ \hat{\gamma}^{k+1} &= \min \left\{ \gamma' \in [0, 1] : \begin{array}{l} (X_c^k + \alpha dX_c^k, Y_c^k + \alpha dY_c^k) \\ \in \tilde{\mathcal{N}}(\gamma', \theta^{k+1} \mu^0) \\ \text{for some } \alpha \in [0, 1] \end{array} \right\}. \end{aligned}$$

Choose a step length $\alpha_c^k \in [0, 1]$ and γ^{k+1} such that

$$\begin{aligned} \hat{\gamma}^{k+1} &\leq \gamma^{k+1} \leq \check{\gamma}^{k+1}, \\ (X_c^k + \alpha_c^k dX_c^k, Y_c^k + \alpha_c^k dY_c^k) &\in \tilde{\mathcal{N}}(\gamma^{k+1}, \theta^{k+1} \mu^0). \end{aligned}$$

(It has been shown in Lemma 3.8 of [3] that the pair of $\alpha_c^k = \hat{\alpha}_c^k$ and $\gamma^{k+1} = \check{\gamma}^{k+1}$ satisfies the relation above.) Let $(X^{k+1}, Y^{k+1}) = (X_c^k, Y_c^k) + \alpha_c^k (dX_c^k, dY_c^k)$.

Step 5. Replace k by $k + 1$. Go to Step 1.

Before ending this section, we remark that the proof of Theorem 5.1 (local convergence theorem) of [3] is not correct since it is based on the claim that $\delta_c^k = \mathcal{O}(1)$, which in turn was incorrectly established in the proof of this result. Indeed, in the first two lines of the proof of Theorem 5.1 of [3], the authors claimed that $\delta_c^k = \mathcal{O}(1)$ holds by (iii) of Lemma 3.1, the definition of δ_c^k , and the fact that $2\theta^k I - (X^k Y^k + Y^k X^k) = \mathcal{O}(\theta^k)$. However, from those arguments we can only conclude $\delta_c^k = \mathcal{O}(\frac{1}{\theta^{k+1}})$. Let us investigate this proof in more detail. From Step 2 of Algorithm 2.1 of [3], we see that

$$(2.4) \quad (X_c^k, Y_c^k) \in \tilde{\mathcal{N}}(\gamma, \theta^{k+1} \mu^0),$$

which, together with (2.1) and Lemma 3.4 of [3], implies that $(X_c^k, Y_c^k) = \mathcal{O}(1)$. Also, by (2.4) and Lemma 3.1 (i) of [3], we have $H_I(X_c^k Y_c^k) = \mathcal{O}(\theta^{k+1})$, which implies that

$$(2.5) \quad \theta^{k+1} \mu^0 I - H_I(X_c^k Y_c^k) = \mathcal{O}(\theta^{k+1}).$$

Now, using (2.4), (2.2), (2.5); Lemma 3.1 (iii) of [3]; and the fact $(X_c^k, Y_c^k) = \mathcal{O}(1)$, we have

$$(2.6) \quad \|dX_c^k\|_F \leq \frac{2\|X_c^k\|_F \|\theta^{k+1} \mu^0 I - H_I(X_c^k Y_c^k)\|_F}{(1-\gamma)\theta^{k+1} \mu^0} = \mathcal{O}(1).$$

Similarly, we have $\|dY_c^k\|_F = \mathcal{O}(1)$, which together with (2.6) and (2.3) implies that $\delta_c^k = \mathcal{O}(\frac{1}{\theta^{k+1}})$. Due to this and [2], we believe that the claim $\delta_c^k = \mathcal{O}(1)$ does not hold for general SDLCPs, even though it holds under a suitable nondegeneracy assumption on the SDLCP, namely Condition 6.1 of [3] (see the proof in section 6 of [3]). Hence, Algorithm 2.1 of [3] can only be claimed to be locally quadratically convergent for nondegenerate SDLCPs. In the next section, we will describe a slight modification of Algorithm 2.1 of [3] which is locally quadratically convergent.

3. Slightly modified algorithm. In this section, we describe a slight modification of Algorithm 2.1 of [3] and establish its local quadratic convergence.

The modified algorithm is the same as before except that the definition of δ_c^k in (2.3) is replaced by

$$(3.1) \quad \delta_c^k = \frac{\|dX_c^k dY_c^k\|_F}{\theta^{k+1} \mu^0}.$$

Accordingly, we refer to the modified algorithm as Algorithm 2.1'. Our main effort from now on will be to establish that the quantity δ_c^k , as defined in (3.1), has the property that $\delta_c^k = \mathcal{O}(1)$.

First, we will argue that Algorithm 2.1' is globally convergent. It can be shown that Lemmas 3.1–3.7 of [3] also hold for Algorithm 2.1'. The next result shows that Lemma 3.8 also holds for Algorithm 2.1' if $\zeta \geq 1/\sqrt{n}$.

LEMMA 3.1. *For Algorithm 2.1', if $\zeta \geq 1/\sqrt{n}$, Lemma 3.8 in [3] holds, where ζ is a constant defined at the beginning of section 2 of [3].*

Proof. Using the fact that $H_I(dX_c^k dY_c^k) \geq -\|dX_c^k dY_c^k\|_F I$ and $dX_c^k \bullet dY_c^k \leq \sqrt{n}\|dX_c^k dY_c^k\|_F$, and the condition $\zeta \geq 1/\sqrt{n}$, we can show that the conclusion holds in a similar way as the proof given in Lemma 3.8 of [3]. $\square$

Using Lemmas 3.1–3.7 of [3] and Lemma 3.1 and following the same proof as the one given in Theorem 2.1 of [3], we see that Theorem 2.1 (global convergence theorem) in [3] also holds for Algorithm 2.1'; namely, Algorithm 2.1' is globally convergent.

We will now show that Algorithm 2.1' is locally quadratically convergent under the following standard condition commonly used in the local convergence analysis of interior-point algorithms for SDLCP.

Condition 5.1 of [3] (strict complementarity). There is a solution (X^*, Y^*) of SDLCP (1.1) such that $X^* + Y^* \succ 0$.

We next state and prove some technical results. The first one is due to Monteiro and Tsuchiya [4].

LEMMA 3.2 (Lemma 2.1 of [4]). *For every $A \in \mathcal{S}_{++}$ and $H \in \mathcal{S}$, the equation $AU + UA = H$ has a unique solution $U \in \mathcal{S}$. Moreover, this solution satisfies $\|AU\|_F \leq \|H\|_F/\sqrt{2}$.*

Under Condition 5.1 of [3], we have a solution (X^*, Y^*) of the SDLCP (1.1) satisfying $X^* + Y^* \succ 0$. Since X^* and Y^* commute, there exists an orthogonal matrix Q such that

$$Q^T X^* Q = \begin{pmatrix} \Lambda_B & 0 \\ 0 & 0 \end{pmatrix}, \quad Q^T Y^* Q = \begin{pmatrix} 0 & 0 \\ 0 & \Lambda_N \end{pmatrix},$$

where Λ_B and Λ_N are positive diagonal matrices with dimension m and $n - m$ for some $m \in \{0, 1, 2, \dots, n\}$, respectively. For each $(X, Y) \in \mathcal{S} \times \mathcal{S}$, define the following optimal partition:

$$Q^T X Q \equiv \hat{X} = \begin{pmatrix} \hat{X}_B & \hat{X}_J \\ \hat{X}_J^T & \hat{X}_N \end{pmatrix}, \quad Q^T Y Q \equiv \hat{Y} = \begin{pmatrix} \hat{Y}_B & \hat{Y}_J \\ \hat{Y}_J^T & \hat{Y}_N \end{pmatrix}.$$

LEMMA 3.3. *Assume that $(X, Y) \in \tilde{\mathcal{N}}(\gamma, \tau)$. Let $(dX(\tau), dY(\tau))$ be a solution of the system of equations*

$$(3.2) \quad H_I(dX \ Y + X \ dY) = \tau I - H_I(XY), \quad (dX, dY) \in \mathcal{F}_0,$$

$\widehat{dX} \equiv Q^T dX Q$ and $\widehat{dY} \equiv Q^T dY Q$. Under Condition 5.1 of [3] and $\zeta \geq 1/\sqrt{n}$, there then holds

$$(3.3) \quad \widehat{dX}_B(\tau) = \mathcal{O}(1), \quad \widehat{dX}_N(\tau) = \mathcal{O}(\tau),$$

$$(3.4) \quad \widehat{dY}_B(\tau) = \mathcal{O}(\tau), \quad \widehat{dY}_N(\tau) = \mathcal{O}(1).$$

Proof. For notational convenience, we will use $\widehat{dX}$ and $\widehat{dY}$ to denote $\widehat{dX}(\tau)$ and $\widehat{dY}(\tau)$, respectively. Using Lemmas 5.3 and 5.5 of [3], we have

$$(3.5) \quad \hat{X} = Q^T X Q = \begin{pmatrix} \Theta(1) & \mathcal{O}(\tau) \\ \mathcal{O}(\tau) & \mathcal{O}(\tau) \end{pmatrix},$$

$$(3.6) \quad \hat{Y} = Q^T Y Q = \begin{pmatrix} \mathcal{O}(\tau) & \mathcal{O}(\tau) \\ \mathcal{O}(\tau) & \Theta(1) \end{pmatrix}.$$

This immediately implies that $X = \mathcal{O}(1)$ and $Y = \mathcal{O}(1)$. In view of Lemma 3.1 (i) of [3] and the definition of $H_I(\cdot)$, we immediately see that

$$(3.7) \quad \tau I - H_I(XY) = \mathcal{O}(\tau).$$

Letting $C = 2(\tau I - H_I(XY))$ and using Lemma 3.1 (iii) of [3], we obtain that

$$\|\widehat{dX}\|_F = \|dX\|_F \leq \frac{\|X\|_F \|C\|_F}{(1 - \gamma)\tau} = \mathcal{O}(1).$$

Hence, $\widehat{dX} = \mathcal{O}(1)$. Similarly, we can show that $\widehat{dY} = \mathcal{O}(1)$. Note that the system (3.2) can be written as

$$(3.8) \quad H_I(\widehat{dX} \hat{Y} + \hat{X} \widehat{dY}) = \tau I - H_I(\hat{X} \hat{Y}), \quad (\widehat{dX}, \widehat{dY}) \in \hat{\mathcal{F}}_0,$$

where $\hat{\mathcal{F}}_0 \equiv \{M = Q^T P Q : P \in \mathcal{F}_0\}$. From (3.7), we easily see that

$$(3.9) \quad \tau I - H_I(\hat{X} \hat{Y}) = \mathcal{O}(\tau).$$

Using this fact and (3.8), we obtain that

$$(3.10) \quad H_I(\widehat{dX}_B \hat{Y}_B + \widehat{dX}_J \hat{Y}_J^T + \hat{X}_B \widehat{dY}_B + \hat{X}_J \widehat{dY}_J^T) = \mathcal{O}(\tau).$$

Using (3.5), (3.6), (3.10), and the fact that $\widehat{dX} = \mathcal{O}(1)$ and $\widehat{dY} = \mathcal{O}(1)$, we have

$$H_I(\hat{X}_B \widehat{dY}_B) = \mathcal{O}(\tau),$$

which together with (3.5) and Lemma 3.2 implies $\widehat{dY}_B = \mathcal{O}(\tau)$. We can show that $\widehat{dX}_N = \mathcal{O}(\tau)$ in a similar way. $\square$

LEMMA 3.4. Assume that $(X, Y) \in \tilde{\mathcal{N}}(\gamma, \tau)$. Let $(\widehat{dX}(\tau), \widehat{dY}(\tau))$ be defined in Lemma 3.3. Under Condition 5.1 of [3] and $\zeta \geq 1/\sqrt{n}$, there then holds

$$\begin{aligned} \|\widehat{dY}_J(\tau)\| &= \Theta(\|\widehat{dX}_J(\tau)\|) + \mathcal{O}(\tau), \\ -\widehat{dX}_J(\tau) \bullet \widehat{dY}_J(\tau) &= \Theta(\|\widehat{dX}_J(\tau)\|^2) + \mathcal{O}(\tau \|\widehat{dX}_J(\tau)\|). \end{aligned}$$

Proof. For notational convenience, we will use $\widehat{dX}$ and $\widehat{dY}$ to denote $\widehat{dX}(\tau)$ and $\widehat{dY}(\tau)$, respectively. Using (3.9) and (3.8), we obtain that

$$\begin{aligned} &\widehat{dX}_B \hat{Y}_J + \widehat{dX}_J \hat{Y}_N + \hat{X}_B \widehat{dY}_J + \hat{X}_J \widehat{dY}_N \\ &+ \widehat{dY}_B \hat{X}_J + \widehat{dY}_J \hat{X}_N + \hat{Y}_B \widehat{dX}_J + \hat{Y}_J \widehat{dX}_N = \mathcal{O}(\tau). \end{aligned}$$

This identity together with (3.5), (3.6), (3.3), and (3.4) implies that

$$(3.11) \quad \widehat{dX}_J \hat{Y}_N + \hat{X}_B \widehat{dY}_J = \mathcal{O}(\tau).$$

Using this identity, we obtain that

$$(3.12) \quad \widehat{dY}_J = -\hat{X}_B^{-1}(\widehat{dX}_J \hat{Y}_N - \mathcal{O}(\tau)).$$

Using this identity ((3.5) and (3.6)), we see that the first conclusion follows. Using (3.12), (3.5), and (3.6), we obtain that

$$\begin{aligned} \widehat{dX}_J \bullet \widehat{dY}_J &= -\text{tr}(\widehat{dX}_J^T \hat{X}_B^{-1} \widehat{dX}_J \hat{Y}_N) + \mathcal{O}(\tau \|\widehat{dX}_J\|) \\ &= -\|(\hat{X}_B)^{-1/2} \widehat{dX}_J (\hat{Y}_N)^{1/2}\|_F^2 + \mathcal{O}(\tau \|\widehat{dX}_J\|), \end{aligned}$$

which together with (3.5) and (3.6) implies the second conclusion. $\square$

LEMMA 3.5. Assume that $(X, Y) \in \tilde{\mathcal{N}}(\gamma, \tau)$. Let $(\widehat{dX}(\tau), \widehat{dY}(\tau))$ be defined in Lemma 3.3. Under Condition 5.1 of [3] and $\zeta \geq 1/\sqrt{n}$, there then holds

$$\widehat{dX}_J(\tau) = \mathcal{O}(\tau), \quad \widehat{dY}_J(\tau) = \mathcal{O}(\tau).$$

Proof. Suppose that $\|\widehat{dX}_J(\tau)\| = \mathcal{O}(\tau)$ does not hold. Then there exists a sequence $\tau_k \downarrow 0$ as $k \rightarrow \infty$ such that $\tau_k = o(\|\widehat{dX}_J(\tau_k)\|)$. For convenience, we omit the index k from τ_k throughout the remaining proof. Then the above identity can be written as $\tau = o(\|\widehat{dX}_J(\tau)\|)$, which together with Lemma 3.4 implies that

$$(3.13) \quad \|\widehat{dY}_J(\tau)\| = \Theta(\|\widehat{dX}_J(\tau)\|),$$

$$(3.14) \quad -\widehat{dX}_J(\tau) \bullet \widehat{dY}_J(\tau) = \Theta(\|\widehat{dX}_J(\tau)\|^2).$$

For any $\tau > 0$, consider the linear system

$$(3.15) \quad (\widehat{dX}, \widehat{dY}) - (\widehat{dX}(\tau), \widehat{dY}(\tau)) \in \hat{\mathcal{F}}_0,$$

$$(3.16) \quad \widehat{dX}_J - \widehat{dX}_J(\tau) = -\widehat{dX}_J(\tau),$$

$$(3.17) \quad \widehat{dX}_N - \widehat{dX}_N(\tau) = -\widehat{dX}_N(\tau),$$

$$(3.18) \quad \widehat{dY}_J - \widehat{dY}_J(\tau) = -\widehat{dY}_J(\tau),$$

$$(3.18) \quad \widehat{dY}_B - \widehat{dY}_B(\tau) = -\widehat{dY}_B(\tau).$$

We see that any $(\widehat{dX}, \widehat{dY}) = (0, 0)$ is a feasible solution to this system. Hence, by Hoffman lemma [1] (see also Lemma A.3, p. 248 of [5]), there exists a sufficiently large constant $\hat{C}$ (independent on τ) such that for any $\tau > 0$, this system has a solution $(\overline{dX}, \overline{dY}) \in \mathcal{S} \times \mathcal{S}$ (dependent on τ) such that

$$(3.19) \quad \|(\overline{dX}, \overline{dY}) - (\widehat{dX}(\tau), \widehat{dY}(\tau))\| \leq \hat{C}(\|\widehat{dX}_N(\tau)\| + \|\widehat{dY}_B(\tau)\| + \|\widehat{dX}_J(\tau)\| + \|\widehat{dY}_J(\tau)\|).$$

Obviously, the monotonicity holds for $\hat{\mathcal{F}}_0$ due to the monotonicity of $\mathcal{F}_0$. Hence, we have

$$(\overline{dX} - \widehat{dX}(\tau)) \bullet (\overline{dY} - \widehat{dY}(\tau)) \geq 0.$$

Hence, it follows that

$$(3.20) \quad -(\overline{dX}_B - \widehat{dX}_B(\tau)) \bullet \widehat{dY}_B(\tau) + 2\widehat{dX}_J(\tau) \bullet \widehat{dY}_J(\tau) - \widehat{dX}_N(\tau) \bullet (\overline{dY}_N - \widehat{dY}_N(\tau)) \geq 0.$$

Note that $\|\overline{dX}_B - \widehat{dX}_B(\tau)\| \leq \|\overline{dX} - \widehat{dX}(\tau)\|$ and $\|\overline{dY}_N - \widehat{dY}_N(\tau)\| \leq \|\overline{dY} - \widehat{dY}(\tau)\|$. Using this fact, (3.20), (3.15)–(3.18), (3.19), (3.13), (3.14), (3.3), and (3.4), we obtain that, for all $\tau > 0$ sufficiently small,

$$\begin{aligned} |\widehat{dX}_J(\tau) \bullet \widehat{dY}_J(\tau)| &\leq \frac{1}{2} \left| (\overline{dX}_B - \widehat{dX}_B(\tau)) \bullet \widehat{dY}_B(\tau) + \widehat{dX}_N(\tau) \bullet (\overline{dY}_N - \widehat{dY}_N(\tau)) \right| \\ &\leq \tilde{C}\tau(\|\overline{dX} - \widehat{dX}(\tau)\| + \|\overline{dY} - \widehat{dY}(\tau)\|) \\ &\leq 2\tilde{C}\hat{C}\tau(\|\widehat{dX}_N(\tau)\| + \|\widehat{dY}_B(\tau)\| + \|\widehat{dX}_J(\tau)\| + \|\widehat{dY}_J(\tau)\|) \\ &\leq \tilde{C}\tau \left(\tau + \sqrt{|\widehat{dX}_J(\tau) \bullet \widehat{dY}_J(\tau)|} \right), \end{aligned}$$

where $\tilde{C}$ and $\hat{C}$ are some constants and the last inequality follows from (3.13) and (3.14). Let $\xi = \sqrt{|\widehat{dX}_J(\tau) \bullet \widehat{dY}_J(\tau)|}$. From the last inequality above, we have

$\xi^2 \leq \tilde{C}\tau(\tau + \xi)$, which together with the fact $\xi > 0$ implies $\xi \leq (\tilde{C} + \sqrt{5\tilde{C}})\tau/2$. Hence, $\xi = \mathcal{O}(\tau)$. Using this result and (3.14), we obtain $\|\widehat{dX}_J(\tau)\| = \mathcal{O}(\tau)$, which contradicts with the assumption $\tau = o(\|\widehat{dX}_J(\tau)\|)$. Therefore, $\|\widehat{dX}_J(\tau)\| = \mathcal{O}(\tau)$ holds. The proof of $\|\widehat{dY}_J(\tau)\| = \mathcal{O}(\tau)$ immediately follows from Lemma 3.4. $\square$

We are now in a position to state the main result of this section, which establishes the local quadratic convergence of Algorithm 2.1'.

THEOREM 3.6. *Assume that Hypothesis 2.1 and Condition 5.1 of [3] hold. If $\zeta \geq 1/\sqrt{n}$, Theorem 5.1 (local convergence theorem) of [3] holds for Algorithm 2.1'.*

Proof. Since (dX_c^k, dY_c^k) satisfies (2.2), it implies that (dX_c^k, dY_c^k) also satisfies the system (3.2) with $\tau = \theta^{k+1}\mu^0$. We also know that $(X_c^k, Y_c^k) \in \tilde{\mathcal{N}}(\gamma, \tau)$. Hence, in view of Lemmas 3.3 and 3.5, we have

$$\begin{aligned}\widehat{dX}_c^k &= \begin{pmatrix} \mathcal{O}(1) & \mathcal{O}(\theta^{k+1}) \\ \mathcal{O}(\theta^{k+1}) & \mathcal{O}(\theta^{k+1}) \end{pmatrix}, \\ \widehat{dY}_c^k &= \begin{pmatrix} \mathcal{O}(\theta^{k+1}) & \mathcal{O}(\theta^{k+1}) \\ \mathcal{O}(\theta^{k+1}) & \mathcal{O}(1) \end{pmatrix}.\end{aligned}$$

It implies that

$$\delta_c^k = \frac{\|dX_c^k dY_c^k\|_F}{\theta^{k+1}\mu^0} = \frac{\|\widehat{dX}_c^k \widehat{dY}_c^k\|_F}{\theta^{k+1}\mu^0} = \mathcal{O}(1).$$

The remaining part of the proof is based on similar arguments as the ones used in the proof of Theorem 5.1 of [3]. $\square$

REFERENCES

- [1] A. J. HOFFMAN, *On approximate solutions of systems of linear inequalities*, J. Research Nat. Bur. Standards, 49 (1952), pp. 263–265.
- [2] M. KOJIMA, *private communication*, 2003.
- [3] M. KOJIMA, M. SHIDA, AND S. SHINDOH, *A predictor-corrector interior-point algorithm for the semidefinite linear complementarity problem using the Alizadeh–Haeberly–Overton search direction*, SIAM J. Optim., 9 (1999), pp. 444–465.
- [4] R. D. C. MONTEIRO AND T. TSUCHIYA, *Polynomial convergence of a new family of primal-dual algorithms for semidefinite programming*, SIAM J. Optim., 9 (1999), pp. 551–577.
- [5] S. WRIGHT, *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia, 1997.

ON THE STABILITY OF THE EXTREME POINT SET IN LINEAR OPTIMIZATION*

MIGUEL A. GOBERNA[†], MERCEDES LARRIQUETA[‡], AND VIRGINIA VERA DE SERIO[§]

Abstract. This paper analyzes the stability properties of the set of extreme points of a closed convex set described by means of a given linear inequality system σ . We assume that all the coefficients of σ can be arbitrarily perturbed maintaining the (possibly infinite) index set as well as the (finite) dimension of the space of variables, and we measure the size of these perturbations by means of the pseudometric of the uniform convergence. The paper characterizes the nonemptiness of the extreme point set under sufficiently small perturbations and the Berge lower semicontinuity of the extreme point set mapping at σ . It also gives necessary as well as sufficient conditions for the closedness and for the Berge upper semicontinuity at σ which provide characterizations of these properties for finite systems.

Key words. extreme point set, feasible set mapping, linear programming, stability

AMS subject classifications. 49K40, 90C34, 15A39, 90C05

DOI. 10.1137/040607927

1. Introduction. The extreme point set plays a crucial role in convex analysis and optimization. In fact, any compact convex set is the convex hull of its extreme point set; simplex-like methods for ordinary and semi-infinite linear programming problems provide sequences contained in such a set, which turn out to be a useful tool in the calculation of the efficient set in multiobjective linear optimization. In all these optimization problems a certain closed convex set $F \subset \mathbb{R}^n$ is described by means of a given system $\sigma = \{a'_t x \geq b_t, t \in T\}$, where T is an arbitrary set and the coefficients are given by two functions $a : T \rightarrow \mathbb{R}^n$ and $b : T \rightarrow \mathbb{R}$. We denote by B and E the boundary and the extreme point set of the solution set F , respectively. In many applications the coefficients of the nominal system σ are approximated by subroutines in the computing process. Stability theory provides a suitable conceptual framework for analyzing the continuity properties of the set-valued mappings associating to each perturbed system its corresponding feasible set, its boundary set, and its extreme point set.

The index set T is not required to be endowed with any topological structure. Each perturbation of the nominal system σ provides an element of the parameter space (only depending on T and n),

$$\Theta = \left\{ \begin{pmatrix} c \\ d \end{pmatrix} \mid c : T \rightarrow \mathbb{R}^n, d : T \rightarrow \mathbb{R} \right\},$$

*Received by the editors May 10, 2004; accepted for publication (in revised form) December 13, 2004; published electronically August 3, 2005. This work was partially supported by the MCYT of Spain and FEDER of UE, grant BFM2002-04114-C02-01 and SECYT-UNCuyo of Argentina, grant 987/02-R.

<http://www.siam.org/journals/siopt/15-4/60792.html>

[†]Departamento de Estadística e Investigación Operativa, Universidad de Alicante, 03071 Alicante, Spain (mgoberna@ua.es).

[‡]Facultad de Ingeniería, Universidad Nacional de Cuyo, 5500 Mendoza, Argentina (mlarriqueta@fcemail.uncu.edu.ar).

[§]Facultad de Ciencias Económicas, Universidad Nacional de Cuyo, 5500 Mendoza, Argentina (vvera@fcemail.uncu.edu.ar).

where we identify $\begin{pmatrix} c \\ d \end{pmatrix} \in \Theta$ (the product space $(\mathbb{R}^{n+1})^T$) with the system $\sigma_1 = \{c'_t x \geq d_t, t \in T\}$, and consequently we write $\sigma_1 \in \Theta$ (all the objects corresponding to σ_1 will be distinguished with the same subscript: F_1, B_1, E_1 , etc.). The subset of consistent systems is referred to as Θ_c and we denote by Ξ the family of systems such that the corresponding feasible set has at least one extreme point, i.e., $\Xi = \{\sigma_1 \in \Theta \mid E_1 \neq \emptyset\}$.

Throughout the paper we make the following assumption.

Assumption 1.1. The nominal system σ is consistent ($\sigma \in \Theta_c$) and nontrivial ($F \neq \mathbb{R}^n$).

Consequently, $\sigma \neq \theta := \{0'_n x \geq 0, t \in T\}$ (the null element in Θ). Moreover, if $\sigma_1 = \{c'_t x \geq d_t, t \in T\}$ is the result of perturbing σ , the size of this perturbation is measured by means of the uniform pseudometric, i.e.,

$$d(\sigma_1, \sigma) = \sup_{t \in T} \left\| \begin{pmatrix} c_t \\ d_t \end{pmatrix} - \begin{pmatrix} a_t \\ b_t \end{pmatrix} \right\|_{\infty}.$$

The feasible set mapping $\mathcal{F} : \Theta \rightrightarrows \mathbb{R}^n$ is the point-to-set mapping that associates with each system σ_1 its solution set F_1 . Similarly, the boundary set mapping $\mathcal{B} : \Theta \rightrightarrows \mathbb{R}^n$ associates to each system σ_1 the boundary of F_1 , denoted by B_1 . The main purpose of this paper is to characterize the topological interior of Ξ as well as some stability properties of the extreme point set mapping $\mathcal{E} : \Theta \rightrightarrows \mathbb{R}^n$ associating to σ_1 the set E_1 of extreme points of F_1 .

We proceed by recalling the basic concepts on stability of set-valued mappings involved in this paper and the main relationships among them. More details can be found, e.g., in [1] and [12] (where the reader can find the inner and outer semicontinuity theory).

A mapping $\mathcal{A} : \Theta \rightrightarrows \mathbb{R}^n$ is closed (or outer semicontinuous) at σ if for any $\bar{x} \in \mathbb{R}^n$ and any two sequences, $\{\sigma_r\} \subset \Theta$ and $\{x^r\} \subset \mathbb{R}^n$ such that $\lim_r \sigma_r = \sigma$, $x^r \in \mathcal{A}(\sigma_r)$, $r = 1, 2, \dots$, and $\lim_r x^r = \bar{x}$, one gets $\bar{x} \in \mathcal{A}(\sigma)$.

A mapping $\mathcal{A} : \Theta \rightrightarrows \mathbb{R}^n$ is lower semicontinuous in Berge sense (or inner semicontinuous), in short B-lsc, at $\sigma \in \Theta$ if for each open set W such that $W \cap \mathcal{A}(\sigma) \neq \emptyset$ there exists an open set V , $\sigma \in V \subset \Theta$, such that $W \cap \mathcal{A}(\sigma_1) \neq \emptyset$ for every $\sigma_1 \in V$. One can prove that $\mathcal{A}$ is B-lsc at σ if and only if there exists $r_0 \in \mathbb{N}$ such that $\mathcal{A}(\sigma) = \lim_{r \geq r_0} \mathcal{A}(\sigma_r)$, i.e., for any sequence $\{\sigma_r\}$ that converges to σ , $\mathcal{A}(\sigma)$ is the set of limit points, as well as the set of cluster points, of all the possible sequences $\{x^r\}$, such that $x^r \in \mathcal{A}(\sigma_r)$, $r \geq r_0$. $\mathcal{A}$ is lower semicontinuous in Hausdorff sense (H-lsc) at $\sigma \in \Theta$ if for each real number $\varepsilon > 0$ there exists an open set V , $\sigma \in V \subset \Theta$, such that $\mathcal{A}(\sigma) \subset \mathcal{A}(\sigma_1) + B(0_n, \varepsilon)$ for any $\sigma_1 \in V$, where $B(0_n, \varepsilon)$ denotes the open ball centered at 0_n with radius $\varepsilon > 0$. It is known that, if $\mathcal{A}$ is H-lsc at σ , then $\mathcal{A}$ is B-lsc at σ ; the converse statement holds if $\mathcal{A}(\sigma)$ is bounded.

A mapping $\mathcal{A} : \Theta \rightrightarrows \mathbb{R}^n$ is upper semicontinuous in Berge sense (B-usc) at $\sigma \in \Theta$ if for each open set W such that $\mathcal{A}(\sigma) \subset W$ there exists an open set V , $\sigma \in V \subset \Theta$, such that $\mathcal{A}(\sigma_1) \subset W$ for every $\sigma_1 \in V$. If such neighborhoods V exist at least for those open sets W of the form $\mathcal{A}(\sigma) + B(0_n, \varepsilon)$, then $\mathcal{A}$ is said to be upper semicontinuous in Hausdorff sense (H-usc) at $\sigma \in \Theta$. Obviously, if $\mathcal{A}$ is B-usc at σ , then $\mathcal{A}$ is H-usc at σ ; the converse statement holds if $\mathcal{A}(\sigma)$ is compact. It is also known that, if $\mathcal{A}$ is H-usc (B-usc) at σ and $\mathcal{A}(\sigma)$ is closed, then $\mathcal{A}$ is closed at σ . Moreover, if $\mathcal{A}$ is closed and locally bounded at σ (i.e., there exists an open set V , $\sigma \in V \subset \Theta$, such that the set $\cup\{\mathcal{A}(\sigma_1) \mid \sigma_1 \in V\}$ is bounded), then $\mathcal{A}$ is B-usc at σ . $\mathcal{A} : \Theta \rightrightarrows \mathbb{R}^n$ is

Berge (Hausdorff) continuous if it is B-lsc and B-usc (H-lsc and H-usc, respectively) at $\sigma \in \Theta$. Since H-lsc and B-usc are usually considered very strong properties, Bouligand continuity (B-lsc and closedness) appears to be a more convenient continuity concept (see, e.g., [12]).

As in most of the papers dealing with the stability analysis of systems and optimization problems, we shall analyze the closedness and the Berge semicontinuity of $\mathcal{E}$ at the nominal system. None of the previous works on the subject ([13], [14], [10], and [3]) considers Berge semicontinuity or infinite systems, and only one of them [3] allows perturbations of all the data (and not only of the right-hand side). In more detail, [3] considers the closedness and the Hausdorff semicontinuity of the restriction of $\mathcal{E}$ to Θ_c , that we denote as $\mathcal{E}_c : \Theta_c \rightrightarrows \mathbb{R}^n$. Obviously, $\mathcal{E}_c$ is closed (B-usc, H-usc) at σ if and only if $\mathcal{E}$ satisfies the same property. Furthermore, $\mathcal{E}_c$ inherits the B-lsc (H-lsc) property from $\mathcal{E}$ and we shall prove that the converse is true except when F is a singleton.

The paper is organized as follows. Section 2 introduces some additional notation and recalls some results on linear systems and stability properties of the mappings $\mathcal{F}$ and $\mathcal{B}$ which will be used later. Section 3 deals with the subspace Ξ , which is empty when $|T| < n$, and so $\mathcal{E}$ has empty image everywhere (trivial case). Hence we assume $|T| \geq n$ and we characterize $\text{int } \Xi$ and prove that this set almost fills Θ_c (in a topological sense). Three cases are possible for the nominal system σ according to the cardinalities of E and F . Section 4 analyzes the two simplest cases: $E = \emptyset$ (first case, where F is unbounded) and $|F| = 1$ (second case, where $E \neq \emptyset$). Finally, section 5 considers the third and most important case: $|F| > 1$ and $E \neq \emptyset$.

2. Preliminaries. Let us introduce the necessary notation. Given a nonempty subset X of a certain real vector space, we denote by $\text{span } X$, $\text{cone } X$, $\text{conv } X$, and $\text{lin } X$ the linear subspace spanned by X , the convex cone spanned by X (including the null vector), the convex hull of X , and the lineality space of the convex cone X , respectively. $0^+ X$ and $\dim X$ denote the recession cone and the dimension of a convex set X . When the vector space is equipped with a certain topology, we denote by $\text{cl } X$, $\text{int } X$, $\text{rint } X$, and $\text{bd } X$ the closure, the interior, the relative interior, and the boundary of X , respectively. If the topology is associated with a certain pseudometric or metric, this will be represented by d . Finally, e_i denotes the i th element of the canonical basis, $e = (1, \dots, 1)$, and $B(x; \varepsilon)$ is the open ball centered at x with radius ε ; $\lim_r$ stands for $\lim_{r \rightarrow \infty}$.

The characteristic cone of σ , $K = \text{cone}\left\{\begin{pmatrix} a_t \\ b_t \end{pmatrix}, t \in T; \begin{pmatrix} 0_n \\ -1 \end{pmatrix}\right\} \subset \mathbb{R}^{n+1}$, captures the most relevant information on F . The nonhomogeneous Farkas lemma establishes that a given inequality $a'x \geq b$ is a consequence of σ if and only if $\begin{pmatrix} a \\ b \end{pmatrix} \in \text{cl } K$, F is bounded if and only if $\begin{pmatrix} 0_n \\ -1 \end{pmatrix} \in \text{int cl } K$, $\dim F = n - \dim \text{lin cl } K$, and so $|F| = 1$ if and only if $\dim \text{lin cl } K = n$ (see [6]). Operative tests for $F = \{\bar{x}\}$ can be found in [9]. On the other hand, it can be shown that F is strictly convex (i.e., B does not contain segments) if and only if $E = B$ if and only if there exists a unique nonvertical supporting hyperplane to $\text{cl } K$ at any point of intersection of $\text{bd } K$ with a vertical line different from $x = 0_n$.

$\bar{x} \in \mathbb{R}^n$ is a strong Slater (SS) point for σ if there exists a positive scalar ε such that $a'_t \bar{x} \geq b_t + \varepsilon$ for all $t \in T$. If $\bar{x}$ is an SS point of σ and $\{a_t, t \in T\}$ is bounded, then $\bar{x} \in \text{int } F$ [5]. It is known that the stability of $\mathcal{F}$ is basically the existence of an SS point (the SS condition), which holds if and only if the following linear optimization

problem has a negative value:

$$\min x_{n+1} \quad \text{s.t.} \quad a'_t x + x_{n+1} \geq b_t, \quad t \in T.$$

If $|T| < \infty$, the SS condition is equivalent to the classical Slater condition, i.e., the linear version of the Mangasarian–Fromovitz constraint qualification [11].

In this paper we show that the stability of $\mathcal{E}$ is basically nondegeneracy of σ , a term inspired in the concept of a nondegenerate extreme point of a polyhedron: in the case $|T| < \infty$, $x \in E$ if and only if $\dim \text{span}\{a_t, t \in T(x)\} = n$, where $T(x) = \{t \in T \mid a'_t x = b_t\}$ is the set of active indices at x , and it is degenerate if $|T(x)| > n$ (there is an excess of active constraints at x). In the same vein, we present the following definition.

DEFINITION 2.1. σ is nondegenerate if $|T(x)| < n$ for all $x \in F \setminus E$.

If σ is such that F is not strictly convex, then in the previous definition the condition “ $x \in F \setminus E$ ” can be replaced by “ $x \in B \setminus E$.” Observe that replacing each inequality $a'_t x \geq b_t$ by the countable system $a'_t x \geq b_t - r^{-1}$, $r \in \mathbb{N}$, we get a nondegenerate representation of F which is also an element of Θ_c if T is infinite. In other words, degeneracy is the effect of a bad representation of F (from the point of view of the stability of $\mathcal{E}$).

In general, there is no connection between degeneracy and existence of redundant constraints, at least for $n \geq 4$, as the following example (used in [3, Example 5.2] with a different purpose) shows.

Example 2.2. Consider the following system in $\mathbb{R}^4$:

$$\sigma = \left\{ \begin{array}{l} -x_1 - x_2 \geq 0, \quad -x_2 - x_3 \geq 0, \quad x_1 - x_2 \geq 0 \\ -x_2 + x_3 \geq 0, \quad x_2 \geq -1, \quad x_4 \geq -1, \quad -x_4 \geq 0 \end{array} \right\}.$$

It can be shown that $(0, -\frac{1}{2}, 0, -\frac{1}{2})$ is a Slater point, so that $\dim F = 4$. On the other hand, for $\bar{x} = (0, 0, 0, -\frac{1}{2})$ we have $T(\bar{x}) = \{1, 2, 3, 4\}$ and $\dim \text{span}\{a_t, t \in T(\bar{x})\} = 3$. Thus $\bar{x} \in F \setminus E$ and σ turns out to be degenerate. Nevertheless, it can be realized that σ does not contain redundant constraints. Observe that the coincidence of n different facets at $\bar{x}$ does not entail that $\bar{x} \in E$.

Consider now the following list of conditions on σ :

- (A) F is strictly convex.
- (B) $\{a_t, t \in T(x)\}$ is linearly independent for all $x \in F \setminus E$ such that $T(x) \neq \emptyset$.
- (C) σ is nondegenerate.
- (D) $|T(\frac{x+y}{2})| < n$ for all $x, y \in E$ such that $x \neq y$.

It can be easily seen that (A) $\Rightarrow$ (B) $\Rightarrow$ (C) $\Rightarrow$ (D) whenever σ does not contain the trivial inequality. In particular, if $|T| < \infty$ and F is bounded (i.e., F is a polytope), then (B) $\Leftrightarrow$ (C). If $|T| < \infty$, then (D) can be checked once E has been calculated. We focus on (C), which turns out to be closely related with the B-usc and closedness properties of $\mathcal{E}$ whereas the SS condition still determines its B-lsc behavior.

For the case $|T| < \infty$, conditions (B), (C), and (D) were already considered in [3], where (B) is emphasized. Concerning the stability in Hausdorff sense, [3, Theorem 4.1] establishes that $\mathcal{E}_c$ is H-continuous at σ if and only if σ satisfies Slater condition, i.e., $\mathcal{E}$ is B-lsc at σ . We show that the nondegeneracy of σ is equivalent to the closedness and to Bouligand continuity of $\mathcal{E}$ at σ (Theorem 5.9). Moreover, if F is bounded, then the nondegeneracy of σ is also equivalent to the B-usc and B-continuity property of $\mathcal{E}$ at σ . By [3, Theorems 5.1 and 5.2], the following properties are also equivalent: H-usc and H-continuity of $\mathcal{E}_c$ at σ . If F is unbounded, [3, Theorem 5.3] establishes

that $\mathcal{E}_c$ is H-continuous at σ if and only if σ and its associated homogeneous system σ_H satisfy condition (B). (See also Remark 4.7.)

The next two results provide computable stability criteria for $\mathcal{F}$ and $\mathcal{B}$, and the proofs can be found in [7, Theorem 3.1] and [5, Theorem 2], respectively.

THEOREM 2.3. *The following statements are equivalent:*

- (i) $\mathcal{F}$ is B-lsc at σ .
- (ii) $\sigma \in \text{int } \Theta_c$.
- (iii) σ satisfies the SS condition.
- (iv) $0_{n+1} \notin \text{cl conv}\{(a_i, b_i), t \in T\}$.

Obviously, $\mathcal{F}$ is closed everywhere. Different conditions for the B-usc of $\mathcal{F}$ at $\sigma \in \Theta_c$ in terms of the coefficients can be found in [8] and [2], but they are not easy to be checked, with the exception of the boundedness of F , which is sufficient. Consequently, $\mathcal{F}$ is locally bounded at σ if and only if F is bounded. Obviously, if $\mathcal{F}$ is locally bounded at σ , then $\mathcal{B}$ and $\mathcal{E}$ enjoy the same property.

THEOREM 2.4. *The following statements hold:*

- (i) $\mathcal{B}$ is B-lsc at σ if and only if $\mathcal{F}$ is B-lsc at σ .
- (ii) If $\mathcal{B}$ is B-lsc at σ , then $\mathcal{B}$ is closed at σ .

In contrast with $\mathcal{F}$, $\mathcal{E}$ is not necessarily closed on Θ . Even more, the images of $\mathcal{E}$ are not necessarily closed if T is infinite.

PROPOSITION 2.5. (i) *If T is infinite, then $\mathcal{E}$ has always closed images if and only if $n \leq 2$.*

(ii) *If $|T| \geq n$, then $\mathcal{E}$ is neither closed nor B-usc at a certain $\sigma \in \Theta$ such that $\emptyset \neq F \neq \mathbb{R}^n$. Moreover, if $|T| \geq n+1$, then there exists $\sigma \in \Theta$ with $\emptyset \neq F \neq \mathbb{R}^n$ such that $\mathcal{E}$ is not B-lsc at σ .*

(iii) *If $|T| \geq n$, then $\{\sigma_1 \in \Theta \mid E_1 \neq \emptyset\}$ contains an open and dense subset of Θ_c .*

Proof. (i) For $n = 1$, the solution sets are intervals and so the extreme point sets have at most two elements. Suppose now that for $n = 2$, there is some $\sigma_1 \in \Theta$ such that E_1 is not closed. Let $\{x^r\} \subset E_1$ be a sequence of all different points with $\lim_r x^r = \bar{x} \notin E_1$ (obviously, $\bar{x} \in F_1$). Let $y^1 \neq y^2$ in F_1 such that $\bar{x} = \frac{1}{2}y^1 + \frac{1}{2}y^2$. The line $L = \{(1-\mu)y^1 + \mu y^2 \mid \mu \in \mathbb{R}\}$ determines two open half-planes. L cannot contain more than three points of $\{x^r\}$. Hence, at least one of the open half-planes, say H , contains a subsequence which we can consider to be $\{x^r\}$ itself. Since

$$\bar{x} = \frac{1}{4}y^1 + \frac{1}{4}y^2 + \frac{1}{4}x^1 + \frac{1}{4}(2\bar{x} - x^1),$$

one gets $\bar{x} \in \text{int conv}\{y^1, y^2, x^1, 2\bar{x} - x^1\}$ (by [4, Theorem 2.3]), so that x^r belongs to $\text{int conv}\{y^1, y^2, x^1, 2\bar{x} - x^1\}$ for r large enough. Since we may assume that $x^r \neq x^1$ for $r > 1$, the fact that $x^r \in H$ gives $x^r \in (\text{conv}\{y^1, y^2, x^1\}) \setminus \{y^1, y^2, x^1\}$ and hence $x^r \notin E_1$. This contradicts the choice of x^r as an extreme point of F_1 .

Now we will show that for $n = 3$, there is some $\sigma_2 \in \Theta$ such that E_2 is not closed. Assume without loss of generality that $|T| = \aleph_0$ (by adding trivial inequalities $0'_n x \geq 0$ if necessary) or, even more, that $T = \mathbb{N}$. We shall define a closed convex set F_2 in $\mathbb{R}^3$ that admits a minimal tight representation σ_2 , whose extreme point set E_2 is not closed. To do this consider

$$E_2 = \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}; \begin{pmatrix} (2t-1)/t^2 \\ 0 \\ (t-1)/t \end{pmatrix}, t = 1, 2, \dots \right\}$$

and take $F_2 = \text{conv } E_2$. Now, E_2 is not closed because

$$\lim_{t \rightarrow \infty} \begin{pmatrix} (2t-1)/t^2 \\ 0 \\ (t-1)/t \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

and

$$\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}.$$

Finally, for $n > 3$, consider the countable system in $\mathbb{R}^n$

$$\sigma_3 = \sigma_2 \cup \{x_4 = \cdots = x_n = 0\},$$

with solution and extreme point sets

$$F_3 = \left\{ \begin{pmatrix} x \\ 0_{n-3} \end{pmatrix}, x \in F_2 \right\} \text{ and } E_3 = \left\{ \begin{pmatrix} x \\ 0_{n-3} \end{pmatrix}, x \in E_2 \right\},$$

E_3 being nonclosed either.

(ii) Let $\{t_1, \dots, t_n\} \subset T$ such that $t_i \neq t_j$ for all $i \neq j$ and let $\sigma = \{a'_t x \geq b_t, t \in T\}$ with $\begin{pmatrix} a_t \\ b_t \end{pmatrix} = \begin{pmatrix} e_1 \\ 1 \end{pmatrix}$ for $t = t_1$ and 0_{n+1} otherwise. Obviously, $E = \emptyset$. Consider now the sequence $\{\sigma_r\} \subset \Theta$, where σ_r is obtained from σ by replacing the t_i -inequality by $r^{-1}x_i \geq r^{-1}$, $i = 2, \dots, n$. Since $\lim_r \sigma_r = \sigma$ and $E_r = \{e\}$ for all r , $\mathcal{E}$ is neither closed nor B-usc at σ .

Now let $\{t_1, \dots, t_{n+1}\} \subset T$ such that $t_i \neq t_j$ for all $i \neq j$ and let σ be such that

$$\begin{pmatrix} a_t \\ b_t \end{pmatrix} = \begin{cases} (e_i, 0)', & t = t_i, i = 1, \dots, n, \\ (-e, 0)', & t = t_{n+1}, \\ (0_n, -1)' & \text{otherwise.} \end{cases}$$

Since $\sigma \notin \text{int } \Theta_c$ (the unique solution of σ is not an SS point), $\mathcal{E}$ cannot be B-lsc at σ .

(iii) The set $\{\sigma_1 \in \text{int } \Theta_c \mid E_1 \neq \emptyset\}$ is open because given any $\sigma_1 = \{c'_t x \geq d_t, t \in T\}$, the fact that $E_1 \neq \emptyset$ is equivalent to asserting that $\{c_t, t \in T\}$ contains a basis of $\mathbb{R}^n$, and this property is preserved by sufficiently small perturbations of the coefficients. Now, denote by $\sigma = \{a'_t x \geq b_t, t \in T\}$ a generic element of Θ_c (with maybe $F = \mathbb{R}^n$) such that $E = \emptyset$. Take a basis of $\text{span}\{a_t, t \in T\}$, say $\{a_{t_i}, i = 1, \dots, m\}$ ($m < n$), and vectors with small norms $u_{m+1}, \dots, u_n$ in $\mathbb{R}^n$ such that $\{a_{t_i}, i = 1, \dots, m; u_{m+1}, \dots, u_n\}$ is a basis. Let $\varepsilon > 0$ be given. Defining $c_{t_i} = a_{t_i} + \varepsilon u_i$, $i = m+1, \dots, n$, and $c_t = a_t$ otherwise, and $d_t = b_t - \varepsilon$ for all $t \in T$, the system $\sigma_1 = \{c'_t x \geq d_t, t \in T\}$ satisfies the SS condition (i.e., $\sigma_1 \in \text{int } \Theta_c$), $\{c_{t_i}, i = 1, \dots, n\}$ is a basis of $\mathbb{R}^n$ (i.e., $E_1 \neq \emptyset$) and $d(\sigma_1, \sigma) = \varepsilon$. $\square$

Notice that statement (iii) is not true if $|T| = n$ as a consequence of Theorem 2.3 and the forthcoming Proposition 4.2 and Theorem 5.2.

Let us observe that the closedness of E does not imply the closedness of $\mathcal{E}$ at σ . Even more, [5, Example 2] shows that any closed convex full dimensional set admits a linear representation where the extreme point set mapping fails to be closed.

3. The interior of Ξ . The following result characterizes $\text{int } \Xi$ and shows that if $|T| \geq n$, then $\text{int } \Xi$ almost fills Θ_c in the sense that any consistent system is the limit of systems for which sufficiently small perturbations of the data preserve the nonemptiness of the extreme point set.

PROPOSITION 3.1. (i) $\Xi \neq \emptyset$ if and only if $\Xi \cap \text{int } \Theta_c \neq \emptyset$ if and only if $|T| \geq n$.

(ii) Ξ is open in Θ_c . Hence $\text{int } \Xi = \Xi \cap \text{int } \Theta_c$.

(iii) $\text{int } \Xi$ is dense in Θ_c if and only if $|T| \geq n$.

Proof. (i) Assume that $|T| \geq n$. We shall prove that $\Xi \cap \text{int } \Theta_c \neq \emptyset$, so that $\Xi \neq \emptyset$ too. Take $\{t_1, \dots, t_n\} \subset T$ with $t_i \neq t_j$ if $i \neq j$, and define

$$\begin{pmatrix} c_t \\ d_t \end{pmatrix} = \begin{cases} \begin{pmatrix} e_i \\ 0 \end{pmatrix}, & t = t_i, i = 1, \dots, n, \\ \begin{pmatrix} 0_n \\ -1 \end{pmatrix}, & t \in T \setminus \{t_1, \dots, t_n\}; \end{cases}$$

hence for $\sigma_1 = \{c'_t x \geq d_t, t \in T\}$, one has $F_1 = \mathbb{R}_+^n$ and $E_1 = \{0_n\}$. Moreover, e is an SS point for σ_1 and $\sigma_1 \in \Xi \cap \text{int } \Theta_c$ according to Theorem 2.3.

Alternatively, if $|T| < n$, then $\dim \text{span}\{c_t, t \in T\} < n$, i.e., all the solution sets of systems in Θ_c contain lines and so $\Xi = \emptyset$.

(ii) We can assume that $|T| \geq n$. Let $\sigma_1 = \{c'_t x \geq d_t, t \in T\} \in \Xi$. Since F_1 does not contain lines, $\dim \text{span}\{c_t, t \in T\} = n$ and there exists a set $S \subset T$ such that $\{c_t, t \in S\}$ is a basis of $\mathbb{R}^n$. By continuity of the determinant with respect to the entries, there exists $\delta > 0$ such that $\{u_t, t \in S\}$ is a basis of $\mathbb{R}^n$ if $\|u_t - c_t\|_\infty < \delta$ for all $t \in S$. Then $\sigma_2 = \{u'_t x \geq v_t, t \in T\} \in \Theta_c$ and $d(\sigma_1, \sigma_2) < \delta$ yield $\dim \text{span}\{u_t, t \in T\} = n$ and so $E_2 \neq \emptyset$, i.e., $\sigma_2 \in \Xi$. Therefore Ξ is open in Θ_c .

(iii) Given $\sigma_0 = \{c'_t x \geq d_t, t \in T\} \in \Theta_c$ we can consider $\sigma_r := \{c'_t x \geq d_t - r^{-1}, t \in T\} \in \text{int } \Theta_c$, $r \in \mathbb{N}$, so we have $\lim_r \sigma_r = \sigma_0$. Therefore it is enough to prove the density of $\text{int } \Xi$ in $\text{int } \Theta_c$.

Assume that $\sigma_0 \in (\text{int } \Theta_c) \setminus \Xi$ and let $\varepsilon > 0$. Take $0 < \varepsilon_1 \leq \varepsilon$, such that $\sigma_1 \in \text{int } \Theta_c$ if $d(\sigma_1, \sigma_0) < \varepsilon_1$.

Let $\{c_t, t \in T_0\}$ be a basis of $\text{span}\{c_t, t \in T\}$. Then $q_0 := |T_0| < n$ (otherwise $\sigma_0 \in \Xi$). Take $t_1 \in T \setminus T_0$, and let $w^1 \in \mathbb{R}^n$ such that $w^1 \notin \text{span}\{c_t, t \in T_0\}$ and $\|w^1\|_\infty < \frac{1}{n}$. Since $c_{t_1} \in \text{span}\{c_t, t \in T_0\}$, $c_{t_1} + \varepsilon_1 w^1 \notin \text{span}\{c_t, t \in T_0\}$. Let $\sigma_1 = \{u'_t x \geq v_t, t \in T\}$, where

$$\begin{pmatrix} u_t \\ v_t \end{pmatrix} = \begin{cases} \begin{pmatrix} c_t + \varepsilon_1 w^1 \\ d_t \end{pmatrix}, & t = t_1, \\ \begin{pmatrix} c_t \\ d_t \end{pmatrix}, & t \in T \setminus \{t_1\}. \end{cases}$$

Obviously, $d(\sigma_1, \sigma_0) < \frac{\varepsilon_1}{n}$ and $\{u_t, t \in T_1\}$ is a basis of $\text{span}\{u_t, t \in T\}$ if $T_1 = T_0 \cup \{t_1\}$. If $q_1 := |T_1| = n$, then $\sigma_1 \in \Xi \cap \text{int } \Theta_c$ and we have finished. Otherwise we can construct inductively a sequence of systems $\sigma_r \in \Theta$, $r = 1, 2, \dots$, with corresponding sequence of dimensions q_r for the linear spans of the left-hand side vectors, such that $d(\sigma_{r+1}, \sigma_r) < \frac{\varepsilon_1}{n}$ and $q_r < q_{r+1}$, $r = 1, 2, \dots$. After at most n iterations we obtain a system σ_p such that $q_p = n$ and $d(\sigma_p, \sigma_0) < \varepsilon_1$. Then $\sigma_p \in \Xi \cap \text{int } \Theta_c$, with $d(\sigma_p, \sigma_0) < \varepsilon$ (observe that the existence of $t_r \in T \setminus T_{r-1}$ is guaranteed by $q_{r-1} < n \leq |T|$). $\square$

According to part (i) of the last proposition, the interesting case is when $|T| \geq n$, so we assume from now on the following assumption.

Assumption 3.2. The nominal system σ satisfies $|T| \geq n$, together with $\emptyset \neq F \neq \mathbb{R}^n$.

4. Two simple cases: $E = \emptyset$ and $|F| = 1$. First, we will consider the case $E = \emptyset$. We will show that $\sigma \in \text{bd } \Xi$, $\mathcal{E}$ is B-lsc at σ , and $\mathcal{E}$ is not B-usc at σ under Assumption 3.2. However, regarding the closedness of $\mathcal{E}$, it is possible for $\mathcal{E}$ to be closed or not closed at σ even for $|T| < \infty$, as the systems in Example 4.1 show. Also, Proposition 4.2 gives a necessary condition, (iv), and a sufficient one, (v), for $\mathcal{E}$ being closed at a given system σ .

Example 4.1. Let $n = 2$ and $\sigma = \{x_1 \geq -1, -x_1 \geq -1\}$. Then, $F = [-1, 1] \times \mathbb{R}$, $E = \emptyset$, and σ is nondegenerate. Consider any pair of sequences $\{\sigma_r\} \subset \Theta$ and $\{x^r\}$ such that $x^r \in E_r$ and $\lim_r \sigma_r = \sigma$. We will show that $\{x^r\}$ cannot converge. Writing

$$\sigma_r = \{\alpha_r x_1 + \beta_r x_2 \geq \gamma_r, \delta_r x_1 + \zeta_r x_2 \geq \eta_r\},$$

since $x^r \in E_r$, $r = 1, 2, \dots$, we have $x_2^r = \frac{\alpha_r \eta_r - \delta_r \gamma_r}{\alpha_r \zeta_r - \beta_r \delta_r}$. The fact that $\{\sigma_r\}$ converges to σ yields $\lim_r (\alpha_r \zeta_r - \beta_r \delta_r) = 0$ and $\lim_r (\alpha_r \eta_r - \delta_r \gamma_r) = -2$. In this way, $\{x_2^r\}$ does not converge and nor does $\{x^r\}$. Therefore $\mathcal{E}$ is closed at σ .

Nevertheless $\mathcal{E}$ is not closed at $\tilde{\sigma} = \{x_1 \geq 0; x_1 \geq 0\}$; indeed the sequences $\{\sigma_r\}$, $\{x^r\}$, with $\sigma_r = \{x_1 \geq 0; x_1 + \frac{1}{r}x_2 \geq 0\}$ and $x^r = 0_2 \in E_r$, $r = 1, 2, \dots$, verify that $\lim_r \sigma_r = \tilde{\sigma}$ and $\lim_r x^r = 0_2 \notin \tilde{E} = \emptyset$.

PROPOSITION 4.2. *If $E = \emptyset$, then the following statements hold:*

- (i) $\sigma \in \text{bd } \Xi$.
- (ii) $\mathcal{E}$ is B-lsc at σ .
- (iii) $\mathcal{E}$ is not B-usc at σ (and so $\mathcal{E}$ is not B-continuous).
- (iv) If $\mathcal{E}$ is closed at σ , then σ is nondegenerate.
- (v) If σ is nondegenerate and $|T| < \infty$, then $\mathcal{E}$ is closed (and so Bouligand-continuous) at σ .

Proof. (i) Since $\sigma \notin \Xi$ and $\text{int } \Xi$ is dense in Θ_c (by Proposition 3.1), $\sigma \in \text{bd } \Xi$.

(ii) It is trivial.

(iii) Let $W = E = \emptyset$. By (i) there exists a sequence $\{\sigma_r\} \subset \Theta$ such that $\lim_r \sigma_r = \sigma$ and $E_r \neq \emptyset$ (i.e., E_r is not included in W) for all $r \in \mathbb{N}$. Thus $\mathcal{E}$ is not B-usc at σ .

(iv) Assume that σ is degenerate, and let $\bar{x} \in F \setminus E$ such that $|T(\bar{x})| \geq n$. $\bar{x} \notin E$ implies that $k := \dim \text{span}\{a_t, t \in T(\bar{x})\} < n$. Let $\{a_{t_1}, \dots, a_{t_k}\}$ be a basis of this subspace and let $\{u_{k+1}, \dots, u_n\}$ be a basis of its orthogonal complement. Take $\{t_{k+1}, \dots, t_n\} \subset T(\bar{x}) \setminus \{t_1, \dots, t_k\}$ arbitrarily and consider the sequence $\{\sigma_r\} \subset \Theta_c$ defined as follows:

$$\sigma_r := \left\{ \left(a_{t_i} + \frac{u_i}{r} \right)' (x - \bar{x}) \geq 0, i = k+1, \dots, n; a'_i x \geq b_t, t \in T \setminus \{t_{k+1}, \dots, t_n\} \right\}.$$

We have $\bar{x} \in F_r$, with $\{t_1, \dots, t_n\} \subset T_r(\bar{x})$, the corresponding left-hand side vectors forming a basis of $\mathbb{R}^n$ (so that $\bar{x} \in E_r$) for all $r \in \mathbb{N}$. Since $\lim_r \sigma_r = \sigma$, taking $x^r := \bar{x}$ for all $r \in \mathbb{N}$, we conclude that $\mathcal{E}$ is not closed at σ .

(v) Assume that σ is nondegenerate and $|T| < \infty$. We suppose also that $\mathcal{E}$ is not closed at σ in order to get a contradiction.

Let $\{\sigma_r\} \subset \Theta$ and $\{x^r\} \subset \mathbb{R}^n$ such that $x^r \in E_r$ for all $r \in \mathbb{N}$, $\lim_r \sigma_r = \sigma$, and $\lim_r x^r = \bar{x} \notin E$. The closedness of $\mathcal{F}$ and the nondegeneracy of σ yield $\bar{x} \in F$ and $|T(\bar{x})| < n$.

For each $t \in T \setminus T(\bar{x})$, we have

$$\lim_r (a'_t x^r - b_t) = a'_t \bar{x} - b_t > 0,$$

and then there exists $r_t \in \mathbb{N}$ such that $a'_t x^r - b_t > 0$ for all $r \geq r_t$. Since $|T| < \infty$, $\bar{r} := \max\{r_t, t \in T \setminus T(\bar{x})\} \in \mathbb{N}$. Then, for every $r \geq \bar{r}$ we have $T(x^r) \subset T(\bar{x})$ and so $n = |T(x^r)| \leq |T(\bar{x})| < n$. $\square$

Remark 4.3. A shorter (but less constructive) argument for the first part of (iv) consists of observing that $\mathcal{E}$ is locally bounded and closed at σ . Then $\mathcal{E}$ is B-usc at σ in contradiction with (iii). The next example shows that the finiteness assumption in (v) is not superfluous.

Example 4.4. Let $n = 2$ and $\sigma = \{-x_2 \geq -\frac{1}{t}, t \in \mathbb{N}\}$. One has $F = \mathbb{R} \times]-\infty, 0]$ and $E = \emptyset$. Replacing in σ the inequality of index r by $r^{-1}x_1 - x_2 \geq 0$ the perturbed system σ_r satisfies $E_r = \{0_2\}$. Taking $x^r = 0_2$ for all $r \in \mathbb{N}$ we conclude that $\mathcal{E}$ is not closed at σ , even though σ is nondegenerate.

PROPOSITION 4.5. *If $|F| = 1$, then the following statements hold:*

- (i) $\sigma \in \text{int } \Xi$ if and only if σ satisfies the SS condition (so, $\sigma \in \text{bd } \Xi$ if $|T| < \infty$).
- (ii) $\mathcal{E}$ is B-lsc at σ if and only if $\sigma \in \text{int } \Theta_c$.
- (iii) $\mathcal{E}$ is B-usc at σ .
- (iv) $\mathcal{E}$ is closed at σ .

Proof. (i) The equivalence is a consequence of Proposition 3.1(ii) and Theorem 2.3. If $F = \{\bar{x}\}$ and we assume $\sigma \in \text{int } \Xi$ and $|T| < \infty$, then $\bar{x}$ is a Slater point and so $\bar{x} \in \text{int } F$, which is not possible.

(ii) The direct statement is trivial. So we assume that $\sigma \in \text{int } \Theta_c$. Let W be an open set in $\mathbb{R}^n$ such that $W \cap E \neq \emptyset$. If $F = \{\bar{x}\}$, then we can suppose that $W = B(\bar{x}; \rho)$, $\rho > 0$. Since $\mathcal{F}$ is B-usc at σ , there exists $\varepsilon_1 > 0$ such that $F_1 \subset W$ if $d(\sigma_1, \sigma) < \varepsilon_1$. Let $\varepsilon_2 > 0$, $\varepsilon_2 < \varepsilon_1$, such that $F_1 \neq \emptyset$ if $d(\sigma_1, \sigma) < \varepsilon_2$. For such a system σ_1 one has $\emptyset \neq F_1 \subset W$, and so $E_1 \subset \text{conv } E_1 = F_1 \subset W$. Hence $W \cap E_1 = E_1 \neq \emptyset$.

(iii) Let W be an open set in $\mathbb{R}^n$ such that $E \subset W$. Then $F = E \subset W$ and the conclusion follows again from the B-usc property of $\mathcal{F}$ at σ .

(iv) Since $E = F$ is closed, and $\mathcal{E}$ is B-usc at σ , $\mathcal{E}$ is closed at σ . $\square$

COROLLARY 4.6. *If $|F| = 1$, then $\mathcal{E}$ is B-continuous at σ if and only if $\sigma \in \text{int } \Theta_c$ if and only if $\mathcal{E}$ is Bouligand-continuous at σ .*

Remark 4.7. It can be easily realized that Proposition 4.2 is valid with $\mathcal{E}_c$ instead of $\mathcal{E}$. On the other hand, from Proposition 4.5 it is possible to prove that $|F| = 1$ implies that $\mathcal{E}_c$ is H-continuous at σ . In fact, since $\mathcal{E}$ is B-usc at σ , it is also H-usc at σ . In order to prove that $\mathcal{E}$ is H-lsc at σ let $\varepsilon > 0$. F is a singleton, so we can write $F = \{\bar{x}\}$. Since $\mathcal{E}$ is B-usc at σ , there exists $\delta > 0$ such that $E_1 \subset B(\bar{x}; \varepsilon)$ if $d(\sigma_1, \sigma) < \delta$. If such a system σ_1 is consistent, we have $d(\bar{x}, E_1) < \varepsilon$, so that $\bar{x} \in E_1 + \text{cl } B(\bar{x}; \varepsilon)$, i.e., $E = \{\bar{x}\} \subset E_1 + \text{cl } B(\bar{x}; \varepsilon)$. Consequently, all the statements in Proposition 4.5 are still valid with $\mathcal{E}_c$ instead of $\mathcal{E}$ except (ii), which should be replaced by

- (ii) $\mathcal{E}_c$ is B-lsc at σ .

5. The complex case: $E \neq \emptyset$, $|F| > 1$. In this section we will consider the following.

Assumption 5.1. The nominal system σ satisfies $|T| \geq n$, $E \neq \emptyset$, and $|F| > 1$.

We associate with σ the homogeneous system $\sigma_H = \{a'_t x \geq 0, t \in T\} \in \Theta_c$. The solution set of σ_H is $F_H = 0^+ F$ so that F is bounded if and only if $F_H = \{0_n\}$.

The next result provides two new conditions that can be aggregated to the list in Theorem 2.3 (under the assumptions we have made).

THEOREM 5.2. *The following statements are equivalent to each other:*

- (i) $\mathcal{E}$ is B -lsc at σ .
- (ii) $\sigma \in \text{int } \Theta_c$.
- (iii) $\sigma \in \text{int } \Xi$.

Proof. We shall prove (i) $\Leftrightarrow$ (ii) $\Leftrightarrow$ (iii).

(i) $\Rightarrow$ (ii) Let $W := B(\bar{x}, 1)$, where $\bar{x}$ is any point of E . Since $W \cap E \neq \emptyset$, there exists a positive number ε such that $W \cap E_1 \neq \emptyset$ whenever $d(\sigma_1, \sigma) < \varepsilon$. In this case $W \cap F_1 \neq \emptyset$ and $F_1 \neq \emptyset$. Therefore $\sigma \in \text{int } \Theta_c$.

(ii) $\Rightarrow$ (i) Suppose that $\sigma \in \text{int } \Theta_c$ and $\mathcal{E}$ is not B -lsc at σ ; then there exist an open set W and a sequence $\{\sigma_r\}$ such that

$$E \cap W \neq \emptyset, \quad \lim_r \sigma_r = \sigma \quad \text{and} \quad E_r \cap W = \emptyset, \quad r = 1, 2, \dots$$

Let $\bar{x} \in E \cap W$ and consider $\delta > 0$ such that $\text{cl } B(\bar{x}, \delta) \subset W$. The fact that $\sigma \in \text{int } \Theta_c$ yields that $\liminf_{r \geq r_0} F_r = F$ for some r_0 (Theorem 2.3). Hence there is a sequence $\{x^r\}$ with $x^r \in F_r$ and $\lim_r x^r = \bar{x}$. The closed convex sets $F_r \cap \text{cl } B(\bar{x}, \delta)$ are nonempty for r large enough and are the convex hulls of their extreme point sets. None of these extreme points can belong to the interior of $\text{cl } B(\bar{x}, \delta)$ because $E_r \cap \text{cl } B(\bar{x}, \delta) \subset E_r \cap W = \emptyset$; in fact if z belongs to $F_r \cap B(\bar{x}, \delta)$, then $z \notin E_r$ and so there are $u \neq v \in F_r$ and $0 < \lambda < 1$, such that

$$z = \lambda u + (1 - \lambda)v.$$

By taking $\tilde{u} \neq \tilde{v}$ in the segment $[u, v] \subset F_r$ close enough to z , we can express

$$z = \alpha \tilde{u} + (1 - \alpha)\tilde{v},$$

for some $0 < \alpha < 1$, with $\tilde{u}, \tilde{v} \in F_r \cap B(\bar{x}, \delta)$. So z is not an extreme point of $F_r \cap \text{cl } B(\bar{x}, \delta)$. Hence any point of $F_r \cap \text{cl } B(\bar{x}, \delta)$, in particular x^r , can be written as a convex combination of $(n + 1)$ points z_i^r in $F_r \cap \text{bd } B(\bar{x}, \delta)$:

$$x^r = \sum_{i=1}^{n+1} \lambda_i^r z_i^r,$$

$\lambda_i^r \geq 0$, $\sum_{i=1}^{n+1} \lambda_i^r = 1$. By a compactness argument, there are subsequences $\{\lambda_i^{r_k}\}$ and $\{z_i^{r_k}\}$ such that

$$\lim_{r_k} \lambda_i^{r_k} = \lambda_i, \quad \lim_{r_k} z_i^{r_k} = z_i, \quad i = 1, \dots, n + 1.$$

So we have $\lambda_i \geq 0$, $\sum_{i=1}^{n+1} \lambda_i = 1$, $z_i \in F$, $\|z_i - \bar{x}\| = \delta$, $i = 1, \dots, n + 1$, and

$$\bar{x} = \sum_{i=1}^{n+1} \lambda_i z_i,$$

which is a contradiction with $\bar{x}$ being an extreme point of F .

(ii) $\Rightarrow$ (iii) Let $\varepsilon_1 > 0$ such that $F_1 \neq \emptyset$ for all σ_1 , $d(\sigma_1, \sigma) < \varepsilon_1$. Since $E \neq \emptyset$, $\text{span}\{a_t, t \in T\} = \mathbb{R}^n$ and there exists a set $S \subset T$ such that $\{a_t, t \in S\}$ is a basis for $\mathbb{R}^n$. Due to the continuity of the determinant function, there exists $\varepsilon_2 > 0$ such

that if $\|c_t - a_t\| < \varepsilon_2$ for all $t \in S$, then $\{c_t, t \in S\}$ is also a basis for $\mathbb{R}^n$. Then, for $\sigma_1 = \{c'_t x \geq d_t, t \in T\}$ such that $d(\sigma_1, \sigma) < \min\{\varepsilon_1, \varepsilon_2\}$ we have that $F_1 \neq \emptyset$ and it does not contain any lines, i.e., $E_1 \neq \emptyset$. Then $\sigma \in \text{int } \Xi$.

(iii) $\Rightarrow$ (ii) is immediate. $\square$

Remark 5.3. This proposition is valid replacing $\mathcal{E}$ by $\mathcal{E}_c$ (and so all the results in this section are also valid for $\mathcal{E}_c$). The proof is as follows.

$\sigma \in \text{int } \Theta_c$ implies that $\mathcal{E}$ is B-lsc at σ and this entails that $\mathcal{E}_c$ is B-lsc at σ , so we have just to prove that $\mathcal{E}_c$ is not B-lsc at σ if $\sigma \notin \text{int } \Theta_c$. Since $|F| > 1$ and $E \neq \emptyset$, we can take $y \in E$ and $z \in F \setminus E$. Let $a := z - y \neq 0_n$ and $W := \{x \in \mathbb{R}^n \mid a'(x - z) < 0\}$. Obviously, $y \in E \cap W$.

By Theorem 2.3, $0_{n+1} \in \text{cl conv}\{(a_t, b_t), t \in T\}$, so that there exists a sequence $\{\lambda^k\} \subset \mathbb{R}_+^{(T)}$ (the cone of nonnegative functions on T with finite support), with $\sum_{t \in T} \lambda_t^k = 1$, for all $k \in \mathbb{N}$, such that

$$\lim_k \sum_{t \in T} \lambda_t^k \begin{pmatrix} a_t \\ b_t \end{pmatrix} = 0_{n+1}.$$

Let $\sigma_r := \{(a_t + \frac{a}{r})'x \geq b_t + \frac{a'z}{r}, t \in T\}$, $r \in \mathbb{N}$. Obviously, $\{\sigma_r\} \subset \Theta_c$, $z \in F_r$, for all $r \in \mathbb{N}$ and $\lim_r \sigma_r = \sigma$. Given $k \in \mathbb{N}$, we have

$$\frac{1}{r} \begin{pmatrix} a \\ a'z \end{pmatrix} + \sum_{t \in T} \lambda_t^k \begin{pmatrix} a_t \\ b_t \end{pmatrix} = \sum_{t \in T} \lambda_t^k \begin{pmatrix} a_t + \frac{a}{r} \\ b_t + \frac{a'z}{r} \end{pmatrix} \in K_r,$$

so that

$$\frac{1}{r} \begin{pmatrix} a \\ a'z \end{pmatrix} = \lim_k \sum_{t \in T} \lambda_t^k \begin{pmatrix} a_t + \frac{a}{r} \\ b_t + \frac{a'z}{r} \end{pmatrix} \in \text{cl } K_r$$

and $\begin{pmatrix} a \\ a'z \end{pmatrix} \in \text{cl } K_r$, i.e., $a'(x - z) \geq 0$ is a consequence of σ_r . Therefore

$$E_r \subset F_r \subset \{x \in \mathbb{R}^n \mid a'(x - z) \geq 0\},$$

which entails $E_r \cap W = \emptyset$ for all $r \in \mathbb{N}$.

The fact that $\mathcal{E}$ is B-lsc at σ if and only if $\mathcal{F}$ is B-lsc at σ suggests a similar relationship for the upper semicontinuity that does not exist: for $\sigma = \{x_1 \geq 0, x_2 \geq 0\}$, only $\mathcal{E}$ is B-usc at σ . The next example shows that it is also possible that only $\mathcal{F}$ is B-usc at σ .

Example 5.4. Let us represent the set in $\mathbb{R}^2$

$$F = \text{cl conv}\{B(-e_1; 1) \cup B(e_1; 1)\} \cap \{(x_1, x_2)' \mid x_2 \geq 0\}$$

by means of the half-planes determined by tangent lines at all the points of the two arcs of circles contained in $\text{bd } F$, except at the point $(-1, 1)'$. Thus,

$$\sigma = \left\{ \begin{array}{l} -(\cos t)x_1 - (\sin t)x_2 \geq -1 - \cos t, \quad t \in [0, \pi/2]; \\ -(\cos t)x_1 - (\sin t)x_2 \geq -1 + \cos t, \quad t \in (\pi/2, \pi] \end{array} \right\} \cup \{x_2 \geq 0\}.$$

It can be realized that σ is nondegenerate, $\mathcal{F}$ is B-lsc and B-usc at σ (F is bounded), and $\mathcal{E}$ is B-lsc (Theorem 5.2) but neither B-usc nor closed at σ .

THEOREM 5.5. *The following statements are true:*

(i) *If $\mathcal{E}$ is closed at $\sigma \in \text{int } \Xi$, then σ is nondegenerate.*

- (ii) If $\mathcal{E}$ is B-usc at σ , then σ and σ_H are nondegenerate.
- (iii) If $\sigma \in \text{int } \Theta_c$ and F is strictly convex, then $\mathcal{E}$ is Bouligand-continuous at σ .
- (iv) If $\mathcal{E}$ is closed at σ and F is bounded, then $\mathcal{E}$ is B-usc at σ .
- (v) If $\mathcal{E}$ is B-usc at σ and $\{a_t, t \in T\}$ is bounded, then $\mathcal{E}$ is closed at σ .

Proof. (i) Choosing $\bar{x} \in F \setminus E$ the proof follows in the same way as in Proposition 4.2(iv).

(ii) Suppose that σ is degenerate and take $\bar{x} \in F \setminus E$ such that $|T(\bar{x})| \geq n$. Consider the open set $W := \mathbb{R}^n \setminus \{\bar{x}\}$ and follow the same steps as in the proof of Proposition 4.2(iv) to find a sequence $\{\sigma_r\}$ convergent to σ with E_r not contained in W , while $E \subset W$. Therefore $\mathcal{E}$ is not B-usc at σ .

If σ_H is degenerate, there exists a vector $y \in 0^+F = F_H$, $\|y\| = 1$, $|T_H(y)| \geq n$. Let $\{t_1, \dots, t_n\} = S \subset T_H(y)$, and take $\bar{x} \in \text{rint } F$. For each $r = 1, 2, \dots$, change the vectors $\left\{\begin{pmatrix} a_{t_1} \\ b_{t_1} \end{pmatrix}, \dots, \begin{pmatrix} a_{t_n} \\ b_{t_n} \end{pmatrix}\right\}$ slightly in such a way that the perturbed systems σ_r satisfy $x^r = \bar{x} + ry \in E_r$ and $\lim_r \sigma_r = \sigma$. If E is bounded, then take any bounded open set W that contains E . If E is unbounded consider $0 \neq z \in \mathbb{R}^n$ such that $z'y = 0$ and $\bar{x} \pm z \in \text{rint } F$. The open set $W = \mathbb{R}^n \setminus ([\bar{x} - z, \bar{x} + z] + \text{cone}\{y\})$ satisfies that $E \subset W$. Then in either case E_r is not included in W for r large enough. This is a contradiction with $\mathcal{E}$ being B-usc at σ .

(iii) It is enough to show that $\mathcal{E}$ is closed at σ , because $\mathcal{E}$ is B-lsc at σ by Theorem 5.2. Since $\sigma \in \text{int } \Theta_c$, we have that $\mathcal{F}$ is B-lsc at σ which implies that $\mathcal{B}$ is closed at σ ([5, Theorem 2], because $E \neq \emptyset$ gives $F \neq \mathbb{R}^n$). Let $\{\sigma_r\}$ and $\{x^r\}$ be such that $x^r \in E_r$, $r = 1, 2, \dots$, $\lim_r \sigma_r = \sigma$, and $\lim_r x^r = \bar{x}$. We have that $x^r \in B_r$, $r = 1, 2, \dots$, and this yields $\bar{x} \in B = E$ (this equality characterizes closed strictly convex sets).

(iv) The fact that F is bounded makes $\mathcal{F}$ is locally bounded at σ and this implies that $\mathcal{E}$ is locally bounded at σ . Since $\mathcal{E}$ is closed and locally bounded at σ , it is also B-usc at σ .

(v) If E is closed, $\mathcal{E}$ B-usc at σ implies that $\mathcal{E}$ is closed at σ . Now, suppose that E is not closed. Let $\{x^r\} \subset E$ such that $\lim_r x^r = \bar{x} \notin E$ ($\bar{x} \in F$ because $\mathcal{F}$ is closed). Let $W := \mathbb{R}^n \setminus \{\bar{x}\} \supset E$, $y^r := \bar{x} - x^r$ and $\sigma_r := \{a'_t(x - y^r) \geq b_t, t \in T\}$, $r = 1, 2, \dots$. Obviously,

$$x \in F_r \Leftrightarrow x - y^r \in F \Leftrightarrow x \in F + y^r,$$

i.e., $F_r = F + y^r$, and so $E_r = E + y^r$. The boundedness of the coefficients a_t implies that $\lim_r \sigma_r = \sigma$. Finally, $\bar{x} = x^r + y^r \in E + y^r = E_r$, and thus E_r is not included in W , which contradicts the assumption that $\mathcal{E}$ is B-usc at σ . $\square$

Remark 5.6. Example 5.7 shows that Theorem 5.5(iii) does not hold for polyhedrons. Example 5.8 shows that we cannot substitute any of the hypotheses by just σ being nondegenerate, which appears to be the condition on the external representation of F closest to $\mathcal{E}$ being closed at σ .

Example 5.7. Let $\sigma = \{a'_t x \geq b_t, t \in \mathbb{N} \cup \{0\}\}$, where

$$\begin{pmatrix} a_t \\ b_t \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix}, \quad t \in \mathbb{N}, \quad \begin{pmatrix} a_0 \\ b_0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

Then σ is degenerate, and $\mathcal{F}$ and $\mathcal{E}$ are B-lsc and not B-usc at σ . Moreover, $\mathcal{E}$ is not closed at σ (consider the systems σ_r obtained by changing $\begin{pmatrix} a_1 \\ b_1 \end{pmatrix}$ by $(1/r, -1, 1/r)'$). Finally observe that $\sigma_r = \{a'_t x \geq b_t, t = 0, \dots, r-1; t^{-2}x_1 - x_2 \geq t^{-1}, t =$

$r, r+1, \dots\}$ verifies that $d(\sigma_r, \sigma) = r^{-1}$ and E_r is unbounded with no cluster points, even though $|E| = 1$.

Example 5.8. Let F be the convex hull of the points

$$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \text{ and } \begin{pmatrix} \cos t \\ 0 \\ \sin t \end{pmatrix}, \quad t \in \left[0, \frac{\pi}{2}\right].$$

Then, σ is nondegenerate and $\{a_t, t \in T\}$ is bounded; $\mathcal{F}$ and $\mathcal{E}$ are B-lsc at σ , $\mathcal{F}$ is B-usc at σ while $\mathcal{E}$ is not; $\mathcal{E}$ is not closed at σ and E is not closed.

THEOREM 5.9. *If $|T| < \infty$, then the following statements hold:*

- (i) $\mathcal{E}$ is B-lsc at σ if and only if σ satisfies Slater condition.
- (ii) $\dim F = n$ if σ is nondegenerate. In such a case Slater condition holds.
- (iii) $\mathcal{E}$ is closed at σ if and only if σ is nondegenerate.
- (iv) $\mathcal{E}$ is Bouligand-continuous at σ if and only if σ is nondegenerate.
- (v) $\mathcal{E}$ is B-usc at σ if and only if σ and σ_H are nondegenerate.
- (vi) $\mathcal{E}$ is B-continuous at σ if and only if σ and σ_H are nondegenerate.

Proof. (i) From Theorems 5.2 and 2.3 we get that $\mathcal{E}$ is B-lsc at $\sigma \Leftrightarrow \sigma \in \text{int } \Theta_c \Leftrightarrow \sigma$ satisfies the SS condition. The latter condition is just Slater condition because $|T| < \infty$.

(ii) Suppose $\dim F < n$ and take some $\begin{pmatrix} a \\ b \end{pmatrix} \in \mathbb{R}^{n+1}$, $a \neq 0_n$, such that $a'x = b$ for all $x \in F$. The assumptions of this section guarantee the existence of an edge of F . Take $\bar{x} \in F \setminus E$ contained in some edge of F . By [6, Theorem 9.23], $\dim\{a_t \mid t \in T(\bar{x})\} = n-1$ and σ does not contain the trivial inequality. Since σ is nondegenerate, $|T(\bar{x})| < n$, so it must be $|T(\bar{x})| = n-1$, and hence the set $\{a_t \mid t \in T(\bar{x})\}$ is linearly independent.

The inequalities $a'x \geq b$ and $-a'x \geq -b$ are consequences of σ , so $\begin{pmatrix} a \\ b \end{pmatrix} \in \text{cl } K = K$ and $-\begin{pmatrix} a \\ b \end{pmatrix} \in \text{cl } K = K$, i.e.,

$$(5.1) \quad \begin{pmatrix} a \\ b \end{pmatrix} = \sum_{t \in T} \lambda_t \begin{pmatrix} a_t \\ b_t \end{pmatrix} + \mu \begin{pmatrix} 0_n \\ -1 \end{pmatrix}$$

and

$$(5.2) \quad -\begin{pmatrix} a \\ b \end{pmatrix} = \sum_{t \in T} \gamma_t \begin{pmatrix} a_t \\ b_t \end{pmatrix} + \delta \begin{pmatrix} 0_n \\ -1 \end{pmatrix}$$

for some $\lambda_t, \gamma_t, \mu, \delta \geq 0$ (not all λ_t s and γ_t s null). Multiplying by $\begin{pmatrix} \bar{x} \\ -1 \end{pmatrix}$, we get

$$0 = a'\bar{x} - b = \sum_{t \in T} \lambda_t (a'_t \bar{x} - b_t) + \mu$$

and

$$0 = -a'\bar{x} + b = \sum_{t \in T} \gamma_t (a'_t \bar{x} - b_t) + \delta$$

which implies that $\mu = \delta = 0$ and $\lambda_t = \gamma_t = 0$ for $t \in T \setminus T(\bar{x})$. In this way, by adding (5.1) and (5.2), we have

$$0_n = \sum_{t \in T(\bar{x})} (\lambda_t + \gamma_t) a_t.$$

This yields $\lambda_t = \gamma_t = 0$ for all t because $\{a_t, t \in T(\bar{x})\}$ is linearly independent. This is a contradiction with $0_n \neq a = \sum_{t \in T(\bar{x})} \lambda_t a_t = 0_n$. So $\dim F = n$. (In particular, σ satisfies the Slater condition.)

(iii) The proof that $\mathcal{E}$ is closed at σ when σ is nondegenerate proceeds as in Proposition 4.2(v), by considering that $\lim_r x^r = \bar{x} \notin E$. Theorem 5.5(i) completes the proof.

(iv) It is a consequence of (i), (ii), and (iii).

(v) The direct implication is Theorem 5.5(ii). To prove the converse assume that $\mathcal{E}$ is not B-usc at σ and take an open set W and a sequence $\{\sigma_r\}$ such that $\lim \sigma_r = \sigma$ and $E_r \subsetneq W$ for all $r \in \mathbb{N}$. We will show that σ_H is degenerate. For each $r \in \mathbb{N}$, take $x^r \in E_r \setminus W$ and let $T_r(x^r)$ be its set of active indices for the system σ_r . Since $x^r \in E_r$, $|T_r(x^r)| \geq n$. The fact that $T_r(x^r) \subset T$ gives the existence of a constant subsequence, which we may assume with no loss of generality that it is $\{T_r(x^r)\}$ itself. Hence, $T_r(x^r) = S$ for all $r \in \mathbb{N}$ and for some $S \subset T$, $|S| \geq n$. Now, if $\{x^r\}$ were bounded, we could suppose that it is convergent to some point $\bar{x}$. The assumption on the nondegeneracy of σ and (iii) entail that $\bar{x} \in E \subset W$ which is a contradiction with $x^r \in E_r \setminus W$ for all $r \in \mathbb{N}$. Thus, $\{x^r\}$ is unbounded, in particular assume that $x^r \neq 0$. The sequence has a subsequence (that we suppose is itself) such that $\lim_r x^r / \|x^r\| = y \neq 0$ exists. Now, if $\sigma_r = \{(a_t^r)'x \geq b_t^r, t \in T\}$ we have for each $t \in T$

$$(a_t^r)' x^r / \|x^r\| \geq b_t^r / \|x^r\|,$$

and by letting $r \rightarrow \infty$, it follows that $a_t' y \geq 0$. Hence y is a solution of σ_H . Moreover, $S = T_r(x^r) \subset T_H(y)$. Observe that $F_H = 0^+ F$ is a convex cone with no lines (because $E \neq \emptyset$) so the only extreme point of F_H is $0_n \neq y$. Finally, $|T_H(y)| \geq |S| \geq n$ is a contradiction with σ_H nondegenerate.

(vi) It follows from (i), (iii), and (v). $\square$

Remark 5.10. By replacing the Slater condition by the SS condition, statement (ii) can be easily extended to a particular class of semi-infinite systems: those which are Farkas–Minkowski and locally polyhedral with $\inf_{t \in T} \|a_t\| > 0$ (see [6] for the definitions and basic properties of such systems).

REFERENCES

- [1] B. BANK, J. GUDDAT, D. KLATTE, B. KUMMER, AND K. TAMMER, *Non-linear Parametric Optimization*, Birkhäuser Verlag, Basel, Switzerland, 1983.
- [2] M. J. CANOVAS, M. A. LOPEZ, AND J. PARRA, *Upper semicontinuity of the feasible set mapping for linear inequality systems*, Set-Valued Anal., 10 (2002), pp. 361–378.
- [3] M. R. DAVIDSON, *Stability of the extreme point set of a polyhedron*, J. Optim. Theory Appl., 90 (1996), pp. 357–380.
- [4] T. FISCHER, *Contributions to semi-infinite linear optimization*, in Approximation and Optimization in Mathematical Physics, B. Brosowski and E. Martensen, eds., Verlag Peter Lang, Frankfurt-Main, Germany, 1983, pp. 175–199.
- [5] M. A. GOBERNA, M. LARRIQUETA, AND V. N. VERA DE SERIO, *On the stability of the boundary of the feasible set in linear optimization*, Set-Valued Anal., 11 (2003), pp. 203–223.
- [6] M. A. GOBERNA AND M. A. LOPEZ, *Linear Semi-infinite Optimization*, John Wiley & Sons, Chichester, UK, 1998.
- [7] M. A. GOBERNA, M. A. LOPEZ, AND M. I. TODOROV, *Stability theory for linear inequality systems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 730–743.
- [8] M. A. GOBERNA, M. A. LOPEZ, AND M. I. TODOROV, *Stability theory for linear inequality systems II: Upper semicontinuity of the solution set mapping*, SIAM J. Optim., 7 (1997), pp. 1138–1151.

- [9] M. A. GOBERNA, M. A. LOPEZ, AND M. I. TODOROV, *Extended active constraints in linear optimization with applications*, SIAM J. Optim., 14 (2003), pp. 608–619.
- [10] W. LI, *Sharp Lipschitz constants for basic optimal solutions and basic feasible solutions of linear programs*, SIAM J. Control Optim., 32 (1994), pp. 140–153.
- [11] O. L. MANGASARIAN AND S. FROMOVITZ, *The Fritz John necessary optimality conditions in the presence of equality and inequality constraints*, J. Math. Anal. Appl., 17 (1967), pp. 37–47.
- [12] T. ROCKAFELLAR AND R. WETS, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
- [13] D. WALKUP AND R. J.-B. WETS, *A Lipschitzian characterization of convex polyhedra*, Proc. Amer. Math. Soc., 23 (1969), pp. 167–173.
- [14] X. S. ZHANG AND D. G. LIU, *A note on the continuity of the solutions of parametric linear programs*, Math. Program., 47 (1990), pp. 143–153.

GROVER'S QUANTUM ALGORITHM APPLIED TO GLOBAL OPTIMIZATION*

W. P. BARITOMPA[†], D. W. BULGER[‡], AND G. R. WOOD[‡]

Abstract. Grover's quantum computational search procedure can provide the basis for implementing adaptive global optimization algorithms. A brief overview of the procedure is given and a framework called Grover adaptive search is set up. A method of Dürr and Høyer and one introduced by the authors fit into this framework and are compared.

Key words. discrete optimization, global optimization, Grover iterations, Markov chains, quantum computers, random search

AMS subject classifications. 90C30, 68Q99, 68Q25

DOI. 10.1137/040605072

1. Introduction. This paper aims to provide the global optimization community with some background knowledge of quantum computation and to explore the importance of this topic for the future of global optimization.

Quantum computing [7] holds great potential to increase the efficiency of stochastic global optimization methods. Current estimates are that quantum computers are likely to be in commercial production within two or three decades. These devices will be in many respects similar to the computers of today but will utilize circuitry capable of quantum coherence [7], enabling data to be manipulated in entirely new ways.

Grover introduced in [9] a quantum algorithm (that is, an algorithm to be executed on a quantum computer) for locating a “marked” item in a database. This was extended in [2] to a quantum algorithm for locating one of an unknown number of marked items. The latter method was incorporated into a minimization algorithm by Dürr and Høyer in [8] (unpublished, but available electronically—see the reference list).

Dürr and Høyer's algorithm can be viewed as an example of Grover adaptive search (GAS), an algorithmic framework we introduced in [4]. GAS in turn is a quantum-computational implementation of hesitant adaptive search [6], a parameterized pseudoalgorithm whose performance is well understood. Here we analyze Dürr and Høyer's method, present another version of GAS, and explore the relative merits of the two methods via numerical simulation.

Outline. Section 2 presents the general optimization problem and introduces some notation and terminology. Section 3 gives a brief overview of quantum computation and Grover's algorithm. Section 4 describes the GAS framework. Section 5 discusses the considerations involved in choosing the “rotation count sequence,” the parameter distinguishing one GAS algorithm from another. Section 6 presents Dürr

*Received by the editors March 11, 2004; accepted for publication (in revised form) November 22, 2004; published electronically August 3, 2005. This research was supported by the Marsden Fund of the Royal Society of New Zealand.

<http://www.siam.org/journals/siopt/15-4/60507.html>

[†]Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand (b.baritomba@math.canterbury.ac.nz).

[‡]Department of Statistics, Macquarie University, NSW 2109, Australia (dbulger@efs.mq.edu.au, gwood@efs.mq.edu.au).

and Høyer's algorithm, extending and correcting the theoretical analysis in [8]. In section 7, we present a refined version of GAS, and in section 8 this version is compared to that of Dürr and Høyer by numerical simulation. Section 9 concludes the paper.

2. Optimization problem. We consider the following finite global optimization problem:

$$\begin{aligned} & \text{minimize } f(x) \\ & \text{subject to } x \in S, \end{aligned}$$

where f is a real-valued function on a finite set S .

Throughout this paper we associate with the objective function f the following definitions. Let $N = |S|$, the cardinality of the finite set S . We will usually assume N to be a power of two. Let $\ell_1 < \dots < \ell_K$ be the distinct objective function values in the range of f . Notice that there may be more than K points in S . Given the uniform probability measure μ on S , we let π be the range measure given by the stochastic vector $(\pi_1, \dots, \pi_K)$ induced by f . That is, $\pi_j = |f^{-1}(\ell_j)|/N$ for $j = 1, 2, \dots, K$. Let p_j denote $\sum_{i=1}^j \pi_i$, the probability that a random point has value of ℓ_j or less. In particular, $p_K = 1$. Corresponding to each function value is an *improving region*, that part of the domain having a strictly better value, and we call its measure under μ the *improving fraction* p . (Usually the specified function value will be the best yet seen, and thus the improving region will be the set of points with objective function values better than any yet seen.)

3. Quantum computing. This paper concerns optimization algorithms that require the use of a quantum computer. The characteristic feature of a quantum computer is that, in place of conventional computer *bits*, *quantum bits*, or *qubits*, are used. A qubit can be in a simultaneous superposition of “off” and “on” and thus allows *quantum parallelism*, where a single quantum circuit can simultaneously perform a calculation on a superposed input, corresponding to very many conventional inputs.

The Grover mechanism. The quantum procedure germane to our purposes is Grover search [9]. This is one of the major advances to date in the fledgling field of quantum computation. More details are given in [4], but we reiterate the salient features here and give an intuitive discussion.

Consider the following general search problem. Let n be a positive integer, and let $S = \{0, 1\}^n$, so that the domain size $N = 2^n$. Let $h : S \rightarrow \{0, 1\}$. We wish to find a point $u \in S$ such that $h(u) = 1$. We further assume that h is a black-box, that is, that knowledge of h can only be gained by sampling (evaluation), but no structural information is available.

With conventional computing, the Boolean function h could be implemented as a subroutine, i.e., a conventional logic circuit constructed to take an input string of bits, representing a point of S , and output the associated bit value of h . The subroutine could then be applied to all points of S , in succession, to find a required point. Such a conventional program would require on average $N/2$ evaluations to find a marked point.

In quantum computing, the circuit implementing h (using gates that work with qubits) inputs and outputs *superpositions*. Thus it “sees” many possible answers at once. On a quantum computer, observing the output will collapse it into a conven-

tional bit string, according to a probability distribution determined by the superposition; thus quantum computing has a stochastic side. Rather than loop through the N points in S , a quantum computer can operate on superposed states in such a way that the probability distribution governing the collapse can be changed. Grover in [9] showed if exactly one point is marked, then only $\frac{\pi}{4}\sqrt{N}$ such operations are required to find the marked point.

Denote the set of *marked* points by $M = \{u \in S | h(u) = 1\}$ and denote the number of these marked target points by t . We may or may not be aware of the value of t . Let p be the proportion of marked points, t/N .

Grover introduced the *Grover rotation operator*, which incorporates the oracle for h and provides a means of implementing a certain phase-space rotation of the states of a quantum system encoding points in the domain S . Repeated applications of this rotation can be used to move from the equal amplitude state, which is simple to prepare within a quantum computer, toward the states encoding the unknown marked points. For details see [4, 2, 9].

A *Grover search of r rotations* applies the above rotation operator r times, starting from the equal amplitude superposition of states, and then observes (and hence collapses to a point) the output state. The mathematical details in [2] show that executing such a search of r rotations generates domain points according to the following probability distribution γ on S :

$$(3.1) \quad \gamma(\{x\}) = \begin{cases} \frac{g_r(p)}{t}, & x \in M, \\ \frac{1 - g_r(p)}{N - t}, & x \in S \setminus M, \end{cases}$$

where

$$(3.2) \quad g_r(p) = \sin^2 [(2r + 1) \arcsin \sqrt{p}].$$

Note that in the special case of $r = 0$, Grover search observes only the prepared equal amplitude superposition of states and so reduces to choosing a point uniformly from the domain.

Most of the work in implementing the Grover rotation operator is in the oracle query, so the cost of a Grover search of r rotations is taken as the cost of r oracle queries. The output is a point in S , and as one would usually want to know if it is in M or not, a further oracle query (acting on the point) would give the function value under h .

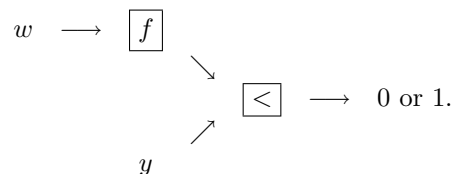
Grover search is sometimes portrayed as a method for the database table lookup problem. This is only one elementary application, however. Other interesting applications concern “marking functions” h which are more than simple tests of indexed data. Examples relating to data encryption and the satisfiability problem are given in [2, 9].

From searching to optimizing. Grover search solves a special global optimization problem: it finds a global maximum of h . For the more general problem introduced in section 2, our intention is to use Grover search repeatedly within a global optimization method of the *adaptive search* variety. Adaptive search methods produce, or attempt to produce, an improving sequence of samples, each uniformly distributed in the improving region of the previous sample (see [16, 15, 5]).

Given an objective function $f : S \rightarrow \mathbb{R}$ and a point $X \in S$ with $f(X) = Y$, we use Grover’s algorithm to seek a point in the improving region $\{w \in S : f(w) < Y\}$.

As described above, Grover's algorithm requires an oracle, a quantum circuit able to classify a point $w \in S$ as inside or outside the target set (see [10]). This will be the oracle for the Boolean function $h(w) = (f(w) < Y)$.

We denote by a "boxed" name the oracle for a given function. Symbolically $\boxed{h}$ is found as shown:



The additional comparison logic circuitry $\boxed{<}$ to construct $\boxed{h}$ is minimal, and we will take the cost of $\boxed{h}$ and $\boxed{f}$ to be the same.

As far as Grover's algorithm is concerned, $\boxed{h}$ is simply a black box quantum circuit, inputting a point w in S (or a superposition of such points) and outputting

$$\begin{cases} 1, & f(w) < y, \\ 0, & f(w) \geq y \end{cases}$$

(or the appropriate superposition of such bits).

Grover search of r rotations, using the compound oracle depicted above, will require r uses of the objective function suboracle $\boxed{f}$ and will output a random domain point. An additional oracle query is required to determine whether the output is an improvement or not. Therefore, for practical purposes, we can consider the cost of running Grover's algorithm to be $r + 1$ objective function evaluations (plus additional costs, such as the cost of the comparisons, which we will ignore).

As a point of departure for the mathematics to follow, we can condense this subsection into the following axiom, and henceforth dispense with any direct consideration of quantum engineering. (Note that the content of this axiom is taken for granted in [2] and many other recent publications on quantum searching.)

Axiom 1. Given $f : S \rightarrow \mathbb{R}$ and $Y \in \mathbb{R}$, there is a search procedure on a quantum computer, which we shall call a "Grover search of r rotations on f with threshold Y ," outputting a random point $x \in S$ distributed uniformly in

$$\begin{aligned} &\{w \in S : f(w) < Y\} \quad \text{with probability } g_r(p), \text{ or uniformly in} \\ &\{w \in S : f(w) \geq Y\} \quad \text{otherwise,} \end{aligned}$$

where $p = |\{w \in S : f(w) < Y\}|/|S|$. The procedure also outputs $y = f(x)$. The cost of the procedure is $r + 1$ objective function evaluations.

4. Grover adaptive search. This section presents the GAS algorithm introduced in [4]. The algorithm requires as a parameter a sequence $(r_n : n = 1, 2, \dots)$ of *rotation counts*. Initially, the algorithm chooses a sample uniformly from the domain and evaluates the objective function at that point. At each subsequent iteration, the algorithm samples the objective function at a point determined by a Grover search. The Grover search uses the best function value yet seen as a threshold. Here is the algorithm in pseudocode form:

GROVER ADAPTIVE SEARCH (GAS).

1. Generate X_1 uniformly in S , and set $Y_1 = f(X_1)$.
2. For $n = 1, 2, \dots$, until a termination condition is met, do:
 - (a) Perform a Grover search of r_n rotations on f with threshold Y_n , and denote the outputs by x and y .
 - (b) If $y < Y_n$, set $X_{n+1} = x$ and $Y_{n+1} = y$; otherwise, set $X_{n+1} = X_n$ and $Y_{n+1} = Y_n$.

GAS fits into the adaptive search framework developed in [5, 6, 15, 16, 17] which has proved useful for theoretical studies of convergence of stochastic global optimization methods. All adaptive algorithms assume “improving” points can be found (at some cost). If Grover’s algorithm were only applicable to database lookup, one might get the impression that GAS would require all function values to be first computed and tabled before they could then be marked. However, Grover’s algorithm can find points in an unknown target set, specified by an oracle. GAS exploits this ability by constructing, at each iteration, an oracle targeting the current improving region. In this way, it builds a sequence of domain points, each uniformly distributed in the improving region of the previous point. Such a sequence is known to converge to the global optimum very quickly; for instance, a unique optimum in a domain of size N will be found after $1 + \ln N$ such improvements, in expectation (see [17]).

Unfortunately this does not mean that GAS can find the global optimum for a cost in proportion to $\ln N$. The reason is that as the improving fraction p decreases, larger rotation counts become necessary to make improvements probable; thus the cost of GAS varies superlinearly in the number of improvements required. Note also that not every iteration finds a point in the improving region. The probability of finding an improvement is given by (3.2), and for a known $p \ll 1$, a rotation count r can be found making this probability very nearly 1. But since in general we can only guess at p , lower probabilities result.

Readers may wonder why we use the best value yet seen as the threshold in the Grover search. In a sense, all of the work of the algorithm is done in the last step, when a Grover search is performed using a threshold only a little larger than the global minimum. This final Grover search is not made any easier by the information gained in earlier steps. In the general case, however, where we have no prior knowledge of the objective function’s range, these earlier steps are an efficient way of finding a good value to use as a threshold in the final step. The earlier steps are not great in number. Moreover, the cost of each step is roughly inversely proportional to the square root of the improving fraction; thus, if the sequence of rotation counts is chosen suitably, most of the earlier steps will be much quicker than the final one.

5. Choosing the rotation count sequence. This section provides a general discussion of the selection of the rotation count sequence used in the GAS algorithm as a precursor to sections 6 and 7, each of which presents a specific selection method.

Why the rotation count should vary. In [4] we considered the possibility of using the same rotation count at each iteration. Although it is easy to construct objective functions for which this method works well, they are exceptional, and in general it is preferable to vary the rotation count as the algorithm progresses.

To see why, suppose that at a certain point in the execution of the GAS algorithm, the best value seen so far is Y , and the improving fraction is $p = |\{w : f(w) < Y\}|/N$. For any given rotation count r , the probability of success of each single iteration of the algorithm is given by $g_r(p)$. Although the rationale for using Grover’s algorithm is to

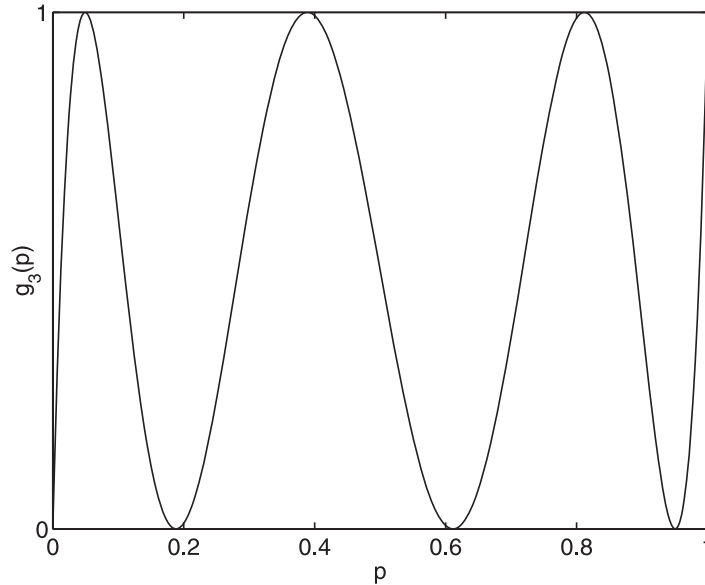


FIG. 1. The probability of a step of three Grover rotations finding an improvement, as a function of the improving fraction p .

increase the probability of finding improving points, there are combinations of values of r and p where the opposite effect occurs. For instance, Figure 1 plots $g_3(p)$ versus p . If $p = 0.2$, then the step is almost guaranteed *not* to find an improvement. If the rotation count varies from each iteration to the next, then this is only an occasional nuisance. But if it is fixed at r , and if the algorithm should happen to sample a point x such that the improving fraction p for $Y = f(x)$ has $g_r(p)$ zero or very small, then the algorithm will become trapped.

How the rotation count should vary. In fact, at each iteration during the execution of the algorithm, some optimal rotation count r is associated with the improving fraction p of the domain (assuming $p > 0$). If it is used for the next Grover search, then an improving point will almost certainly be found. This r is the first positive solution to the equation $g_r(p) = 1$. (Actually of course we must round this to the nearest integer, and therefore success is not absolutely guaranteed, but this would contribute little to the expected cost of the algorithm.)

Unfortunately, in the general case the improving fraction p is unknown, so we are somewhat in the dark as to the choice of rotation counts. In order to make the most use of all the information available to us at each iteration, we could take a Bayesian approach and keep track of a sequence of posterior distributions of the improving fraction at each iteration and choose each rotation count to optimize the change in some statistic of this posterior distribution. As might be expected, this kind of approach appears to be very complex and unwieldy. The methods outlined in the following two sections, however, strike a happy balance between implementability and optimality of rotation count selection.

6. Dürr and Høyer's random method. In this section we outline a method due to Dürr and Høyer for randomly choosing rotation counts and correct two key arguments in its originators' analysis.

Grover's search algorithm provides a method of finding a point within a subset

of a domain. If the size of the target subset is known, the algorithm's rotation count parameter can easily be tuned to give a negligible failure probability. The case of a target subset of *unknown* size is considered in [2], where the following algorithm is presented:

BOYER ET AL. SEARCH ALGORITHM.

1. Initialize $m = 1$.
2. Choose a value for the parameter λ ($8/7$ is suggested in [2]).
3. Repeat:
 - (a) Choose an integer j uniformly at random such that $0 \leq j < m$.
 - (b) Apply Grover's algorithm with j rotations, giving outcome i .
 - (c) If i is a target point, terminate.
 - (d) Set $m = \lambda m$.

Actually, in [2], the final step updates m to $\min\{\lambda m, \sqrt{N}\}$. It is pointless to allow m to exceed $\sqrt{N}$, because for a target set of *any* size, it is known [2] that the optimal rotation count will be no more than $\lceil \pi\sqrt{N}/4 \rceil$. In the global optimization context, however, this point will usually be immaterial, since the target region, though comprising a small proportion of the domain, will normally be large in absolute terms.

For instance, suppose the domain contains 10^{20} elements and suppose finding one of the smallest 10,000 points is required. The optimal rotation count to find a target set of this size is $10^8\pi/4$, substantially less than $\lceil \pi\sqrt{N}/4 \rceil$. The actual target size will be unknown, and therefore the actual optimal rotation count will be unknown. But when m reaches this magnitude, if not before, each step will have a substantial probability (on the order of $1/2$) of finding a target point. Therefore, unless λ is very large, there will be negligible probability of m reaching $\sqrt{N} = 10^{10}$ before a target point is produced. For simplicity, therefore, in this article we ignore the $\sqrt{N}$ ceiling on the growth of m .

In the **quant-ph** internet archive, Dürr and Høyer [8] propose using the Boyer et al. algorithm as the nucleus of a minimization algorithm. Their paper gives the impression that the algorithm is just for the database problem. They begin with "an unsorted table of N items each holding a value from an ordered set. The minimum searching problem is to find the index y such that $T[y]$ is minimum." Again we stress their algorithm fits in the GAS framework and is thus applicable to the general optimization problem.

In their paper, they indicate that every item that is improving is explicitly marked. However, this is a mistake, as it is incompatible with their complexity analysis later in the paper. We describe a corrected version of their method using the terminology of this paper.

DÜRR AND HØYER'S ALGORITHM.

1. Generate X_1 uniformly in S , and set $Y_1 = f(X_1)$.
2. Set $m = 1$.
3. Choose a value for the parameter λ (as in the previous algorithm).
4. For $n = 1, 2, \dots$ until a termination condition is met, do:
 - (a) Choose a random rotation count r_n uniformly distributed on $\{0, \dots, \lceil m-1 \rceil\}$.
 - (b) Perform a Grover search of r_n rotations on f with threshold Y_n , and denote the outputs by x and y .
 - (c) If $y < Y_n$, set $X_{n+1} = x$, $Y_{n+1} = y$, and $m = 1$; otherwise, set $X_{n+1} = X_n$, $Y_{n+1} = Y_n$, and $m = \lambda m$.

This is the special case of GAS arising when the rotation count r_n is chosen

randomly from an integer interval which is initialized to $\{0\}$ at each improvement but which grows exponentially to a maximum of $\{0, \dots, \lceil \sqrt{N}-1 \rceil\}$ between improvements.

The analysis of the algorithm reported in the archive [8] uses incorrect constants from a preprint of [2]. In the following analysis, we correct this by using the published version of [2]. Because the Boyer et al. algorithm underpins that of Dürr and Høyer, we begin with an analysis of the former algorithm. Theorem 3 in [2] is an order of magnitude result, but inspection of the proof implies that the expected time required by the Boyer et al. algorithm to find one of t marked items among a total of N items is bounded by $8\sqrt{N/t}$. This constant can be improved upon, though, as we shall see after the following theorem.

THEOREM 6.1. *The expected number of oracle queries required by the Boyer et al. algorithm with parameter λ to find and verify a point from a target subset of size t from a domain of size N is*

$$(6.1) \quad \sum_{j=0}^{\infty} \frac{\lceil \lambda^j \rceil}{2} \prod_{i=0}^{j-1} \left(\frac{1}{2} + \frac{\sin(4\theta \lceil \lambda^i \rceil)}{4 \lceil \lambda^i \rceil \sin(2\theta)} \right),$$

where $\theta = \arcsin(\sqrt{t/N})$.

Proof. Conditioned on reaching iteration j , the expected number of oracle queries required at that iteration is $\lceil \lambda^j \rceil / 2$ (including the test of the output of Grover's algorithm for target subset membership). The probability of reaching iteration j is a product of failure rates; the probability of the algorithm failing to terminate at iteration j , having reached this iteration, is

$$\frac{1}{2} + \frac{\sin(4\theta \lceil \lambda^j \rceil)}{4 \lceil \lambda^j \rceil \sin(2\theta)}$$

(this is [2, Lemma 2]). Thus the expected number of oracle queries required at iteration j , *not* conditioned on whether the iteration is reached, is

$$\frac{\lceil \lambda^j \rceil}{2} \prod_{i=0}^{j-1} \left(\frac{1}{2} + \frac{\sin(4\theta \lceil \lambda^i \rceil)}{4 \lceil \lambda^i \rceil \sin(2\theta)} \right),$$

and summing over all possible iterations $j = 0 \dots \infty$ gives the result. $\square$

It is straightforward to evaluate the geometrically convergent series (6.1) numerically. By graphing the ratio of (6.1) to $\sqrt{N/t}$ versus t for a range of λ , empirically λ that gave the lowest maximum is 1.34. The plot of Figure 2 uses this value of λ , and it justifies the following observation.

Observation 1. The expected number of oracle queries required by the Boyer et al. algorithm with parameter $\lambda = 1.34$ to find and verify a point from a target subset of size t from a domain of size N is at most $1.32\sqrt{N/t}$.

Now we can derive a performance bound for Dürr and Høyer's algorithm. This is similar to and extends the result in [8]; the main difference is in our treatment of the coefficient of the order bound. Also we correct another technical error in their argument, which is pointed out in our proof below.

THEOREM 6.2. *Assume the validity of the above observation. Let $1 \leq s \leq N$ and assume that there are s points in the domain with strictly better objective function values than the remaining $N-s$ points. The expected number of oracle queries required by Dürr and Høyer's algorithm with $\lambda = 1.34$ to find one of these s points is bounded*

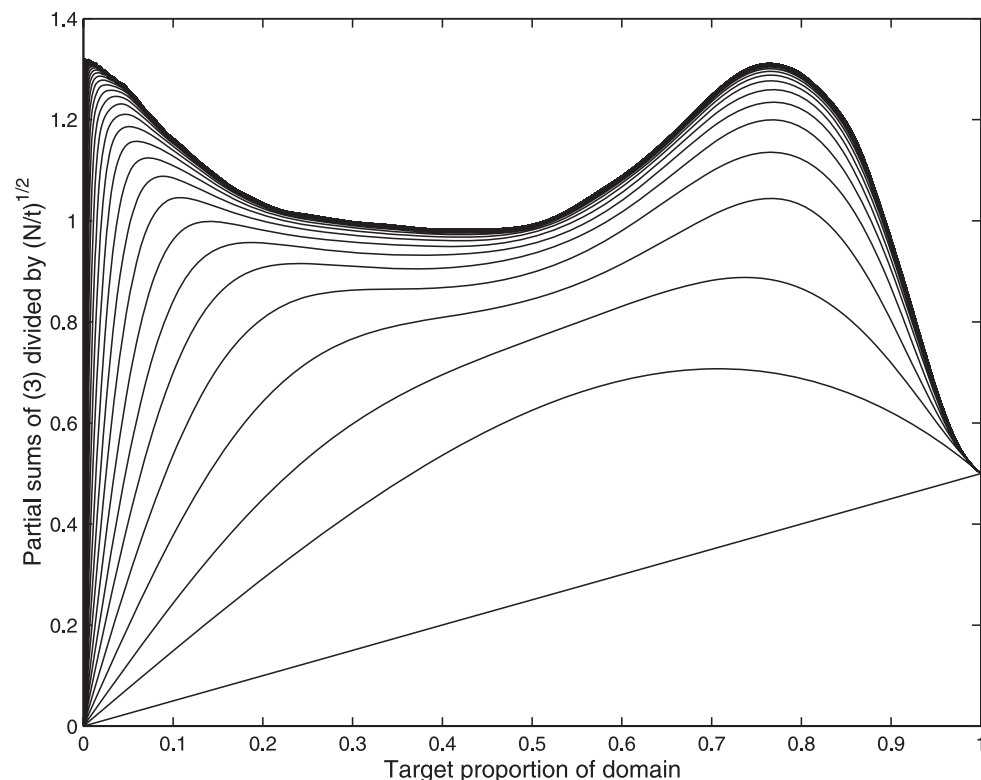


FIG. 2. The ratio between the partial sums of the geometrically convergent series (6.1) and $\sqrt{N/t}$ when $\lambda = 1.34$, plotted against t/N . Note that 1.32 appears to be an upper bound.

above by

$$1.32\sqrt{N} \sum_{r=s+1}^N \frac{1}{r\sqrt{r-1}}.$$

Note that if s is small compared to N , then the above bound approximately equals $2.46\sqrt{N/s}$.

Proof. Assign the domain points ranks from 1 to N , giving the best point rank 1 and so forth. Where several points have equal objective function value, break ties arbitrarily, but let $l(r)$ be the least rank and $h(r)$ the greatest rank among the points with the same value as the rank r point. (In the distinct values case we will have $l(r) = h(r) = r$ for each $r \in \{1, \dots, N\}$.)

Since Dürr and Høyer's algorithm will move through a succession of threshold values with rank above s before finding the desired target point, the bound on the expectation in question is given by

$$(6.2) \quad \sum_{r=s+1}^N p(N, r)B(N, l(r) - 1),$$

where $p(N, r)$ is the probability of the rank r point ever being chosen and $B(N, l(r) - 1)$ is the expected number of iterations required by the Boyer et al. algorithm to find and verify a point from a target subset of size $l(r) - 1$.

The probability $p(N, r) = 1/h(r)$. This is demonstrated in the proof of Theorem 1 in [17] and in Lemma 1 of [8]. Also, by Observation 1, $B(N, l(r) - 1) \leq 1.32\sqrt{N/(l(r) - 1)}$.

In the distinct values case, substitution of the above value for $p(N, r)$ and bound for $B(N, l(r) - 1) = B(N, r - 1)$ into (6.2) gives the theorem immediately. In [8] it is claimed for the case of repeated objective function values that since the equation $p(N, r) = 1/r$ becomes the inequality $p(N, r) \leq 1/r$, the bound still holds. This argument ignores that the value of $B(N, l(r) - 1)$ *increases* (for a given r) when repeated values are allowed. Nevertheless, the theorem holds as follows. Consider $\hat{r} \in \{1, \dots, N\}$ with $l(\hat{r}) < h(\hat{r})$. We examine just that part of the summation in (6.2) with the index going from $l(\hat{r})$ to $h(\hat{r})$,

$$\begin{aligned} \sum_{r=l(\hat{r})}^{h(\hat{r})} p(N, r) B(N, l(r) - 1) &\leq 1.32\sqrt{N} \sum_{r=l(\hat{r})}^{h(\hat{r})} \frac{1}{h(r)\sqrt{l(r) - 1}} \\ &= 1.32\sqrt{N(l(\hat{r}) - 1)} \sum_{r=l(\hat{r})}^{h(\hat{r})} \frac{1}{h(\hat{r})(l(\hat{r}) - 1)} \\ &= 1.32\sqrt{N(l(\hat{r}) - 1)} \sum_{r=l(\hat{r})}^{h(\hat{r})} \frac{1}{r(r - 1)} \\ &\leq 1.32\sqrt{N} \sum_{r=l(\hat{r})}^{h(\hat{r})} \frac{1}{r\sqrt{r - 1}}. \quad \square \end{aligned}$$

Remark 1. Dürr and Høyer's method can be viewed as an implementation of pure adaptive search [17], requiring no more than $1.32(N/t)^{1/2}$ iterations in expectation to find an improvement, when t is the cardinality of the improving region.

7. A new method. In this section we propose an explicit sequence of integers to be used as the GAS rotation count sequence. This gives a special case of GAS that can be identified with an inhomogeneous Markov chain having states $\ell_1, \dots, \ell_K$.

For this paper we have sought an efficient choice for the rotation count sequence used in GAS. This has led us to the special case of GAS arising when the sequence (r_n) is fixed in advanced and determined by the following pseudocode. Note that the sequence of rotation counts it produces is independent of the particular optimization task; its first 33 entries are

$$(7.1) \quad \begin{aligned} &0, 0, 0, 1, 1, 0, 1, 1, 2, 1, 2, 3, 1, 4, 5, 1, 6, 2, 7, 9, \\ &11, 13, 16, 5, 20, 24, 28, 34, 2, 41, 49, 4, 60, \dots \end{aligned}$$

Here is the pseudocode:

ROTATION SCHEDULE CONSTRUCTION ALGORITHM.

1. Initialize u to be the polynomial $u(y) = y$.
2. For $i = 1, 2, \dots$, do:
 - (a) Set $E_u = 1 - \int_0^1 u \, dy$.
 - (b) Set $b' = 0$.
 - (c) For $r = 0, 1, \dots$, until $E_u/(r+1) \leq 2b'$, do:
 - i. Set $v = u + y \int_y^1 (g_r(t)/t) du(t)$.
 - ii. Set $E_v = 1 - \int_0^1 v \, dy$.
 - iii. Set $b = (E_u - E_v)/(r+1)$.
 - iv. If $b > b'$, then,
 - A. Set $r' = r$.
 - B. Set $b' = b$.
 - C. Set $v' = v$.
 - (d) Set $u = v'$.
 - (e) Output i th rotation count r' .

The resulting sequence (7.1) is heuristically chosen to maximize a benefit-to-cost ratio, denoted b in the pseudocode, at each GAS iteration. The reader can verify that u and E_u are the cumulative distribution function and expectation, respectively, of the improving fraction of the domain, after the first $i-1$ iterations of the GAS algorithm. The symbols v and E_v denote the corresponding cumulative distribution function and expectation after a further GAS step of r rotations. The benefit is (somewhat arbitrarily) taken to be the expected decrease in the improving fraction of the domain, $E_u - E_v$. The cost is $r+1$, where r is the number of rotations chosen, as per Axiom 1. The inner loop at (2c) terminates since even if g_r were identically one, the expected improving region measure would halve. Thus, higher rotation counts need not be considered once we pass the point where half the expected improving region measure, divided by the cost, exceeds the current best found benefit-to-cost ratio.

8. Computational results. In section 6 we presented a corrected version of Dürr and Høyer's demonstration of a performance bound for their algorithm. This readily establishes the $O(\sqrt{N/s})$ complexity, inherited from Grover's algorithm. However, even the improved coefficient of 2.46 suggested by Theorem 6.2 is based on an upper bound and may be a poor indicator of the algorithm's actual performance. In this section we study the methods described in sections 6 and 7 using numerical simulation. Our aim is twofold: to tune the parameter λ appearing in Dürr and Høyer's algorithm and then to compare their tuned method against the method of section 7.

Our simulations will determine the length of time each algorithm requires to sample a point in a target region, constituting a certain proportion of the domain. Intuitively, the algorithm terminates upon finding a value equal to or lower than the quantile determined by a proportion α .

Recall that the proportion of the domain with value lower than or equal to ℓ_j is p_j . More precisely, we specify an *intended* quantile proportion α_{nominal} and set $k = \min\{j : p_j \geq \alpha_{\text{nominal}}\}$. We require the algorithm to find a point with value less than or equal to ℓ_k . The target set is $f^{-1}(\{\ell_1, \ell_2, \dots, \ell_k\})$. Let s be its cardinality. So $\alpha = p_k = s/N$ and gives the quantile the algorithm will find. Note that it is the measure under π of $\{\ell_1, \ell_2, \dots, \ell_k\}$. It may be inevitable that α and α_{nominal} differ since it is possible that $p_{k-1} < \alpha_{\text{nominal}} < \alpha = p_k$.

Thus the quantity α is often unknown in practice and is a "global" piece of information. The dependence of performance on global information is unavoidable [14], but we will see that for certain methods, the dependence is primarily on α . For the

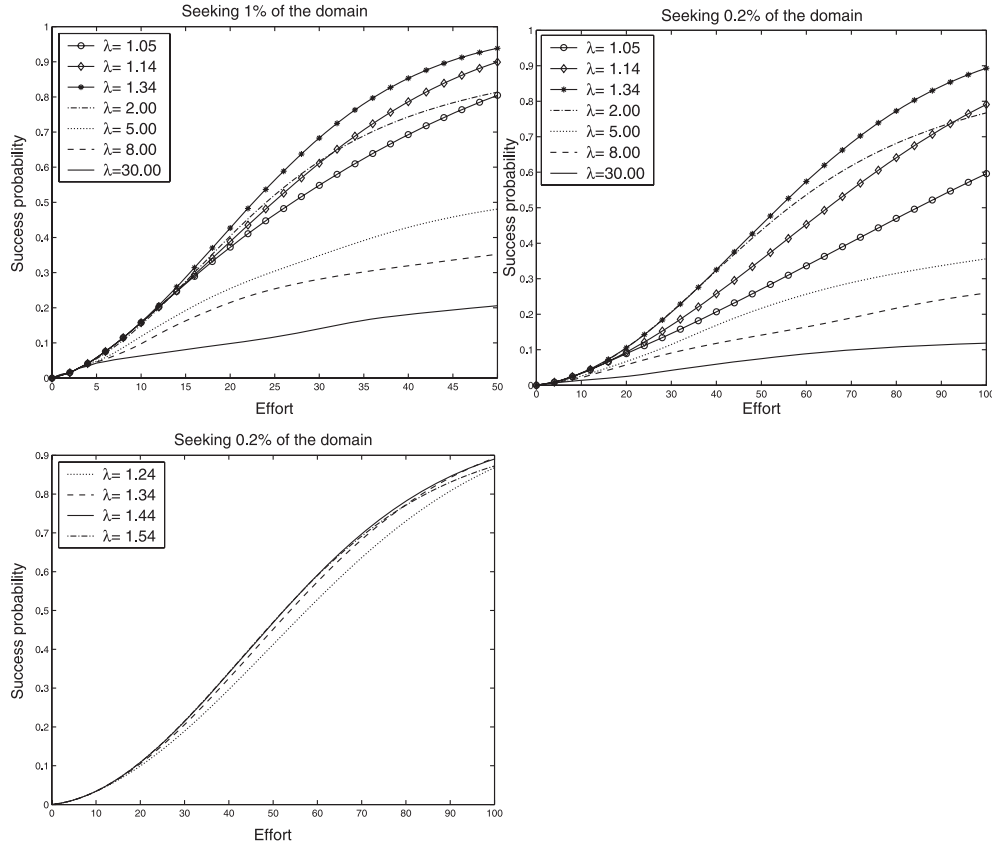


FIG. 3. Performance graphs for Dürr and Høyer's algorithm for various values of the parameter λ and two domain sizes. The third graph repeats the second with a finer mesh of λ values.

rest of this paper we assume α is close to α_{nominal} .

Methodology. For the performance of either algorithm under consideration, the distribution of objective function values influences performance only via the range measure π . Our primary focus here will be the case where π is uniformly distributed over a finite set of distinct function values. Without loss of generality we can take this finite set to be $\{1, \dots, K\}$. For example, to explore seeking the best 1% of the domain under a uniform range distribution (i.e., $\alpha_{\text{nominal}} = 0.01$), using $K = 100$ will be fairly representative. At the end of this section we look briefly at other distributions.

To compare the algorithms, we plot their *performance graphs* [11] which relate practical computational effort to the probability of finding a point in the target set. The performance graph is simply the cumulative distribution function of the effort to success, defined as the number of objective function evaluations before a point in the target set is sampled. We compute these with MATLAB, using standard techniques for Markov chains and stochastic processes.

Tuning λ . Observation 1 suggests the parameter choice $\lambda = 1.34$ for Dürr and Høyer's algorithm. Numerical experimentation agrees with this choice. Figure 3 shows the performance graphs, seeking 1% ($K = 100$) or 0.2% ($K = 500$) of the domain, of Dürr and Høyer's algorithm using a selection of values of λ ranging from 1.05 to 30,

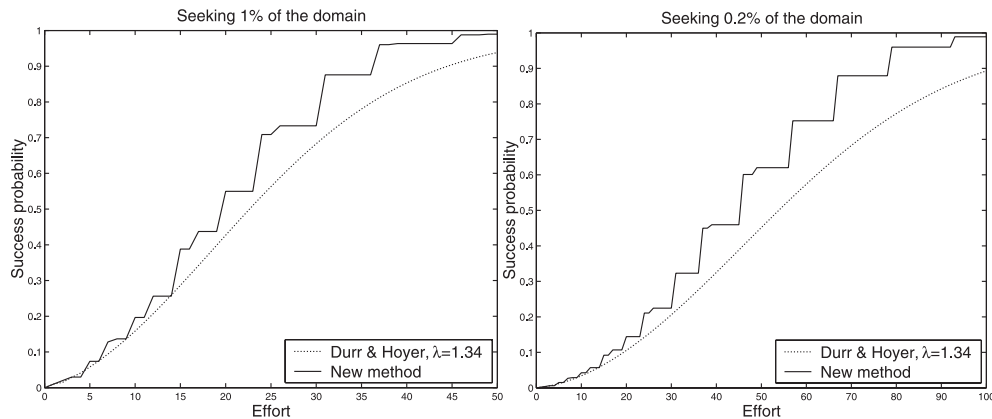


FIG. 4. Performance graphs comparing Dürr and Høyer's method to the method of section 7, for a uniform range distribution.

and including the values $8/7$ and 1.34 suggested by [2] and Figure 2. Performance deteriorates slowly outside of the range from 1.34 to 1.44 , but within that range there is no visible performance gradient. The value of λ may become more important for smaller values of α , but for the remainder of this section we shall use the value $\lambda = 1.34$.

Comparing the new method to Dürr and Høyer. Having settled on the parameter value $\lambda = 1.34$ for Dürr and Høyer's method, we can compare it to the method of section 7. Figure 4 shows that, in the two cases studied, the new method dominates that of Dürr and Høyer. For instance, to sample a target comprising 0.2% of the domain with probability 90% or more, Dürr and Høyer's method requires more than 100 units of effort, whereas the new method requires only 79 (and in fact it then samples the target with probability 96%).

Note also, in the two situations depicted in Figure 4, the estimated bound of $2.46\sqrt{N/s}$ on the expected time required by Dürr and Høyer's algorithm, mentioned following Theorem 6.2, amounts to 24.6 and 55.0 . While the true expectations cannot be computed from any finite portion of the performance graphs, these figures do appear visually to be in approximate agreement with the numerical results.

Nonuniform range distributions. Until now in this section we have assumed a uniform range distribution. This corresponds to the assumption of injectivity of the objective function, that is, that different points in the domain map to different values in the range. In many cases, however, for instance in combinatorial optimization, there may be a unique optimum, or a small number of optimal domain points, but large sets of the domain sharing values in the middle of the range; this results in a nonuniform range distribution.

Experimentation indicates that nonuniformity of the range distribution improves the performance of both methods under study. To produce Figure 5, we randomly created five stochastic vectors of length 20 with first element 0.002 (the remainder of each vector was a point uniformly distributed in $[0, 1]^{19}$ and then scaled to sum to 0.998) and simulated the performance of both methods. Compare this with the last plot of Figure 4. Nonuniformity has improved the performance of the method of section 7 somewhat. However, a greater improvement in Dürr and Høyer's method has allowed it to overtake the method of section 7. Here, for most of the five sample

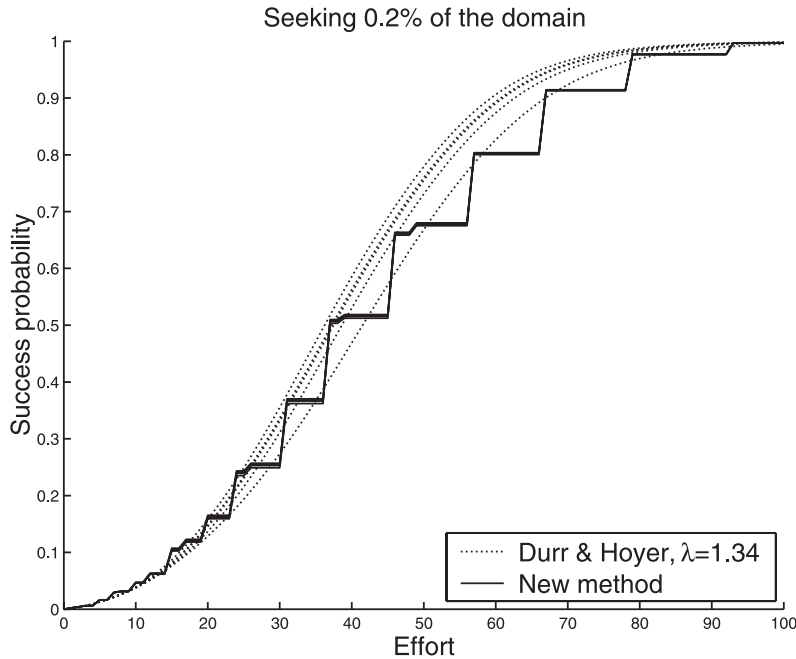


FIG. 5. Performance graphs comparing Dürr and Høyer's method to the method of section 7, for a nonuniform range distribution.

range distributions, Dürr and Høyer's method reaches the target with probability 90% or more after 61 or fewer units of effort, whereas the new method now requires 67.

9. Conclusion. This paper outlines the significance of Grover's quantum search algorithm (with its performance characteristics implying $O(\sqrt{N/t})$ performance taken as an axiom) for global optimization. Grover search can provide the basis of implementing adaptive global optimization algorithms. One example is an algorithm of Dürr and Høyer's introduced as a method for finding minimum values in a database. An improved analysis of Dürr and Høyer's algorithm suggests increasing its parameter λ from $8/7$ to 1.34 . Also, that algorithm fits the Grover adaptive search framework, and thus is applicable to the more general global optimization problem. A new algorithm within the same framework is proposed in section 7. Our numerical experiments in section 8 show that the algorithms have similar performance. The method proposed in section 7 had its parameters tuned for the distinct objective function value case and shows superior performance to that of Dürr and Høyer's in that case. On the other hand, Dürr and Høyer's method (with $\lambda = 1.34$) overtakes the new method if there is a great deal of repetition in objective function values.

A final comment concerning implementation on a quantum computer. This is work mainly for computer engineers of the future, but some indications are known at the present time. A fully functional quantum computer would be able to evaluate an objective function in just the same way as a conventional computer, by executing compiled code. A technical requirement to control quantum coherence, which we have not mentioned previously, is that the gates must implement reversible operations. The code implementing the objective function must be run in the forward direction and then in the reverse direction. This obviously at most doubles the computational effort for a function evaluation compared to a conventional computer.

REFERENCES

- [1] W. P. BARITOMPA, *Handling Different Cost Algorithms*, Department of Mathematics and Statistics, Technical Report, University of Canterbury Christchurch, New Zealand, preprint, 2001.
- [2] M. BOYER, G. BRASSARD, P. HØYER, AND A. TAPP, *Tight bounds on quantum searching*, Fortschr. Phys., 46 (1998), pp. 493–506.
- [3] S. H. BROOKS, *A discussion of random methods for seeking maxima*, Oper. Res., 6 (1958), pp. 244–251.
- [4] D. W. BULGER, W. P. BARITOMPA, AND G. R. WOOD, *Implementing pure adaptive search with Grover's quantum algorithm*, J. Optim. Theory Appl., 116 (2003), pp. 517–529.
- [5] D. W. BULGER, D. L. J. ALEXANDER, W. P. BARITOMPA, G. R. WOOD, AND Z. B. ZABINSKY, *Expected hitting time for backtracking adaptive search*, Optimization, 53 (2004), pp. 189–202.
- [6] D. W. BULGER AND G. R. WOOD, *Hesitant adaptive search for global optimisation*, Math. Programming, 81 (1998), pp. 89–102.
- [7] Centre for Quantum Computation, Oxford, <http://www.qubit.org>.
- [8] C. DÜRR AND P. HØYER, *A quantum algorithm for finding the minimum*, <http://lanl.arxiv.org/abs/quant-ph/9607014>, version 2 (7 Jan. 1999).
- [9] L. K. GROVER, *A fast quantum mechanical algorithm for database search*, in Proceedings of the 28th Annual ACM Symposium on the Theory of Computing, Philadelphia, 1996, pp. 212–219.
- [10] L. K. GROVER, *A framework for fast quantum mechanical algorithms*, in Proceedings of the 30th Annual ACM Symposium on the Theory of Computing, Dallas, 1998, pp. 53–62.
- [11] E. M. T. HENDRIX AND O. KLEPPER, *On uniform covering, adaptive random search and raspberries*, J. Global Optim., 18 (2000), pp. 143–163.
- [12] R. LAFLAMME, *Los Alamos scientists make seven bit quantum leap*, <http://www.lanl.gov/worldview/news/releases/archive/00-041.html> (2000).
- [13] P. W. SHOR, *Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer*, SIAM J. Comput., 26 (1997), pp. 1484–1509.
- [14] C. P. STEPHENS AND W. P. BARITOMPA, *Global optimization requires global information*, J. Optim. Theory Appl., 96 (1998), pp. 575–588.
- [15] G. R. WOOD, Z. B. ZABINSKY, AND B. P. KRISTINSDOTTIR, *Hesitant adaptive search: The distribution of the number of iterations to convergence*, Math. Programming, 89 (2001), pp. 479–486.
- [16] Z. B. ZABINSKY AND R. L. SMITH, *Pure adaptive search in global optimization*, Math. Programming, 53 (1992), pp. 323–338.
- [17] Z. B. ZABINSKY, G. R. WOOD, M. A. STEEL, AND W. P. BARITOMPA, *Pure adaptive search for finite global optimization*, Math. Programming, 69 (1995), pp. 443–448.
- [18] C. ZALKA, *Grover's quantum searching algorithm is optimal*, Phys. Rev. A, 60 (1999), pp. 2746–2751.

FLOWS OVER TIME WITH LOAD-DEPENDENT TRANSIT TIMES*

EKKEHARD KÖHLER[†] AND MARTIN SKUTELLA[‡]

Abstract. More than forty years ago, Ford and Fulkerson studied maximum s - t -flows over time (also called “dynamic” flows) in networks with fixed transit times on the arcs and a fixed time horizon. Here, flow on arcs may change over time and transit times specify the amount of time it takes for flow to travel through a particular arc. Ford and Fulkerson proved that there always exists an optimal solution which sends flow on certain s - t -paths at a constant rate as long as there is enough time left for the flow along a path to arrive at the sink; a flow over time featuring this simple structure is called “temporally repeated.”

Although this result does not hold for the more general and also more realistic setting where transit times depend on the current flow situation, we show that there always exists a provably good temporally repeated solution. Moreover, such a solution can be determined very efficiently by only one minimum convex cost flow computation. Our results rest upon a new model of flow-dependent transit times. It is based on two assumption on the pace of flow on a particular arc. First, the pace of flow on an arc is assumed to be uniform for all flow units on an arc for each point in time. Second, this uniform pace is for each moment determined by the actual amount of flow on this arc. Finally, we show that the resulting flow-over-time problem is strongly NP-hard and cannot be approximated with arbitrary precision in polynomial time, unless $P=NP$.

Key words. network flow, dynamic flow, congestion, approximation algorithm, traffic routing

AMS subject classifications. 90B06, 90B10, 90B20, 90C27, 90C35, 90C59, 68Q25, 68W25

DOI. 10.1137/S1052623403432645

1. Introduction. Flow variation over time is an important feature in network flow problems arising in various applications such as road or air traffic control, production systems, communication networks (e.g., the Internet), and financial flows. The common characteristic are “dynamic” networks with capacities and transit times on the arcs. In contrast to static flow problems, flow values on arcs may change with time in these networks. Moreover, flow does not progress instantaneously but can only travel at a certain pace through the network, which is determined by transit times of arcs.

Another crucial phenomenon in many of those applications is the variation of time taken to traverse an arc with the current (and maybe also past) flow situation on this arc. Since it is already a highly nontrivial problem to map these two aspects into an appropriate and tractable mathematical network flow model, there are hardly any algorithmic techniques known which are capable of providing reasonable solutions even for networks of rather modest size. The main aim of this paper is to make a first step in this direction by providing new insights and algorithmic results which will

*Received by the editors August 1, 2003; accepted for publication (in revised form) November 17, 2004; published electronically August 3, 2005. This work was partially supported by DFG Focus Program 1126, “Algorithmic Aspects of Large and Complex Networks,” grant SK 58/4-1, by EU Thematic Network APPOL II, “Approximation and Online Algorithms,” grant IST-2001-30012, and by Bundesministerium für Bildung und Forschung, Förderkennzeichen 03-MOM4B1. An extended abstract appeared in *Proceedings of the Thirteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SIAM, Philadelphia, 2002, pp. 174–183.

<http://www.siam.org/journals/siopt/15-4/43264.html>

[†]Technische Universität Berlin, Fakultät II—Mathematik und Naturwissenschaften, Institut für Mathematik, Sekr. MA 6–1, Straße des 17. Juni 136, D–10623 Berlin, Germany (ekoehler@math.tu-berlin.de).

[‡]Universität Dortmund, Institut für Angewandte Mathematik, D–44221 Dortmund, Germany (martin.skutella@uni-dortmund.de).

hopefully turn out to have the potential to contribute to practically efficient solution methods.

Problem definition and notation. We consider a directed network $G = (V, E)$ with node set V and arc set E . There is a source node $s \in V$, a sink $t \in V$, and a positive demand value D . Our aim is to find a *quickest flow over time* which satisfies demand D within minimal time horizon (or makespan) T while respecting the following restrictions. Each arc $e \in E$ has a positive capacity u_e which is interpreted as an upper bound on the rate of flow entering e , i.e., a capacity per unit time. Moreover, arc e has an associated positive *transit time* τ_e which determines the amount of time it takes for flow to travel from the tail to the head of that arc. In many real-world applications, a difficult but crucial aspect is that the amount of time needed to traverse an arc of the network increases as the arc becomes more congested. Thus, we consider the case where τ_e is not necessarily fixed but may depend on the amount of flow currently sent through arc e .

In the setting of flows over time, flow conservation constraints require that, for any point θ in time and for any node $v \in V \setminus \{s\}$, the total inflow into node v until time θ is an upper bound on the total outflow out of node v until time θ . In particular, the fact that the inflow may exceed the outflow means that flow can be stored in nodes. However, for a flow over time with finite time horizon T we require that, for any node $v \in V \setminus \{s, t\}$, the total inflow into node v until time T is equal to the total outflow out of node v until time T . In other words, the network must be empty again at time T , as it was at time 0.

Modeling flow-dependent transit times. The crucial parameter for modeling temporal dynamics of flows is the presumed dependency of the actual transit time τ_e on the current (and maybe also past) flow situation on arc e . Unfortunately, there is a tradeoff between the need of modeling this usually highly complex correlation as realistically as possible and the requirement of retaining tractability of the resulting mathematical program.

Due to the latter condition, many models in the literature rely on relatively simple assumptions. For example, the transit time of an arc is often treated as a function of only the flow rate at time of entry to the arc; see, e.g., [7]. However, this assumption is in many cases unrealistic since it does not, for example, preserve the first-in-first-out property encountered in most applications.

In contrast, a fully realistic model of flow-dependent transit times on arcs must take density, speed, and flow rate evolving along the arc into consideration [16]. Unfortunately, even the solution of mathematical models relying on simplifying assumptions is in general still impracticable, i.e., beyond the means of state-of-the-art computers, for problem instances of realistic size (as those occurring in real-world applications such as road traffic control).

Existing models and results. In what follows, we discuss some approaches which can be found in the literature. For a more detailed account and further references we refer the reader to [2, 8, 19, 24, 29, 31]. The existing approaches can be assigned to four groups: Simulation-based approaches like, for example, traffic simulation (see, e.g., [3, 28]); models based on fluid dynamics (see, e.g., [30]) or variational inequalities (see, e.g., [9, 10, 15]); and, finally, mathematical programming-based approaches, which we discuss in more detail below.

While simulation is a powerful tool to evaluate complex flow scenarios, it misses the optimization potential. On the other hand, fluid models and other models based on differential equations capture very well the dynamical behavior of flows, but cannot

currently handle large networks. Mathematical programming and, in particular, network flow theory have the potential to overcome this drawback at the cost of a less precise modeling of real flow behavior.

Merchant and Nemhauser [26] formulate a nonlinear and nonconvex program where time is being discretized. In their model, the outflow out of an arc in each time period solely depends on the amount of flow on that arc at the beginning of the time period. However, the nonconvexity of their model causes analytical and computational problems. In [27] and [5], special constraint qualifications are described which are necessary to guarantee optimality of a solution in this model. Carey [6] introduces a slight revision of the model of Merchant and Nemhauser which transforms the nonconvex problem into a convex one.

Carey and Subrahmanian [7] introduce a time-expanded network¹ with fixed transit times on the arcs. However, for each time period there are several copies of an arc of the underlying “static” network corresponding to different transit times. In this setting, flow-dependent transit times can implicitly be modeled by introducing appropriate capacities on the copies of an arc corresponding to different transit times. More precisely, these arc-capacities approximately model *inflow*-dependent transit times, that is, the situation when the transit time on an arc depends only on the current rate of flow into the arc. As discussed above, this model provides only a rough description of flow-dependent transit times in typical real-world situations.

Note that the model of Carey and Subrahmanian is defined for the multicommodity case. However, also when restricted to a single commodity, their model requires general linear programming techniques. Köhler, Langkau, and Skutella [23] present a refined time-expanded model that can be solved by standard network flow computations. While these algorithmic techniques are typically very efficient, the size of the time-expanded graph itself causes problems when the number of discrete time steps gets large. Building upon [7, 23], Hall, Langkau, and Skutella [18] present an improved analysis of these models which also works for the multicommodity case.

Results for fixed transit times. The problem concerning the size of the time-expanded graph was already addressed by Ford and Fulkerson [13, 14] when they studied the “maximal dynamic flow problem”: Given a directed network with capacities and fixed transit times on the arcs and a fixed time horizon T , send as much flow as possible from the source vertex s to the sink vertex t within time T . This problem can be solved by one max-flow computation on the corresponding time-expanded network. Notice, however, that the size of the time-expanded network is only pseudopolynomial in the size of the input.

Nevertheless, Ford and Fulkerson were able to show that the problem can be solved by essentially one min-cost flow computation on the given “static” network, where transit times of arcs are interpreted as cost coefficients. An optimal solution to this min-cost flow problem can be turned into a flow over time by decomposing it into flows on paths. The optimal flow over time starts to send flow on each path at time zero and repeats each so long as there is enough time left in the T periods for the flow along the path to arrive at the sink. Such a flow over time is called *temporally repeated*.

¹Time-expanded networks are often used for computing flows over time with fixed (integral) transit times on the arcs. Such a time-expanded network contains a copy of the node set of the underlying network for each discrete time step (building a *time layer*). Moreover, for each arc with transit time τ in the given network, there is a copy between each pair of time layers of distance τ in the time-expanded network.

In particular, this result of Ford and Fulkerson implies that the quickest s - t -flow problem with fixed transit times can be solved in polynomial time. Burkard, Dlaska, and Klinz [4] give a strongly polynomial algorithm for the quickest s - t -flow problem which is based on the parametric search method of Megiddo [25]. However, if costs are added the resulting minimum cost quickest flow problem is NP-hard [22]; the same hardness result holds for the case of multiple commodities, even without costs [17]. On the other hand, Hoppe [20] and Hoppe and Tardos [21] show that there is a nontrivial generalization of the result of Ford and Fulkerson to the case of multiple sources and sinks.

Recently, and prior to this work, Fleischer and Skutella [11] showed that the technique of Ford and Fulkerson can be generalized to yield approximate solutions to the NP-hard quickest flow problem with costs and also to the more general quickest multicommodity flow problem with costs. Their approach is based on the computation of a static *length-bounded flow*² which is then turned into a flow over time, similar to Ford and Fulkerson's result. This leads to an approximation algorithm with performance guarantee $2 + \varepsilon$ for the quickest multicommodity flow problem with costs. Moreover, Fleischer and Skutella [11, 12] introduce the concept of so-called *condensed* time-expanded networks. They are obtained by rounding transit times of arcs and scaling time such that the size of the resulting time-expanded network is polynomially bounded in the input size. As a consequence, one obtains fully polynomial-time approximation schemes for the quickest multicommodity flow problem with costs and for related problems.

New results and models. Our work is inspired by the results of Ford and Fulkerson [13, 14] and Fleischer and Skutella [11] on temporally repeated flows. Although their techniques cannot directly be translated to the more general setting of flow-dependent transit times, we can show that a similar approach yields provably good temporally repeated flows in this setting.

This result is based on the following fairly general model of flow-dependent transit times. We assume that, at each point in time, the uniform speed on an arc depends only on the amount of flow or *load* which is currently on that arc. This assumption captures, for example, the behavior of road traffic when an arc corresponds to a rather short street (notice that longer streets can be replaced by a series of short streets); similar transit time functions are used in standard traffic simulation systems.

For the case of steady state flows which do not vary over time, the constant load of an arc can be determined by the constant flow rate on the arc, i.e., by the number of flow units traversing the arc per time unit. Therefore, the transit time of an arc is a function of its flow rate in this case. Throughout the paper we assume that this dependency is given by an increasing and convex function. This assumption is satisfied for many applications; see, e.g., [32] and Figure 1. We propose an algorithm which is similar to the one of Ford and Fulkerson and thus also very efficient. However, since the transit times are no longer fixed, the linear min-cost flow problem considered by Ford and Fulkerson now turns into a *convex cost* flow problem. Under the assumptions on the transit time functions τ_e stated above, the resulting optimal static flow can be turned into a temporally repeated flow which needs at most twice as long as a quickest flow over time.

²A static (multicommodity) flow is called *length-bounded* if it can be decomposed into flows on paths whose transit times are bounded from above by a given value.

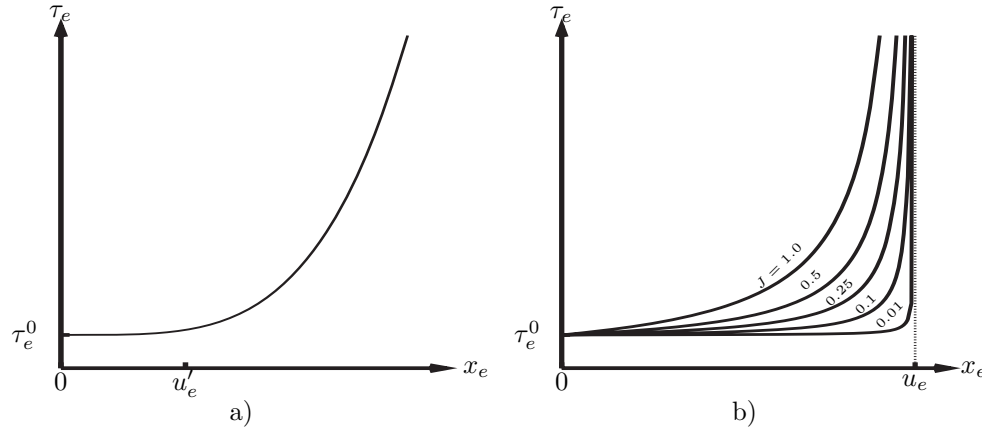


FIG. 1. a) For the case of static road traffic, the U.S. Bureau of Public Roads developed a simplified function describing the dependency of the transit time on the flow. This function is given by $\tau_e = \tau_e^0(1 + 0.15(x_e/u'_e)^4)$, where τ_e^0 is the free-flow transit time, x_e is the flow rate, and u'_e is the “practical capacity” of arc e . It follows from the given equation that the practical capacity of an arc is the flow rate at which the transit time is 15% higher than the free-flow transit time. b) Davidson proposed a function describing the transit time based on queuing theory considerations: $\tau_e = \tau_e^0(1 + Jx_e/(u_e - x_e))$. Again, τ_e^0 is the free-flow transit time (i.e., the transit time at zero flow), x_e is the flow rate, u_e is the capacity of the arc, and J is a parameter of the model. We depict the function for various choices of J . In contrast to the function depicted in a), Davidson’s function is obviously asymptotic to the capacity u_e of the arc. More details can be found in [32].

Finally, we show that the quickest flow problem under consideration is strongly NP-hard and even cannot be approximated with arbitrary precision in polynomial time, unless $P=NP$. We give reductions from the NP-hard problems PARTITION and SATISFIABILITY. Notice that the flow-over-time problem considered by Ford and Fulkerson is solvable in polynomial time and that the more general problems considered by Fleischer and Skutella [11, 12] possess fully polynomial-time approximation schemes and are therefore not strongly NP-hard, unless $P=NP$.

The paper is organized as follows. In section 2 we discuss the load-dependent transit time model which our results are based on. Our main result on the existence and efficient computation of provably good temporally repeated flows is presented in section 3. Section 4 contains results on the complexity of the quickest flow problem under consideration.

2. A model for load-dependent transit times. Our research is motivated by traffic routing problems. In this application, the transit time τ_e of an arc (street) e is typically given for the case of static flows, that is, flows which do not vary over time. We thus interpret $\tau_e(x_e)$ as the transit time on arc e for the static flow rate x_e , which is the number of flow units (cars etc.) traversing the arc per time unit; see Figure 1 for examples.

Unfortunately, for a flow varying over time, the latter model is only of limited use. First of all, it is not clear how to define the flow rate on an arc at a specific moment in time; there are several possibilities depending on where exactly it is measured. For example, the *inflow rate* is measured at the tail and the *outflow rate* is measured at the head of an arc; moreover, a flow rate can also be determined at any other position on an arc. Second, in real-world applications, it is not reasonable to assume

that the transit time of an arc depends only on the current flow rate on this arc, no matter where it is measured. For example, even if the number of cars which currently enter a street is small compared to the capacity of the street, the transit time might nevertheless be huge due to traffic congestion caused by a large number of cars which have entered the street earlier.

2.1. Description of the model. In what follows, we assume that we are given a network and for every arc of the network there is a function τ_e which describes the dependency of the arc's transit time on the *static* flow rate. Since we are interested in the general setting where the flow on an arc may vary over time, we have to come up with a model which enables us to at least approximately determine transit times for flows over time. The small example of a congested street discussed above suggests that the transit time on an arc depends on its current load, which is the amount of flow (number of cars) currently on that arc. We thus assume in our model that

- (i) at each point in time, the entire flow on an arc travels with uniform speed;
- (ii) this speed depends only on the current load of that arc.

It is clear that this model is, of course, not able to capture all properties of traffic flows. One such limitation is, for example, the incapability of mapping tailback effects properly, i.e., traffic that is delayed on some arc from x to y because of a traffic jam on some arc going out of y . Including effects like this causes severe complications and considerably weakens the model's efficiency in terms of computing good or even optimal solutions. The approximation algorithm that we propose in section 3 relies on methods from the area of classical network flow theory. In this context, the independence of flows on different arcs (apart from flow conservation, of course) is a crucial feature.

Before we discuss the implications of the above assumptions for flows varying over time, we first consider the simpler case of static flows. If we let y_e denote the load of arc e and x_e its flow rate, it is easy to see that, for a static flow, the following relation holds:

$$(1) \quad y_e = x_e \tau_e(x_e).$$

Observation 2.1. If the function τ_e is monotonically increasing and convex (as, e.g., in Figure 1), then, in a static flow, the flow rate x_e is a strictly increasing and concave function of the load y_e of arc e .

Proof. Since both τ_e and the identity are nonnegative, monotonically increasing, and convex, it follows from (1) that the load is also a nonnegative, strictly increasing, and convex function of the flow rate x_e . Thus, the inverse function exists and is strictly increasing and concave. $\square$

It follows from the proof that Observation 2.1 already holds under the weaker assumption that the function $x_e \mapsto x_e \tau_e(x_e)$ is strictly increasing and convex. In the remainder of the paper we assume that this weaker assumption holds for all arcs $e \in E$.

Assumption 2.2. For all arcs $e \in E$, the function $x_e \mapsto x_e \tau_e(x_e)$ is nonnegative, strictly increasing, and convex.

Observation 2.1 yields that, for the case of static flows, the transit time τ_e can also be interpreted as a function of the load y_e ; to avoid ambiguity, it is then denoted by $\hat{\tau}_e(y_e)$. Notice that

$$(2) \quad \tau_e(x_e) = \hat{\tau}_e(y_e)$$

if the flow rate x_e and the load y_e satisfy (1). The function $\hat{\tau}_e$ can now be used in order to model the setting of flows over time with load-dependent transit times.

A flow over time on arc e with time horizon T can be described by its flow rate $f_e : (0, T] \rightarrow \mathbb{R}^+$. The considerations and results in the remainder of this paper do not depend on the precise definition of f_e but work for all possible types of flow rates (like, e.g., inflow or outflow rate etc.). In the flow model given by assumptions (1) and (2), the mutual dependency of the flow rate $f_e : (0, T] \rightarrow \mathbb{R}^+$ and the load $\ell_e : (0, T] \rightarrow \mathbb{R}^+$ is implicitly given as follows: At any point in time θ , the speed of the flow on arc e is proportional to the inverse of the “current transit time” $\hat{\tau}_e(\ell_e(\theta))$. This concludes the precise description of our model of load-dependent transit times on one arc.

An s - t -flow over time with load-dependent transit times is given by flow rate functions $(f_e)_{e \in E}$ satisfying the following constraints:

- (i) $0 \leq f_e(\theta) \leq u_e$ for all $e \in E$ and $\theta \in (0, T]$ (capacity constraints);
- (ii) for every node $v \neq s$ and every point in time $\theta \in (0, T]$, the total amount of flow that has arrived in v until time θ is an upper bound on the total amount of flow that has left v until time θ (flow conservation constraints);
- (iii) equality holds in (ii) for $\theta = T$ and all $v \in V \setminus \{s, t\}$; moreover, $\ell_e(T) = 0$ for all $e \in E$ (i.e., all flow must have arrived at the sink at time T).

Notice that it is a nontrivial task to check whether flow conservation constraints (ii) are fulfilled for given inflow rate functions $(f_e)_{e \in E}$; the outflow rate of an arc is not a simple function of its inflow rate. Flow entering arc e at time ν arrives at the head of e at time

$$(3) \quad \inf \left\{ \xi \mid \int_{\nu}^{\xi} \frac{1}{\hat{\tau}_e(\ell_e(\theta))} d\theta \geq 1 \right\}.$$

This follows since, by definition of our flow model, $1/\hat{\tau}_e(\ell_e(\theta))$ is proportional to the speed on arc e at time θ .

2.2. Basic results and observations. We are interested in two basic characteristics of flows over time on an arc e . The total transit time is the total amount of time spent by all units of flow on that arc. If we think of cars driving along a street, it is the sum of the individual transit times of all cars. Formally, the total transit time is the integral of the load of arc e over time. Thus every unit of flow accounts to that particular arc as long as it contributes to the load of this arc. This is given by

$$\int_0^T \ell_e(\theta) d\theta.$$

The total amount of flow shipped through arc e is the integral over the flow rate $f_e(\theta)$, $0 < \theta \leq T$. This value can also be written in terms of the load $\ell_e(\theta)$:

$$(4) \quad \int_0^T f_e(\theta) d\theta = \int_0^T \frac{\ell_e(\theta)}{\hat{\tau}_e(\ell_e(\theta))} d\theta.$$

We give an intuitive interpretation of (4): If $f_e(\theta)$ denotes the inflow rate into arc e at time θ , the left-hand side simply counts the number of cars entering the street e over time. In contrast, the right-hand side counts at every point in time the number of cars currently traveling on e weighted by the inverse of the “current transit time” on e . In particular, the overall contribution of every car to the right-hand side is one which yields the claimed equation.

Notice, however, that (4) does not hold pointwise (with the integrals removed) since the load $\ell_e(\theta_0)$ at some point in time θ_0 does not only depend on the current flow rate $f_e(\theta_0)$ but also on the whole history $f_e(\theta)$, $0 < \theta \leq \theta_0$.

For an arbitrary flow over time on arc e with time horizon T and flow rate $f_e(\theta)$, for $\theta \in (0, T]$, we define a corresponding static flow with flow rate

$$(5) \quad x_e := \frac{1}{T} \int_0^T f_e(\theta) d\theta$$

and load $y_e := x_e \tau_e(x_e)$ according to (1). We refer to this static flow as the *average rate flow* corresponding to the flow over time given by f_e (or ℓ_e).

LEMMA 2.3. *If the function τ_e is monotonically increasing and convex, the total transit time of a flow over time with time horizon T , flow rate f_e , and load ℓ_e is at least as big as the total transit time of the corresponding static average rate flow over T time units, that is,*

$$(6) \quad \int_0^T \ell_e(\theta) d\theta \geq T y_e.$$

Proof. The left-hand side of (6) motivates the consideration of another static flow with load

$$\tilde{y}_e := \frac{1}{T} \int_0^T \ell_e(\theta) d\theta$$

and flow rate

$$\tilde{x}_e := \frac{\tilde{y}_e}{\hat{\tau}_e(\tilde{y}_e)}.$$

We refer to this static flow as the *average load flow* corresponding to the flow over time given by f_e (or ℓ_e). Thus, (6) can now be rewritten as $\tilde{y}_e \geq y_e$. Since the load of a static flow is a monotonically increasing function of its flow rate, it suffices to show that $\tilde{x}_e \geq x_e$. Using (4) and (5), this inequality can be rewritten as

$$(7) \quad \frac{\tilde{y}_e}{\hat{\tau}_e(\tilde{y}_e)} \geq \frac{1}{T} \int_0^T \frac{\ell_e(\theta)}{\hat{\tau}_e(\ell_e(\theta))} d\theta.$$

It follows from Observation 2.1 that the function $\xi \mapsto \xi/\hat{\tau}_e(\xi)$ is concave. Thus, inequality (7) is a result of Jensen's inequality. This concludes the proof. $\square$

In other words, Lemma 2.3 says that, for an arbitrary flow over time on arc e , the corresponding average load flow has a higher value than the corresponding average rate flow.

3. An approximation algorithm.

3.1. A related static flow problem. In order to determine a flow over time which satisfies demand D in close to optimal time, we consider the following static maximum flow problem with bounded convex cost. In this problem, the cost of flow x_e on arc e is $x_e \tau_e(x_e)$ and the total cost must not exceed D . More formally, the problem

can be written as follows:

$$\begin{aligned}
 & \max \quad \sum_{e \in \delta^-(t)} x_e - \sum_{e \in \delta^+(t)} x_e \\
 & \text{s. t.} \quad \sum_{e \in \delta^-(v)} x_e - \sum_{e \in \delta^+(v)} x_e = 0 \quad \text{for all } v \in V \setminus \{s, t\}, \\
 & \quad \sum_{e \in E} x_e \tau_e(x_e) \leq D, \\
 & \quad 0 \leq x_e \leq u_e \quad \text{for all } e \in E.
 \end{aligned}$$

Here, $\delta^+(v)$ and $\delta^-(v)$ denote the set of arcs leaving and entering node v , respectively.

LEMMA 3.1. *If there exists a flow over time f with load-dependent transit times which sends D units of flow from s to t within time T , then there exists a static flow x of value at least D/T for the static flow problem stated above.*

Proof. Consider the static average rate flow $x = (x_e)_{e \in E}$ as defined in (5) which corresponds to the given flow over time with flow rate $(f_e)_{e \in E}$ and load $(\ell_e)_{e \in E}$. By the assumption on the amount of flow sent, the value of the static flow x is D/T . Moreover, since the given flow over time obeys capacity constraints, the same is true for the average rate flow x . It thus remains to show that the cost of x is bounded by D .

Since in the flow over time f every unit of flow needs at most time T to travel through the network from s to t , the corresponding total transit time is bounded by DT , that is,

$$\sum_{e \in E} \int_0^T \ell_e(\theta) d\theta \leq DT.$$

By Lemma 2.3 and (1), this concludes the proof. $\square$

An optimal integral solution to the static constrained maximum flow problem stated above can be computed in polynomial time, for example by the capacity scaling algorithm of Ahuja and Orlin [1]. We argue that the difference between the value OPT of an optimal fractional flow and the value OPT' of an optimal integral flow is at most $|E|$, i.e., the number of arcs: It is a classical result from network flow theory that an optimal fractional flow can be decomposed into flows on at most $|E|$ paths. Rounding down the flow value on all of these paths to the nearest integer decreases the flow value by at most $|E|$ and yields a feasible integral flow. Thus, $\text{OPT} - \text{OPT}' \leq |E|$.

By appropriately scaling capacities, flow values, and the demand value D (i.e., multiplying all of these parameters by a large enough constant M), we can make sure that $\text{OPT} \geq D/T \geq |E|/\varepsilon$, for any given $\varepsilon > 0$. As a consequence, the value of an optimal integral solution fulfills $\text{OPT}' \geq \text{OPT} - |E| \geq (1 - \varepsilon)\text{OPT}$.

Summarizing, we can compute a static flow x of value at least $(1 - \varepsilon)\text{OPT}$ and cost at most D in polynomial time, where OPT is the value of an optimal static flow.

LEMMA 3.2. *If there is a flow over time with load-dependent transit times which sends D units of flow from s to t within time T , then a static flow x of value at least $(1 - \varepsilon)D/T$ and cost at most D can be computed in polynomial time.*

Proof. The result follows from Lemma 3.1 and the discussion above. $\square$

3.2. A 2-approximation algorithm. Although x is a static flow, it contains some structural information on how to construct a provably good flow over time. We

decompose x into a sum of static path-flows on a set of s - t -paths $\mathcal{P}$. The flow value on path $P \in \mathcal{P}$ is denoted by x_P such that, for each arc $e \in E$,

$$(8) \quad x_e = \sum_{P \in \mathcal{P} : e \in P} x_P.$$

Notice that we can assume without loss of generality that no cycles are needed in the flow decomposition; otherwise, the solution x can be improved by canceling flow on those cycles. Moreover, it is well known that the number of paths in $\mathcal{P}$ can be bounded by the number of arcs $|E|$.

For the case of fixed transit times, a temporally repeated flow over time with time horizon T' can be generated from a path-decomposition of x by starting each path-flow at time zero, and repeatedly sending flow on each of them, as long as there is enough time left in the T' periods for the flow along the path to arrive at the sink. In what follows, we argue that basically the same approach can be used in the setting of load-dependent transit times. Here, however, some care has to be taken in order to avoid undesirable congestion. If all transit times are fixed, it follows from (8) that at any point in time the total flow rate on arc e in the temporally repeated flow over time is bounded by x_e . This is no longer true if transit times on arcs may vary over time. We use the following trick to solve this problem.

As soon as we have computed the static flow x , we assume that the transit time of every arc $e \in E$ in the network is fixed to $\tau_e(x_e)$. This assumption is justified if we can assure that the rate of flow into arc e is always bounded by x_e and thus its load never exceeds $x_e \tau_e(x_e)$. In this case, we can enforce the constant transit time $\tau_e(x_e)$ by introducing waiting times at the head node v of arc e in order to compensate for a potentially smaller transit time on that arc. Thereby we emulate the fixed transit time $\tau_e(x_e)$ and, at the same time, make sure that the rate of flow into every arc $e' \in \delta^+(v)$ also stays below $x_{e'}$. The same technique is used in [11] in order to round up transit times of arcs.

Under these assumptions, the amount of flow that can be sent within time horizon T' over path $P \in \mathcal{P}$ of length $\tau_P := \sum_{e \in P} \tau_e(x_e)$ with $\tau_P \leq T'$ is $x_P (T' - \tau_P)$. Therefore, the total amount of flow which we can send on paths $\mathcal{P}$ within time horizon T' is

$$(9) \quad d(T') := \sum_{P \in \mathcal{P} : \tau_P \leq T'} x_P (T' - \tau_P).$$

From this expression one can easily determine the minimal T' that is needed to satisfy demand D : Simply order the paths in the set $\mathcal{P}$ by nondecreasing lengths $\tau_{P_1} \leq \tau_{P_2} \leq \dots \leq \tau_{P_k}$ and observe that the function d is affine linear and increasing within every interval $[\tau_{P_i}, \tau_{P_{i+1}}]$, $1 \leq i < k$.

THEOREM 3.3. *If there is a flow over time with load-dependent transit times which sends D units of flow from s to t within time T , then there exists a temporally repeated flow satisfying demand D within time horizon at most $2T$. Moreover, for every $\varepsilon > 0$, one can compute a temporally repeated flow in polynomial time which satisfies demand D within time horizon at most $(2 + \varepsilon)T$.*

Proof. By Lemma 3.2, we can compute a static flow x of value at least $(1 - \varepsilon/3)D/T$ and cost at most D in polynomial time, for any $\varepsilon > 0$ (in what follows, we assume that $\varepsilon \leq 1$). Moreover, by Lemma 3.1, there even exists such a flow for $\varepsilon = 0$.

We decompose x into path-flows as discussed above and get

$$\begin{aligned}
 d((2 + \varepsilon)T) &\stackrel{(9)}{=} \sum_{P \in \mathcal{P} : \tau_P \leq (2 + \varepsilon)T} x_P ((2 + \varepsilon)T - \tau_P) \\
 &\geq \sum_{P \in \mathcal{P}} x_P ((2 + \varepsilon)T - \tau_P) \\
 &= (2 + \varepsilon)T \sum_{P \in \mathcal{P}} x_P - \sum_{e \in E} x_e \tau_e(x_e) \\
 &\geq (2 + \varepsilon)(1 - \varepsilon/3)D - D \\
 &\geq D.
 \end{aligned}$$

Since the function d is increasing, this concludes the proof. $\square$

The presented analysis of the algorithm is similar to that of the $(2 + \varepsilon)$ -approximation algorithm in [11]. The common underlying idea is to prove that averaging a flow over time with time horizon T yields a feasible static flow which can then again be turned into a flow over time with time horizon $2T$. However, while Fleischer and Skutella use the fact that the average static flow is length-bounded by T , it is easy to observe that this is no longer true if we allow flow-dependent transit times on the arcs.

3.3. Lower bounds on the performance of the algorithm. It can be shown that our analysis is tight, even for the case of a network consisting of only one uncaped arc e from s to t . In this example, the transit time of arc e is given by

$$\tau_e(x_e) := \begin{cases} 2 & \text{if } x_e \leq 1, \\ \infty & \text{if } x_e > 1. \end{cases}$$

From (2) one can determine the corresponding load-dependent function

$$\hat{\tau}(y_e) = \begin{cases} 2 & \text{if } y_e \leq 2, \\ \infty & \text{if } y_e > 2. \end{cases}$$

Thus, a quickest flow over time sending demand $D := 2$ from s to t needs exactly 2 time units since it can put the 2 units of flow onto arc e at time 0 such that they arrive at time 2. However, an optimal solution to the static maximum flow problem stated at the beginning of this section sets the flow rate x_e to 1. Thus, in the resulting temporally repeated flow the last piece of flow leaves s only at time 2 and therefore does not arrive before time 4.

In this example, the gap of 2 between the optimal solution and the solution arising from the static maximum flow problem obviously originates from the incapability of the static flow problem to capture the possibility of sending flow at a very high flow rate for a very short period of time.

Although the optimal flow over time is temporally repeated here, a slight modification of the instance shows that every temporally repeated flow can be bad compared to an optimal flow over time. If we double the demand value to $D := 4$, an optimal flow over time needs 4 time units since it sends 2 packets of flow, each containing 2 units of flow, at time 0 and time 2, respectively. However, in a temporally repeated flow, the flow rate cannot be chosen bigger than 1 since otherwise the arc would

become completely congested as soon as there are more than 2 units of flow on it. Therefore, every temporally repeated flow needs at least 6 time units and is thus at least a factor of $3/2$ away from the optimal value 4.

3.4. A bicriteria generalization. The results in Theorem 3.3 can be generalized in the following direction. One can decrease the factor of 2 in time at the cost of a decrease of the amount of flow that can be delivered. This leads to the following bicriteria results.

COROLLARY 3.4. *If there is a flow over time which sends D units of flow from s to t within time T , then, for every $\alpha \geq 1$, there exists a temporally repeated flow satisfying demand $(\alpha - 1)D$ within time horizon at most αT . Moreover, for every $\varepsilon > 0$, one can compute a temporally repeated flow in polynomial time which satisfies demand $(\alpha - 1)D$ within time horizon at most $(\alpha + \varepsilon)T$.*

The proof of Corollary 3.4 is almost identical to the proof of Theorem 3.3 (replacing 2 by α).

3.5. An alternative view on the 2-approximation. We close this section with the following alternative view of the presented approximation result which highlights its close relation to the algorithm of Ford and Fulkerson. A temporally repeated flow with load-dependent transit times and time horizon T can be obtained from a solution to the following static convex cost flow problem:

$$\begin{aligned} \max \quad & T \left(\sum_{e \in \delta^-(t)} x_e - \sum_{e \in \delta^+(t)} x_e \right) - \sum_{e \in E} x_e \tau_e(x_e) \\ \text{s. t.} \quad & \sum_{e \in \delta^-(v)} x_e - \sum_{e \in \delta^+(v)} x_e = 0 && \text{for all } v \in V \setminus \{s, t\}, \\ & 0 \leq x_e \leq u_e && \text{for all } e \in E. \end{aligned}$$

For the special case of fixed transit times, this is exactly the static flow problem considered by Ford and Fulkerson. It is easy to observe that the value of the resulting temporally repeated flow is equal to the value of the objective function (compare (9)). For the special case of fixed transit times, Ford and Fulkerson showed that this temporally repeated flow over time is maximal.

This is no longer true in the setting with load-dependent transit times. However, it follows from Lemma 3.1 that for $T = 2T^*$ (where T^* denotes the makespan of a quickest flow) this value is at least D . Thus, we get an alternative $(2 + \varepsilon)$ -approximation algorithm for the quickest flow problem by embedding this approach into a binary search framework for T . The main drawback of this alternative algorithm is that it requires more than one convex cost flow computation. Moreover, the simple example discussed above also shows that its performance guarantee is not better than 2.

4. Complexity results. While the corresponding problem with fixed transit times can be solved efficiently [4, 13, 14], the quickest s - t -flow problem with load-dependent transit times is NP-hard. We start with a simple reduction from the well-known NP-complete PARTITION problem.

PARTITION

Given: A set of n items with associated sizes $a_1, \dots, a_n \in \mathbb{N}$ such that $\sum_{i=1}^n a_i = 2L$ for some $L \in \mathbb{N}$.

Question: Is there a subset $I \subset \{1, \dots, n\}$ with $\sum_{i \in I} a_i = L$?

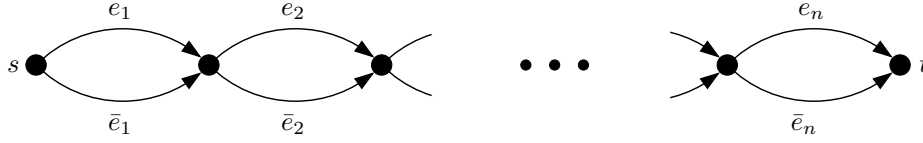


FIG. 2. Reduction of the problem PARTITION to a flow-over-time problem with load-dependent transit times.

Given an instance of PARTITION, we construct a network with load-dependent transit times as follows. Take a chain of length n where each link $i = 1, \dots, n$ consists of a pair of two parallel arcs e_i and $\bar{e}_i$ (see Figure 2) with the following transit times:³

$$\tau_{e_i}(x) := \begin{cases} 2L & \text{if } x \leq 1/(2L), \\ \infty & \text{if } x > 1/(2L), \end{cases} \quad \hat{\tau}_{e_i}(y) = \begin{cases} 2L & \text{if } y \leq 1, \\ \infty & \text{if } y > 1, \end{cases}$$

and

$$\tau_{\bar{e}_i}(x) := \begin{cases} 2L + a_i & \text{if } x \leq 1/(2L + a_i), \\ \infty & \text{if } x > 1/(2L + a_i), \end{cases} \quad \hat{\tau}_{\bar{e}_i}(y) = \begin{cases} 2L + a_i & \text{if } y \leq 1, \\ \infty & \text{if } y > 1. \end{cases}$$

The task is to send $D := 2$ units of flow from s to t .

LEMMA 4.1. *There exists a flow over time which sends two units of flow from s to t in time $(2n + 1)L$ if and only if the underlying instance of PARTITION is a “yes”-instance.*

Proof. If: Let I be a subset of $\{1, \dots, n\}$ such that $\sum_{i \in I} a_i = L$. The flow is sent in two packets, each containing one flow unit. The packets use two arc-disjoint paths of length (transit time) $(2n + 1)L$ that are induced by the subset I and its complement, respectively.

Only if: It follows from the definition of the transit time functions that there can never be more than one unit of flow on any arc. Therefore the construction of the network yields that no arc can be traversed by more than one flow unit unless the flow takes at least $(2n + 2)L$ units of time. As a consequence, in a flow over time with makespan $(2n + 1)L$, every arc is traversed by exactly one flow unit and the total transit time is thus $\sum_{i=1}^n (2L + 2L + a_i) = 2(2n + 1)L$. In particular, an arbitrary piece of flow needs exactly $(2n + 1)L$ units of time to travel from s to t and the corresponding path therefore induces a subset $I \subset \{1, \dots, n\}$ with $\sum_{i \in I} a_i = L$. $\square$

So far we have shown that the problem under consideration is NP-hard in the weak sense. Next we give a more involved reduction from the NP-complete SATISFIABILITY problem.

SATISFIABILITY

Given: n Boolean variables $z_1, \dots, z_n$ and m disjunctive clauses $C_1, \dots, C_m$.

Question: Does there exist a truth-assignment which satisfies all clauses?

The aim of the following reduction is to create a gap between those instances of the flow problem corresponding to “yes”-instances of SATISFIABILITY and those

³Notice that the functions τ_{e_i} and $\tau_{\bar{e}_i}$ are only given to derive the load-dependent functions $\hat{\tau}_{e_i}$ and $\hat{\tau}_{\bar{e}_i}$ that are then actually used in our flow model.

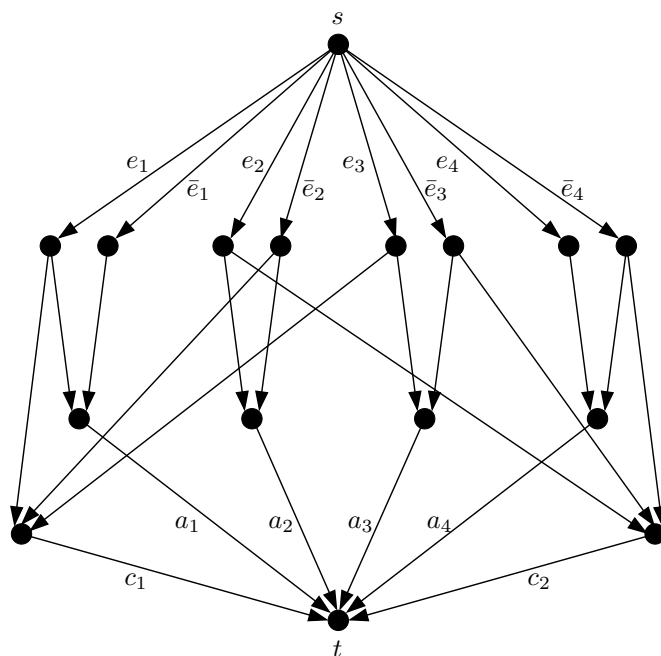


FIG. 3. Reduction of the problem SATISFIABILITY to a flow-over-time problem with load-dependent transit times. In the depicted example, the instance of SATISFIABILITY contains four variables z_1, z_2, z_3 , and z_4 and two clauses $C_1 = (z_1 \vee \neg z_2 \vee z_3)$ and $C_2 = (z_2 \vee \neg z_3 \vee \neg z_4)$.

corresponding to “no”-instances. This gap then yields a nonapproximability result for the flow problem under consideration.

For every variable z_i of the SATISFIABILITY instance, we introduce two outgoing arcs e_i and $\bar{e}_i$ from the source s and one ingoing arc a_i to the sink t . Moreover, for every clause C_j , there is an ingoing arc c_j to the sink. There are additional arcs from the head of e_i and $\bar{e}_i$ to the tail of a_i . Finally, for every variable z_i occurring unnegated (negated) in clause C_j , there is an arc from the head of e_i ($\bar{e}_i$) to the tail of c_j .

An illustration of this construction is given in Figure 3. All arcs are uncapacitated and have the following transit times (let $0 < \varepsilon \ll 1/(9m)$ be a small constant):

$$\begin{aligned} \tau_{e_i}(x) &:= \tau_{\bar{e}_i}(x) := \begin{cases} \frac{1}{1-x} & \text{if } x < 1, \\ \infty & \text{if } x \geq 1, \end{cases} & \hat{\tau}_{e_i}(y) &= \hat{\tau}_{\bar{e}_i}(y) = 1 + y, \\ \tau_{a_i}(x) &:= \begin{cases} 2 & \text{if } x \leq 1/2, \\ \infty & \text{if } x > 1/2, \end{cases} & \hat{\tau}_{a_i}(y) &= \begin{cases} 2 & \text{if } y \leq 1, \\ \infty & \text{if } y > 1, \end{cases} \\ \tau_{c_j}(x) &:= \begin{cases} 3 & \text{if } x \leq \varepsilon/3, \\ \infty & \text{if } x > \varepsilon/3, \end{cases} & \hat{\tau}_{c_j}(y) &= \begin{cases} 3 & \text{if } y \leq \varepsilon, \\ \infty & \text{if } y > \varepsilon. \end{cases} \end{aligned}$$

The transit times of all remaining arcs are fixed to 0.

LEMMA 4.2. *If the underlying instance of SATISFIABILITY is a “yes”-instance, then there exists a flow over time which sends $n + \varepsilon m$ units of flow from s to t in time $4 + \varepsilon m$. However, if it is a “no”-instance, then every flow over time needs at least $4 + 1/9$ units of time.*

Proof. Given a satisfying truth-assignment for the underlying instance of SATISFIABILITY, we construct a flow over time as follows. For every $i = 1, \dots, n$, if the variable z_i is set to true (false), then we route one unit of flow from s over $\bar{e}_i$ (e_i) and a_i to t . Since this is the only flow routed across these arcs, it will arrive at t at time 4.

For every clause C_j , $j = 1, \dots, m$, we choose a literal ℓ_j which is set to true. If ℓ_j is the unnegated (negated) variable z_i , then we route ε units of flow from s over e_i ($\bar{e}_i$) and c_j to t . Since at most εm units of flow are routed across e_i ($\bar{e}_i$) and exactly ε units of flow are routed across c_j , this flow arrives at t not later than at time $4 + \varepsilon m$. We have thus constructed a flow over time which sends $n + \varepsilon m$ units of flow from s to t in time $4 + \varepsilon m$.

On the other hand, we have to show that the existence of a flow over time with makespan less than $4 + 1/9$ yields a satisfying truth-assignment for the underlying instance of SATISFIABILITY. We first claim that, in such a flow, exactly one unit of flow is sent to t across arc a_i , for every $i = 1, \dots, n$, and exactly ε units of flow are sent over arc c_j , for every $j = 1, \dots, m$.

Notice that every unit of flow which is sent across a_i enters this arc between time 1 and $2 + 1/9$. Since the transit time of a_i is at least 2, these flow units can simultaneously be found on the arc at time $2 + 1/9$. The construction of $\hat{\tau}_{a_i}$ yields that the total amount of flow which is sent across a_i is bounded by 1. A similar argument shows that at most ε units of flow can be sent to t across arc c_j , for every $j = 1, \dots, m$. Since exactly $n + \varepsilon m$ units of flow are sent from s to t , these bounds are tight which proves the claim.

In what follows, we refer to the unit of flow sent across arc a_i as *commodity i* . For every $i = 1, \dots, n$, if at most one half of commodity i is sent across e_i , we set variable z_i to true; otherwise, we set it to false.⁴ It remains to show that this is a satisfying truth-assignment.

Consider some clause C_j and the flow which uses the corresponding arc c_j ; we refer to this flow as *commodity j* . Moreover, consider an arc e leaving the source s which is used by commodity j . By construction of the network, e corresponds to a literal (z_i or $\neg z_i$) of clause C_j . It therefore suffices to show that the amount of flow of commodity i which is sent across arc e is less than $1/2$. In this case, the corresponding literal in clause C_j is set to true such that the clause is fulfilled.

Notice that in order to arrive in time at the sink t , all flow of commodity j using arc e must arrive at the head of e before time $10/9$. In particular, arc e must not be congested too much in the time interval $[0, 10/9]$; that is, the “speed” $1/\hat{\tau}_e(\ell_e(\theta))$ has to be large enough such that flow of commodity j can arrive at the head of e before time $10/9$. More formally, this implies the following condition on the load $\ell_e(\theta)$ of arc e :

$$(10) \quad \int_0^{10/9} \frac{1}{\hat{\tau}_e(\ell_e(\theta))} d\theta \geq 1.$$

Notice that flow entering arc e at or after time 0 cannot arrive at the head of e before time $\inf\{\xi \mid \int_0^\xi 1/\hat{\tau}_e(\ell_e(\theta)) d\theta \geq 1\}$, by (3).

By contradiction, we assume that the amount d of flow of commodity i which is sent across arc e is at least $1/2$. In order to arrive in time at the sink t , all flow of

⁴In fact, as we will show below, we can assume that either e_i or $\bar{e}_i$ carries strictly less than one half unit of flow.

commodity i using arc e must arrive at the head of e before time $19/9$. This yields

$$(11) \quad \int_0^{19/9} \frac{\ell_{e,i}(\theta)}{\hat{\tau}_e(\ell_e(\theta))} d\theta \geq d,$$

where $\ell_{e,i}(\theta)$ denotes the load of commodity i on arc e at time θ (also compare (4)). Since $\ell_{e,i}(\theta) \leq \min\{\ell_e(\theta), d\}$, for all θ , and $d \geq 1/2$, we get

$$\int_{10/9}^{19/9} \frac{\ell_{e,i}(\theta)}{\hat{\tau}_e(\ell_e(\theta))} d\theta \leq \int_{10/9}^{19/9} \frac{\ell_{e,i}(\theta)}{1 + \ell_{e,i}(\theta)} d\theta \leq \int_{10/9}^{19/9} \frac{d}{1 + d} d\theta = \frac{d}{1 + d} \leq \frac{2d}{3}.$$

The second inequality is based on the fact that the function $\xi \mapsto \xi/(1 + \xi)$ is monotonically increasing in the interval $[0, \infty)$.

Together with (11), this yields

$$(12) \quad \int_0^{10/9} \frac{\ell_{e,i}(\theta)}{\hat{\tau}_e(\ell_e(\theta))} d\theta \geq \frac{d}{3}.$$

Putting together (10) and (12), we get the following contradiction:

$$\begin{aligned} \frac{10}{9} &= \int_0^{10/9} \frac{\ell_e(\theta)}{\hat{\tau}_e(\ell_e(\theta))} d\theta + \int_0^{10/9} \frac{1}{\hat{\tau}_e(\ell_e(\theta))} d\theta \\ &\geq \int_0^{10/9} \frac{\ell_{e,i}(\theta)}{\hat{\tau}_e(\ell_e(\theta))} d\theta + 1 \geq \frac{d}{3} + 1 \geq \frac{7}{6}. \end{aligned}$$

The first equality follows from the definition $\hat{\tau}_e(\ell_e(\theta)) = 1 + \ell_e(\theta)$. This concludes the proof. $\square$

As a consequence of Lemma 4.2 we get the following hardness result for the flow problem under consideration.

THEOREM 4.3. *The problem of finding a quickest flow over time with load-dependent transit times is strongly NP-hard and also APX-hard; i.e., there does not exist a polynomial-time approximation scheme, unless $P=NP$.*

Lemma 4.2 even yields a stronger nonapproximability result.

COROLLARY 4.4. *There does not exist an approximation algorithm with performance guarantee better than $37/36$, unless $P=NP$.*

Certainly, stronger bounds than the one stated in Corollary 4.4 can be obtained using the same reduction with a more careful choice of the crucial parameters. However, we did not pursue this idea further.

Notice that the load-dependent transit times used for the hardness results in this section and the negative results of the last section are artificial and probably not very realistic. However, similar results can be obtained for more natural transit time functions. Unfortunately, the analysis gets much more involved then.

5. Concluding remarks. The following interesting generalization of the quickest flow problem under consideration was pointed out by Lisa Fleischer (personal communication, January 2002). There are k commodities $i = 1, \dots, k$, each given by a source-sink pair (s_i, t_i) . The aim is to find a quickest multicommodity flow over time such that the sum of the flow values of all commodities is at least D . The approximation result described in section 3 can be generalized directly to this setting; the static flow problem stated at the beginning of section 3 is turned into a maximum multicommodity flow problem with cost bounded by D . The analysis remains essentially unchanged.

A problem closely related to the quickest flow problem is the maximum flow-over-time problem with fixed time horizon T . Unfortunately, Corollary 3.4 does not yield any useful result for this variant of the problem. On the other hand, the reduction from the problem PARTITION in section 4 shows that the problem cannot be approximated with performance guarantee strictly better than $1/2$, unless $P=NP$.

Acknowledgments. The authors are grateful to Lisa Fleischer for numerous helpful comments on this manuscript. Moreover, the second author is much indebted to Lisa Fleischer for interesting and helpful discussions on the subject of flows over time in general, which also partly inspired the work at hand. Finally, the authors want to thank Arie Tamir for pointing out the connection to Jensen's inequality in the proof of Lemma 2.3. Interesting comments that helped to improve the presentation of the paper were given by an anonymous referee.

REFERENCES

- [1] R. K. AHUJA AND J. B. ORLIN, *A capacity scaling algorithm for the constrained maximum flow problem*, Networks, 25 (1995), pp. 89–98.
- [2] J. E. ARONSON, *A survey of dynamic network flows*, Ann. Oper. Res., 20 (1989), pp. 1–66.
- [3] M. BEN-AKIVA, H. KOUTSOPOULOS, R. MISHALANI, AND Q. YANG, *Simulation laboratory for evaluating dynamic traffic management systems*, ASCE Journal of Transportation Engineering, 123 (1997), pp. 283–289.
- [4] R. E. BURKARD, K. DLASKA, AND B. KLINZ, *The quickest flow problem*, Z. Oper. Res., 37 (1993), pp. 31–58.
- [5] M. CAREY, *A constraint qualification for a dynamic traffic assignment model*, Transportation Science, 20 (1986), pp. 55–58.
- [6] M. CAREY, *Optimal time-varying flows on congested networks*, Oper. Res., 35 (1987), pp. 58–69.
- [7] M. CAREY AND E. SUBRAHMANIAN, *An approach to modelling time-varying flows on congested networks*, Transportation Research B, 34 (2000), pp. 157–183.
- [8] H.-K. CHEN, *Dynamic Travel Choice Models: A Variational Inequality Approach*, Springer, Berlin, 1999.
- [9] H.-K. CHEN AND C.-F. HSUEH, *A model and an algorithm for the dynamic user-optimal route choice problem*, Transportation Research B, 32 (1998), pp. 219–234.
- [10] S. P. DIRKSE AND M. C. FERRIS, *Traffic modeling and variational inequalities using GAMS*, in Operations Research and Decision Aid Methodologies in Traffic and Transportation Management, NATO Adv. Sci. Inst. Ser. F Comput. Systems Sci. 166, P. L. Toint, M. Labbe, K. Tanczos, and G. Laporte, eds., Springer, Berlin, 1998, pp. 136–163.
- [11] L. FLEISCHER AND M. SKUTELLA, *The quickest multicommodity flow problem*, in Integer Programming and Combinatorial Optimization, Lecture Notes in Comput. Sci. 2337, W. J. Cook and A. S. Schulz, eds., Springer, Berlin, 2002, pp. 36–53.
- [12] L. FLEISCHER AND M. SKUTELLA, *Minimum cost flows over time without intermediate storage*, in Proceedings of the Fourteenth Annual ACM–SIAM Symposium on Discrete Algorithms, Baltimore, MD, 2003, pp. 66–75.
- [13] L. R. FORD AND D. R. FULKERSON, *Constructing maximal dynamic flows from static flows*, Oper. Res., 6 (1958), pp. 419–433.
- [14] L. R. FORD AND D. R. FULKERSON, *Flows in Networks*, Princeton University Press, Princeton, NJ, 1962.
- [15] T. L. FRIESZ, D. BERNSTEIN, T. E. SMITH, R. L. TOBIN, AND B. W. WIE, *A variational inequality formulation of the dynamic network user equilibrium problem*, Oper. Res., 41 (1993), pp. 179–191.
- [16] N. GARTNER, C. J. MESSER, AND A. K. RATHI, EDS., *Traffic Flow Theory: A State of the Art Report*, 1997, <http://www.tfhrc.gov/its/tft/tft.htm>.
- [17] A. HALL, S. HIPPLER, AND M. SKUTELLA, *Multicommodity flows over time: Efficient algorithms and complexity*, in Automata, Languages and Programming, Lecture Notes in Comput. Sci. 2719, J. C. M. Baeten, J. K. Lenstra, J. Parrow, and G. J. Woeginger, eds., Springer, Berlin, 2003, pp. 397–409.
- [18] A. HALL, K. LANGKAU, AND M. SKUTELLA, *An FPTAS for quickest multicommodity flows with inflow-dependent transit times*, in Approximation, Randomization, and Combinatorial

- Optimization, Lecture Notes in Comput. Sci. 2764, S. Arora, K. Jansen, J. D. P. Rolim, and A. Sahai, eds., Springer, Berlin, 2003, pp. 71–82.
- [19] H. W. HAMACHER AND S. A. TJANDRA, *Mathematical modelling of evacuation problems: A state of the art*, in Pedestrian and Evacuation Dynamics, M. Schreckenberg and S. D. Sharma, eds., Springer, Berlin, 2002, pp. 227–266.
 - [20] B. HOPPE, *Efficient Dynamic Network Flow Algorithms*, Ph.D. thesis, Cornell University, Ithaca, NY, 1995.
 - [21] B. HOPPE AND E. TARDOS, *The quickest transshipment problem*, Math. Oper. Res., 25 (2000), pp. 36–62.
 - [22] B. KLINZ AND G. J. WOEGINGER, *Minimum cost dynamic flows: The series-parallel case*, in Integer Programming and Combinatorial Optimization, Lecture Notes in Comput. Sci. 920, E. Balas and J. Clausen, eds., Springer, Berlin, 1995, pp. 329–343.
 - [23] E. KÖHLER, K. LANGKAU, AND M. SKUTELLA, *Time-expanded graphs for flow-dependent transit times*, in Algorithms—ESA '02, Lecture Notes in Comput. Sci. 2461, R. Möhring and R. Raman, eds., Springer, Berlin, 2002, pp. 599–611.
 - [24] H. S. MAHMASSANI AND S. PEETA, *System optimal dynamic assignment for electronic route guidance in a congested traffic network*, in Urban Traffic Networks. Dynamic Flow Modelling and Control, N. H. Gartner and G. Improta, eds., Springer, Berlin, 1995, pp. 3–37.
 - [25] N. MEGIDDO, *Combinatorial optimization with rational objective functions*, Math. Oper. Res., 4 (1979), pp. 414–424.
 - [26] D. K. MERCHANT AND G. L. NEMHAUSER, *A model and an algorithm for the dynamic traffic assignment problems*, Transportation Science, 12 (1978), pp. 183–199.
 - [27] D. K. MERCHANT AND G. L. NEMHAUSER, *Optimality conditions for a dynamic traffic assignment model*, Transportation Science, 12 (1978), pp. 200–207.
 - [28] K. NAGEL, J. ESSER, AND M. RICKERT, *Large-scale traffic simulations for transportation planning*, in Annual Review of Computational Physics VII, D. Stauffer, ed., World Scientific, Singapore, 2000, pp. 151–202.
 - [29] W. B. POWELL, P. JAILLET, AND A. ODONI, *Stochastic and dynamic networks and routing*, in Network Routing, Handbooks Oper. Res. Management Sci. 8, M. O. Ball, T. L. Magnanti, C. L. Monma, and G. L. Nemhauser, eds., North-Holland, Amsterdam, 1995, pp. 141–295.
 - [30] I. PRIGOGINE AND R. HERMAN, *Kinetic Theory of Vehicular Traffic*, Elsevier Science B. V., Amsterdam, 1971.
 - [31] B. RAN AND D. E. BOYCE, *Modelling Dynamic Transportation Networks*, Springer, Berlin, 1996.
 - [32] Y. SHEFFI, *Urban Transportation Networks*, Prentice-Hall, Englewood Cliffs, NJ, 1985.

ON USING THE ELASTIC MODE IN NONLINEAR PROGRAMMING APPROACHES TO MATHEMATICAL PROGRAMS WITH COMPLEMENTARITY CONSTRAINTS*

MIHAI ANITESCU†

Abstract. We investigate the possibility of solving mathematical programs with complementarity constraints (MPCCs) using algorithms and procedures of smooth nonlinear programming. Although MPCCs do not satisfy a constraint qualification, we establish sufficient conditions for their Lagrange multiplier set to be nonempty. MPCCs that have nonempty Lagrange multiplier sets and satisfy the quadratic growth condition can be approached by the elastic mode with a bounded penalty parameter. In this context, the elastic mode transforms MPCC into a nonlinear program with additional variables that has an isolated stationary point and local minimum at the solution of the original problem, which in turn makes it approachable by sequential quadratic programming (SQP) algorithms. One such algorithm is shown to achieve local linear convergence once the problem is relaxed. Under stronger conditions, we also prove superlinear convergence to the solution of an MPCC using an adaptive elastic mode approach for an SQP algorithm recently analyzed in an MPCC context in [R. Fletcher, S. Leyffer, S. Sholtes, and D. Ralph, *Local Convergence of SQP Methods for Mathematical Programs with Equilibrium Constraints*, Tech. report NA 210, University of Dundee, Dundee, UK, 2002]. Our assumptions are more general since we do not use a critical assumption from that reference. In addition, we show that the elastic parameter update rule will not interfere locally with the superlinear convergence once the penalty parameter is appropriately chosen.

Key words. nonlinear programming, elastic mode, sequential quadratic programming, mathematical programs with equilibrium constraints, mathematical programs with complementarity constraints, complementarity constraints

AMS subject classifications. 90C30, 90C33, 90C55, 49M37, 65K10

DOI. 10.1137/S1052623402401221

1. Introduction. Complementarity constraints can be used to model numerous economics or engineering applications [30, 36]. Solving optimization problems with complementarity constraints may prove difficult for classical nonlinear optimization, however, given that, at a solution x^* , such problems cannot satisfy a constraint qualification [30, Chapter 3]. As a result, algorithms based on the linearization of the feasible set, such as sequential quadratic programming (SQP) algorithms, may fail because feasibility of the linearization can no longer be guaranteed in a neighborhood of the solution [30].

Several methods have been recently proposed to accommodate such problems. Nonsmooth and disjunctive programming approaches [31, 30, 36] can be used to successfully solve MPCC. However, for certain problems they may take a number of steps

*Received by the editors January 20, 2002; accepted for publication (in revised form) December 3, 2004; published electronically August 9, 2005.

<http://www.siam.org/journals/siopt/15-4/40122.html>

†Mathematics and Computer Science Division, Argonne National Laboratory, 9700 S. Cass Ave., Argonne, IL, 60439 (anitescu@mcs.anl.gov). This work was supported by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under Contract W-31-109-ENG-38. This work was also supported by award DMS-9973071 of the National Science Foundation. The submitted manuscript has been created by the University of Chicago as Operator of Argonne National Laboratory (“Argonne”) under contract W-31-109-ENG-38 with the U.S. Department of Energy. The U.S. Government retains for itself, and others acting on its behalf, a paid-up, nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government. Copyright is owned by SIAM to the extent not limited by these rights.

that is exponential in the size of the problem. For special cases, such as linear data functions, or bilevel optimization, other successful approaches have been defined [22, 30].

In this work we investigate an elastic mode SQP approach for mathematical programs with complementarity constraints (MPCCs). The elastic mode is a standard technique of approaching infeasible subproblems by relaxing the constraints and introducing a differentiable penalty term in the objective function [24].

Here we use the framework from [40] to determine sufficient conditions for MPCCs to have nonempty Lagrange multiplier sets. As in [40], the first- and second-order optimality properties of an MPCC are compared with the similar properties of two nonlinear programs that involve no complementarity constraints.

Note that the elastic mode formulation is different from regularization approaches [25, 30, 41, 28] (though the latter reference introduces an elastic mode relaxation, but not with respect to all nonlinear constraints, and, in particular, not with respect to the complementarity constraints). The regularization approaches create a family of nonlinear programs that are regular (in the sense that they satisfy a constraint qualification at the solution) for the interior of the domain of the regularization parameter. However, the solution of the regularized problems is different from the one of the original MPCC for any value of the parameter in the interior of the domain. The solution of the MPCC is obtained in the limit of the regularization parameter going towards the boundary of its domain. By contrast, under certain conditions to be developed in this work, the elastic mode formulation transforms the original MPCC into a regular nonlinear program with the same solution as the MPCC.

We also note that SQP approaches have been applied before to MPCC, in connection to regularization methods [21]. However, this is the first work that analyzes in detail issues connected to applying an elastic mode SQP approach to solving MPCC.

The paper is structured as follows.

- In section 1 we review the relevant nonlinear programming concepts.
- In section 2 we discuss sufficient conditions for MPCC to have a nonempty Lagrange multiplier set, in spite of not satisfying a constraint qualification at any point.
- In section 3 we show that the elastic mode applied to an instance of the MPCC class will retrieve a local solution of the problem for a finite value of the penalty parameter, a point which is supported by several numerical examples.
- In section 4 we prove that an adaptive elastic mode approach built around an algorithm recently analyzed in Fletcher et al. [18] in the MPCC context will result in superlinear convergence near the solution of an MPCC under assumptions weaker than in [18]. Specifically, here we do not assume that the iterates are either feasible or satisfy the complementarity constraints for the unrelaxed problem. If the sequence produced by the algorithm is assumed to converge, we show that superlinear convergence follows from even weaker assumptions about the signs of the relevant multipliers. In addition, we show that the elastic parameter update rule will not affect locally the superlinear convergence once the penalty parameter is appropriately chosen.

1.1. Optimality conditions for general nonlinear programming. We review the optimality conditions for a general nonlinear program

$$(1.1) \quad \min_x \tilde{f}(x) \quad \text{subject to } \tilde{g}(x) \leq 0, \quad \tilde{h}(x) = 0.$$

Here $\tilde{g} : \mathcal{R}^n \rightarrow \mathcal{R}^m$, $\tilde{h} : \mathcal{R}^n \rightarrow \mathcal{R}^r$. We assume that $\tilde{f}$, $\tilde{g}$, and $\tilde{h}$ are twice continuously differentiable.

We call x a stationary point of (1.1) if the Fritz–John condition holds: There exist multipliers $0 \neq \tilde{\lambda} = (\tilde{\lambda}_0, \tilde{\lambda}_1, \dots, \tilde{\lambda}_{m+r}) \in \mathcal{R}^{m+r+1}$, such that

$$(1.2) \quad \begin{aligned} \nabla_x \mathcal{L}(x, \tilde{\lambda}) &= 0, \quad \tilde{h}(x) = 0; \quad \tilde{\lambda}_i \geq 0, \quad \tilde{g}_i(x) \leq 0, \quad \text{for } i = 1, 2, \dots, m; \\ \sum_{i=1}^m \tilde{\lambda}_i \tilde{g}_i(x) &= 0. \end{aligned}$$

Here $\mathcal{L}$ is the Lagrangian function

$$(1.3) \quad \mathcal{L}(x, \tilde{\lambda}) = \tilde{\lambda}_0 \tilde{f}(x) + \sum_{i=1}^m \tilde{\lambda}_i \tilde{g}_i(x) + \sum_{j=1}^r \tilde{\lambda}_{m+j} \tilde{h}_j(x).$$

A local solution x^* of (1.1) is a stationary point [37]. We introduce the sets of generalized Lagrange multipliers

$$(1.4) \quad \Lambda^g(x) = \{0 \neq \tilde{\lambda} \in \mathcal{R}^{m+r+1} \mid \tilde{\lambda} \text{ satisfies (1.2) at } x\},$$

$$(1.5) \quad \Lambda_1^g(x) = \{\tilde{\lambda} \in \Lambda^g(x) \mid \tilde{\lambda}_0 = 1\}.$$

The set of active inequality constraints at a stationary point x is

$$(1.6) \quad \tilde{\mathcal{A}}(x) = \{i \in \{1, 2, \dots, m\} \mid \tilde{g}_i(x) = 0\}.$$

The set of inactive inequality constraints at x is the complement of $\tilde{\mathcal{A}}(x)$:

$$(1.7) \quad \tilde{\mathcal{A}}^c(x) = \{1, 2, \dots, m\} - \tilde{\mathcal{A}}(x).$$

With this notation, the complementarity condition from (1.2), $\sum_{i=1}^m \tilde{\lambda}_i \tilde{g}_i(x) = 0$, becomes $\tilde{\lambda}_{\tilde{\mathcal{A}}^c(x)} = 0$.

If certain regularity conditions hold at a stationary point x (discussed below), there exist $\tilde{\mu} = (\tilde{\mu}_1, \tilde{\mu}_2, \dots, \tilde{\mu}_{m+r}) \in \mathcal{R}^{m+r}$ that satisfy the Karush–Kuhn–Tucker (KKT) conditions [3, 4, 15]:

$$(1.8) \quad \begin{aligned} \nabla_x \tilde{f}(x) + \sum_{i=1}^m \tilde{\mu}_i \nabla_x \tilde{g}_i(x) + \sum_{j=1}^r \tilde{\mu}_{m+j} \nabla_x \tilde{h}_j(x) &= 0, \quad \tilde{h}(x) = 0; \\ \tilde{\mu}_i &\geq 0, \quad \tilde{g}_i(x) \leq 0, \quad \tilde{\mu}_i \tilde{g}_i(x) = 0, \quad \text{for } i = 1, 2, \dots, m. \end{aligned}$$

In this case, $\tilde{\mu}$ are referred to as the Lagrange multipliers, and x is called a KKT point. We denote the set of Lagrange multipliers by

$$(1.9) \quad \Lambda(x) = \{\tilde{\mu} \in \mathcal{R}^{m+r} \mid \tilde{\mu} \text{ satisfies (1.8) at } x\}.$$

A simple inspection of the definitions of $\Lambda(x)$ and $\Lambda_1^g(x)$ reveals that

$$\tilde{\mu} \in \Lambda(x) \Leftrightarrow (1, \tilde{\mu}) \in \Lambda_1^g(x).$$

Also, because of the first-order homogeneity of the conditions (1.2), and from (1.8), it immediately follows that

$$(1.10) \quad \Lambda(x) \neq \emptyset \Leftrightarrow \Lambda_1^g(x) \neq \emptyset \Leftrightarrow \exists \tilde{\lambda} \in \Lambda^g(x), \quad \text{such that } \tilde{\lambda}_0 \neq 0.$$

The regularity condition, or constraint qualification, ensures that a linear approximation of the feasible set in the neighborhood of a stationary point x captures the geometry of the feasible set. The regularity condition that we will use at times at a stationary point x is the Mangasarian–Fromovitz constraint qualification (MFCQ) [33, 32]:

- (MFCQ) 1. $\nabla_x \tilde{h}_j(x)$, $j = 1, 2, \dots, r$, are linearly independent and
 2. $\exists p \neq 0$ such that $\nabla_x \tilde{h}_j(x)^T p = 0$, $j = 1, 2, \dots, r$,
 and $\nabla_x \tilde{g}_i(x)^T p < 0$, $i \in \tilde{\mathcal{A}}(x)$.

It is well known [23] that (MFCQ) is equivalent to the fact that the set $\Lambda(x)$ of Lagrange multipliers of (1.1) is nonempty and bounded at a stationary point x of (1.1). Note that $\Lambda(x)$ is certainly polyhedral in any case.

Another condition that we will use on occasion is the strict Mangasarian–Fromovitz constraint qualification (SMFCQ). We say that this condition is satisfied by (1.1) at a KKT point x if

- (SMFCQ) 1) (MFCQ) is satisfied at x and
 2) the Lagrange multiplier set $\Lambda(x)$ contains exactly one element.

The critical cone at a stationary point x is [14, 42]:

$$(1.11) \quad \mathcal{C}(x) = \{u \in \mathcal{R}^n \mid \nabla_x \tilde{h}_j(x)^T u = 0, \ j = 1, 2, \dots, r, \\ \nabla_x \tilde{g}_i(x)^T u \leq 0, \ i \in \tilde{\mathcal{A}}(x); \ \nabla_x \tilde{f}(x)^T u \leq 0\}.$$

We now review the conditions for a point x^* to be a solution of (1.1). The second-order necessary conditions for x^* to be a local minimum are that $\Lambda^g(x^*) \neq \emptyset$ and [26]

$$(1.12) \quad \forall u \in \mathcal{C}(x^*), \ \exists \tilde{\lambda}^* \in \Lambda^g(x^*), \quad \text{such that } u^T \nabla_{xx}^2 \mathcal{L}(x^*, \tilde{\lambda}^*) u \geq 0.$$

The second-order sufficient conditions for x^* to be a local minimum are that $\Lambda^g(x^*) \neq \emptyset$ and [26]

$$(1.13) \quad \forall u \in \mathcal{C}(x^*), \ u \neq 0, \ \exists \tilde{\lambda}^* \in \Lambda^g(x^*), \quad \text{such that } u^T \nabla_{xx}^2 \mathcal{L}(x^*, \tilde{\lambda}^*) u > 0.$$

Stronger second-order conditions are Robinson's conditions. These conditions are that, at a solution x^* , the following condition holds:

$$\forall u \in \mathcal{C}(x^*), \ u \neq 0, \ \forall \tilde{\lambda}^* \in \Lambda^g(x^*), \quad \text{we have that } u^T \nabla_{xx}^2 \mathcal{L}(x^*, \tilde{\lambda}^*) u > 0.$$

In fact, we will invoke Robinson's conditions for the case where $\Lambda_1^g(x^*) \neq \emptyset$. In the latter situation, Robinson's conditions are equivalent to

$$(RSOSC) \quad \forall u \in \mathcal{C}(x^*), \ u \neq 0, \ \forall \tilde{\lambda}^* \in \Lambda_1^g(x^*), \quad \text{we have that } u^T \nabla_{xx}^2 \mathcal{L}(x^*, \tilde{\lambda}^*) u > 0.$$

1.2. Notation. For a mapping $q : \mathcal{R}^n \rightarrow \mathcal{R}^l$, we define

$$q^+(x) = \begin{pmatrix} \max\{q_1(x), 0\} \\ \max\{q_2(x), 0\} \\ \vdots \\ \max\{q_l(x), 0\} \end{pmatrix} \quad \text{and} \quad q^-(x) = \begin{pmatrix} \max\{-q_1(x), 0\} \\ \max\{-q_2(x), 0\} \\ \vdots \\ \max\{-q_l(x), 0\} \end{pmatrix}.$$

With this definition it immediately follows that $q(x) = q^+(x) - q^-(x)$ and $|q_i(x)| = q_i^+(x) + q_i^-(x)$, $i = 1, 2, \dots, l$.

In this work we use the convention that $\nabla_x q(x)$ is a matrix with l rows and n columns. This will allow us to write first-order Taylor expansion of $q(x)$ at some point $\hat{x}$ as $q(\hat{x}) + \nabla_x q(\hat{x})(x - \hat{x})$, without the need to use a transpose sign. In particular, if $q : \mathcal{R}^n \rightarrow \mathcal{R}$, $\nabla_x q(x)$ is a row vector.

We will also use the Landau notation. We say that a is of order b , which we denote by $a = O(b)$, if there exists c such that $a \leq cb$ for all a and b sufficiently small. We denote by $a = \Omega(b)$ quantities a and b that satisfy $a = O(b)$ and $b = O(a)$.

We will use certain symbols twice to denote related data of different programs. However, to avoid confusion, we will use a $\tilde{\cdot}$ sign for the data of the general nonlinear programming problem (1.1), whereas the same objects associated with the MPCC problem (to be defined later) are denoted without the $\tilde{\cdot}$ sign. For instance, $\tilde{f}, \tilde{g}, \tilde{h}$ denote, respectively, the objective, the inequality constraint, and the equality constraints of the general nonlinear programming problem, whereas f, g, h denote, respectively, the objective, the inequality constraints, and the equality constraints of the MPCC problem. The MPCC problem, however, has, in addition, complementarity constraints.

In the literature, problems of the type we treat here are also called mathematical programs with equilibrium constraints (MPEC), an acronym that we may use when we invoke optimality conditions from the respective references.

We denote the L_∞ nondifferentiable penalty function by

$$(1.14) \quad \tilde{P}_\infty(x) = \max\{\tilde{g}_1(x), \tilde{g}_2(x), \dots, \tilde{g}_m(x), |\tilde{h}_1(x)|, |\tilde{h}_2(x)|, \dots, |\tilde{h}_r(x)|, 0\}.$$

We also define the L_1 penalty function as

$$(1.15) \quad \tilde{P}_1(x) = \sum_{i=1}^m \tilde{g}_i^+(x) + \sum_{j=1}^r |\tilde{h}_j(x)|.$$

It is immediate that

$$0 \leq \tilde{P}_\infty(x) \leq \tilde{P}_1(x) \leq (m+r)\tilde{P}_\infty(x).$$

An obvious consequence of (1.15) and (1.14) is that x is a feasible point of (1.1) if and only if $\tilde{P}_1(x) = \tilde{P}_\infty(x) = 0$.

We say that the nonlinear program (1.1) satisfies the quadratic growth condition with a parameter $\tilde{\sigma}$ at x^* if

$$(1.16) \quad \max\{\tilde{f}(x) - \tilde{f}(x^*), \tilde{P}_\infty(x)\} \geq \tilde{\sigma} \|x - x^*\|^2$$

holds for some $\tilde{\sigma} > 0$ and all x in a neighborhood of x^* . The quadratic growth condition is equivalent to the second-order sufficient conditions (1.13) [6, 7, 26, 27, 42], and it is the weakest possible second-order condition.

For the case in which (MFCQ) holds at a solution x^* of (1.1), the quadratic growth condition at x^* is equivalent to [6]

$$(1.17) \quad \tilde{f}(x) - \tilde{f}(x^*) \geq \tilde{\sigma}_{\tilde{f}} \|x - x^*\|^2$$

for some $\tilde{\sigma}_{\tilde{f}} > 0$ and all x feasible in a neighborhood of x^* .

1.3. Exact penalty conditions for degenerate nonlinear programming.

We now assume that at a solution x^* of the nonlinear program (1.1) the following conditions hold:

1. The Lagrange multiplier set at x^* , $\Lambda(x^*)$, is not empty.
2. The quadratic growth condition (1.16) is satisfied.

Then there exists a neighborhood $\mathcal{V}(x^*)$, some penalty parameters $\tilde{c}_1 \geq 0$, $\tilde{c}_\infty \geq 0$, and some growth parameters $\sigma_1 > 0$ and $\sigma_\infty > 0$ such that [7, Theorem 3.113]

$$(1.18) \quad \begin{aligned} \forall x \in \mathcal{V}(x^*), \psi_1(x) &= \tilde{f}(x) + \tilde{c}_1 \tilde{P}_1(x) \geq \tilde{f}(x^*) + \sigma_1 \|x - x^*\|^2 \\ &= \psi_1(x^*) + \sigma_1 \|x - x^*\|^2, \end{aligned}$$

$$(1.19) \quad \begin{aligned} \forall x \in \mathcal{V}(x^*), \psi_\infty(x) &= \tilde{f}(x) + \tilde{c}_\infty \tilde{P}_\infty(x) \geq \tilde{f}(x^*) + \sigma_\infty \|x - x^*\|^2 \\ &= \psi_\infty(x^*) + \sigma_\infty \|x - x^*\|^2. \end{aligned}$$

Therefore, x^* becomes an unconstrained strict local minimum for the nondifferentiable functions $\psi_1(x)$ and $\psi_\infty(x)$. Such functions are called nondifferentiable exact merit functions for the nonlinear program (1.1) [3, 4, 15]. If (1.18) and (1.19) are satisfied, then we say that the functions $\psi_1(x)$ and $\psi_\infty(x)$ satisfy a quadratic growth condition near x^* .

1.4. Formulation of MPCC. We use notation similar to the one in [40] to define a MPCC,

$$(MPCC) \quad \begin{array}{ll} \min_x & f(x) \\ \text{subject to} & g_i(x) \leq 0, \quad i = 1, 2, \dots, n_i, \\ & h_j(x) = 0, \quad j = 1, 2, \dots, n_e, \\ & F_{k,1}(x) \leq 0, \quad k = 1, 2, \dots, n_c, \\ & F_{k,2}(x) \leq 0, \quad k = 1, 2, \dots, n_c, \\ & F_{k,1}(x)F_{k,2}(x) \leq 0, \quad k = 1, 2, \dots, n_c. \end{array}$$

In this work we assume that the data of MPCC ($f(x)$, $h(x)$, $g(x)$ and $F_{k,i}(x)$, for $k = 1, 2, \dots, n_c$, and $i = 1, 2$) are twice continuously differentiable.

For a given k , the constraints $F_{k,1}(x) \leq 0$, $F_{k,2}(x) \leq 0$ imply that $F_{k,1}(x)F_{k,2}(x) \leq 0$ is equivalent to $F_{k,1}(x)F_{k,2}(x) = 0$. The constraints $F_{k,1}(x)F_{k,2}(x) \leq 0$ are thus called complementarity constraints and are active at any feasible point of (MPCC). Therefore (MPCC) included a particular choice of representing the complementarity constraints $F_{k,1}(x) \perp F_{k,2}(x)$ for $k = 1, 2, \dots, n_c$ as constraints of a smooth nonlinear program. We present another one of the several equivalent smooth nonlinear programming formulations in (2.13).

Since we cannot have $F_{k,1}(x) < 0$, $F_{k,2}(x) < 0$, and $F_{k,1}(x)F_{k,2}(x) < 0$ simultaneously, it follows that (MFCQ) cannot hold at any feasible point x [30, 40].

1.5. MPCC notation. In this section, which previews our general convergence results, we use the same notation from [40] to denote certain index sets, because at some point we invoke a theorem from that reference. Later, in our superlinear convergence results we will use notation from [18] to denote similar index sets, because we will use results from the latter reference.

If i is one of $1, 2$ we define $\bar{i} = 2 - i + 1$. Therefore $i = 1 \Rightarrow \bar{i} = 2$, and $i = 2 \Rightarrow \bar{i} = 1$. The complementarity constraints can thus be written as $F_{k,i}(x)F_{k,\bar{i}}(x) \leq 0$, $k = 1, 2, \dots, n_c$. We use the notation

$$(1.20) \quad F(x) = (F_{11}(x), F_{12}(x), F_{21}(x), F_{22}(x), \dots, F_{n_c 1}(x), F_{n_c 2}(x))^T.$$

The active set of the inequality constraints $g_i(x) \leq 0$, $1 \leq i \leq m$, at a feasible point x is

$$(1.21) \quad \mathcal{A}(x) = \{i \in \{1, 2, \dots, n_i\} \mid g_i(x) = 0\}.$$

We use the following notation:

$$(1.22) \quad \mathcal{I}(x) = \{(k, i) \in \{1, 2, \dots, n_c\} \times \{1, 2\} \mid F_{k,\bar{i}}(x) < 0\},$$

$$(1.23) \quad \bar{\mathcal{I}}(x) = \{(k, i) \in \{1, 2, \dots, n_c\} \times \{1, 2\} \mid F_{k,i}(x) < 0\},$$

$$(1.24) \quad \bar{\mathcal{D}}(x) = \{(k, i) \in \{1, 2, \dots, n_c\} \times \{1, 2\} \mid F_{k,i}(x) = F_{k,\bar{i}}(x) = 0\},$$

$$(1.25) \quad \mathcal{I}^c(x) = \{1, 2, \dots, n_c\} \times \{1, 2\} - \mathcal{I}(x),$$

$$(1.26) \quad \mathcal{K}(x) = \{k \in \{1, 2, \dots, n_c\} \mid (k, 1) \in \mathcal{I}(x) \text{ or } (k, 2) \in \mathcal{I}(x)\},$$

$$(1.27) \quad \bar{\mathcal{K}}(x) = \{k \in \{1, 2, \dots, n_c\} \mid F_{k,1}(x) = F_{k,2}(x) = 0\} = \{1, 2, \dots, n_c\} - \mathcal{K}(x).$$

There are two cases for the constraints involved in the complementarity constraints at a feasible point x .

1. $F_{k,1}(x) + F_{k,2}(x) < 0$. In this case there is an $i(k) \in \{1, 2\}$ such that $F_{k,i(k)} = 0$ and $F_{k,\bar{i}(k)} < 0$. Therefore, with our notation $k \in \mathcal{K}(x)$, $(k, i(k)) \in \mathcal{I}(x)$ and $(k, \bar{i}(k)) \in \bar{\mathcal{I}}(x)$. We call $F_{k,1}(x)$, $F_{k,2}(x)$ a nondegenerate (or strictly complementary) pair. In the rest of the paper $i(k)$ and $\bar{i}(k)$ will have the meaning defined in this paragraph, whenever $k \in \mathcal{K}$.

2. $F_{k,1}(x) + F_{k,2}(x) = 0$, or $F_{k,1}(x) = F_{k,2}(x) = 0$. In this case $k \in \bar{\mathcal{K}}(x)$, $(k, 1) \in \bar{\mathcal{D}}(x)$ and $(k, 2) \in \bar{\mathcal{D}}(x)$. We call $F_{k,1}(x)$, $F_{k,2}(x)$ a degenerate pair. Therefore $\mathcal{I}(x)$ and $\bar{\mathcal{I}}(x)$ contain the indices of the active constraints at which strict complementarity occurs, whereas $\bar{\mathcal{D}}(x)$ contains the indices of the constraints that are degenerate at x from the point of view of complementarity. The set $\mathcal{K}(x)$ represents the indices k at which strict complementarity occurs and $\bar{\mathcal{K}}(x)$ the indices k at which complementarity degeneracy occurs.

Since we are interested in the behavior of (MPCC) at a solution point x^* , we may avoid the dependence of these index sets on x . Therefore we denote $\mathcal{I} = \mathcal{I}(x^*)$, $\bar{\mathcal{D}} = \bar{\mathcal{D}}(x^*)$, $\mathcal{K} = \mathcal{K}(x^*)$, and $\mathcal{A} = \mathcal{A}(x^*)$.

For a set of pairs $\mathcal{J} \subset \{1, 2, \dots, n_c\} \times \{1, 2\}$ we denote by $F_{\mathcal{J}}$ a map whose components are $F_{k,i}$ with $(k, i) \in \mathcal{J}$.

1.6. Associated nonlinear programs at x^* . In this section we associate two nonlinear programs to (MPCC). This will help with characterizing the stationarity conditions for (MPCC). The notation is from [40].

At x^* we associate the relaxed nonlinear program (RNLP) to (MPCC),

$$\begin{aligned} (\text{RNLP}) \quad & \min_x \quad f(x) \\ & \text{subject to} \quad g_i(x) \leq 0, \quad i = 1, 2, \dots, n_i, \\ & \quad h_j(x) = 0, \quad j = 1, 2, \dots, n_e, \\ & \quad F_{\bar{\mathcal{D}}}(x) \leq 0, \\ & \quad F_{\mathcal{I}}(x) = 0. \end{aligned}$$

As can be seen, (RNLP) is obtained from (MPCC) by dropping the elements from $F(x)$ that are inactive at x^* , as well as the complementarity constraints, but enforcing the complements of inactive constraints as equality constraints.

Note that (RNLP) typically includes the constraints $F_{\bar{\mathcal{I}}}(x) \leq 0$ in order to have a tighter relaxation of the feasible set of (MPCC). The nonlinear program (RNLP), however, is used only for local analysis near a solution x^* . Since the constraints $F_{\bar{\mathcal{I}}}(x) \leq 0$ are not active in a neighborhood of x^* , we do not consider them in the analysis.

We also associate at x^* the tightened nonlinear program (TNLP), in which all the complementarity constraints in (MPCC) are dropped and all active constraints at x^* connected to complementarity constraints are replaced by equality constraints,

$$\begin{aligned} (\text{TNLP}) \quad & \min_x \quad f(x) \\ & \text{subject to} \quad g_i(x) \leq 0, \quad i = 1, 2, \dots, n_i, \\ & \quad h_j(x) = 0, \quad j = 1, 2, \dots, n_e, \\ & \quad F_{\bar{\mathcal{D}}}(x) = 0, \\ & \quad F_{\mathcal{I}}(x) = 0. \end{aligned}$$

We immediately see that, near x^* , (TNLP) is a more constrained problem than (MPCC), which in turn is more constrained than (RNLP), and all three programs have the same objective function. As a result, if x^* is a local solution of (RNLP), then it must be a local solution of (MPCC). Also, if x^* is a local solution of (MPCC), then it will be a local solution of (TNLP). None of the reverse implications hold in general for either local solutions or stationary points.

However, if (TNLP) satisfies (SMFCQ) at a solution x^* of (MPCC), then x^* is a KKT point of (TNLP) and (RNLP) [40].

2. The Lagrange multiplier set of MPCC. In this section we analyze the relationship between the relevant mathematical objects of (MPCC) and (RNLP) at a solution x^* . The (RNLP) formulation does not immediately violate (MFCQ) the way (MPCC) does. By establishing a correspondence between the Lagrange multiplier sets of (RNLP) and (MPCC) we ensure that, under certain conditions, (MPCC) has a nonempty Lagrange multiplier set, although it does not satisfy a constraint qualification.

2.1. Critical cones. In this section we compare the critical cones of (MPCC) and (RNLP). The active sets play a structural part in the definition of the critical cones. We have that

$$\nabla_x(F_{k,1}F_{k,2})(x^*) = F_{k,1}(x^*)\nabla_x F_{k,2}(x^*) + F_{k,2}(x^*)\nabla_x F_{k,1}(x^*).$$

Using the definition (1.11) we obtain that the critical cone of (MPCC) is

$$(2.1) \quad \mathcal{C}_{\text{MPCC}} = \{u \in R^n \mid \begin{array}{ll} \nabla_x f(x^*)u & \leq 0, \\ \nabla_x g_i(x^*)u & \leq 0, \quad i \in \mathcal{A}, \\ \nabla_x h_j(x^*)u & = 0, \quad j \in 1, 2, \dots, n_e, \\ \nabla_x F_{k,1}(x^*)u & \leq 0, \quad k \in \bar{\mathcal{K}}, \\ \nabla_x F_{k,2}(x^*)u & \leq 0, \quad k \in \bar{\mathcal{K}}, \\ \nabla_x F_{k,i(k)}(x^*)u & \leq 0, \quad (k, i(k)) \in \mathcal{I}, \\ F_{k,i(k)}(x^*)\nabla_x F_{k,i(k)}(x^*)u & \leq 0, \quad (k, i(k)) \in \mathcal{I}. \end{array}$$

Note that the definition we use here of the critical cone corresponds to the nonlinear programming interpretation of (MPCC). There exists another (combinatorial) definition of the critical cone that is used in connection with disjunctive approaches [30, equation 5.3.(2)].

We use (1.11) again to determine the critical cone of the relaxed nonlinear program. It is immediate from the definition of the index sets $\mathcal{I}$, $\mathcal{K}$, and $\bar{\mathcal{D}}$ that all constraints involving components of $F(x)$ are active at x^* for (RNLP). It thus follows that the critical cone of (RNLP) is

$$(2.2) \quad \mathcal{C}_{\text{RNLP}} = \{u \in R^n \mid \begin{array}{ll} \nabla_x f(x^*)u & \leq 0, \\ \nabla_x g_i(x^*)u & \leq 0, \quad i \in \mathcal{A}, \\ \nabla_x h_j(x^*)u & = 0, \quad j \in 1, 2, \dots, n_e, \\ \nabla_x F_{k,1}(x^*)u & \leq 0, \quad k \in \bar{\mathcal{K}}, \\ \nabla_x F_{k,2}(x^*)u & \leq 0, \quad k \in \bar{\mathcal{K}}, \\ \nabla_x F_{k,i(k)}(x^*)u & = 0, \quad (k, i(k)) \in \mathcal{I}. \end{array}$$

LEMMA 2.1. $\mathcal{C}_{\text{MPCC}} = \mathcal{C}_{\text{RNLP}}$.

Proof. The conclusion is immediate, by noting that all the constraints involving the critical cones are the same with the exception of the ones involving indices k for

which $(k, i(k)) \in \mathcal{I}$. For these k , from the definition (1.22) of the index sets it follows that $F_{k, \bar{i}(k)}(x^*) < 0$. We therefore have that

$$\begin{aligned} \nabla_x F_{k, i(k)}(x^*)u \leq 0 \quad \text{and} \quad F_{k, \bar{i}(k)}(x^*)\nabla_x F_{k, i(k)}(x^*)u \leq 0 &\Leftrightarrow \\ \nabla_x F_{k, i(k)}(x^*)u \leq 0 \quad \text{and} \quad \nabla_x F_{k, i(k)}(x^*)u \geq 0 &\Leftrightarrow \\ \nabla_x F_{k, i(k)}(x^*)u = 0. \end{aligned}$$

Since the remaining constraints of (RNLP) and (MPCC) are the same, this equivalence proves the claim. $\square$

2.2. Generalized Lagrange multipliers. The set of generalized Lagrange multipliers of MPCC at x^* is a set of multiples

$$0 \neq (\alpha, \nu, \pi, \mu, \eta) \in \mathcal{R} \times \mathcal{R}^{n_i} \times \mathcal{R}^{n_e} \times \mathcal{R}^{2n_c} \times \mathcal{R}^{n_c}$$

that satisfies the Fritz–John conditions (1.2). Since μ are the multipliers corresponding to the components of $F(x)$, we will index them by elements in $(1, 2, \dots, n_c) \times (1, 2)$. The Fritz–John conditions for (MPCC) at x^* are that x^* is feasible for (MPCC) and that

$$\begin{aligned} (2.3) \quad & \alpha \nabla_x f(x^*) + \sum_{i=1}^{n_i} \nu_i \nabla_x g_i(x^*) + \sum_{j=1}^{n_e} \pi_j \nabla_x h_j(x^*) \\ & + \sum_{k=1}^{n_c} [\mu_{k,1} \nabla_x F_{k,1}(x^*) + \mu_{k,2} \nabla_x F_{k,2}(x^*) + \eta_k \nabla_x (F_{k,1} F_{k,2})(x^*)] = 0, \end{aligned}$$

$$\begin{aligned} (2.4) \quad & F_{k,i}(x^*) \leq 0, \quad \mu_{k,i} \geq 0, \quad \mu_{k,i} F_{k,i}(x^*) = 0, \quad k = 1, 2, \dots, n_c, \quad i = 1, 2, \\ & g_i(x^*) \leq 0, \quad \nu_i \geq 0, \quad \nu_i g_i(x^*) = 0, \quad i = 1, 2, \dots, n_i, \\ & F_{k,1}(x^*) F_{k,2}(x^*) \leq 0, \quad \eta_k \geq 0, \quad \eta_k F_{k,1}(x^*) F_{k,2}(x^*) = 0, \quad k = 1, 2, \dots, n_c. \end{aligned}$$

From our definition of the index sets, it follows that $F_{\bar{\mathcal{I}}}(x^*) < 0$ and $g_{\mathcal{A}^c}(x^*) < 0$. Therefore, from the complementarity conditions (2.4), it follows that $\mu_{\bar{\mathcal{I}}} = 0$ and $\nu_{\mathcal{A}^c} = 0$.

We can also determine the relations satisfied by the generalized Lagrange multipliers of (RNLP). As discussed above, the index sets that define (RNLP) have been chosen such that all constraints involving components of $F(x)$ are active. Therefore the generalized Lagrange multipliers are

$$0 \neq (\tilde{\alpha}, \tilde{\nu}, \tilde{\pi}, \tilde{\mu}, \tilde{\eta}) \in \mathcal{R} \times \mathcal{R}^{n_i} \times \mathcal{R}^{n_e} \times \mathcal{R}^{n_{\bar{\mathcal{D}}}} \times \mathcal{R}^{n_{\mathcal{I}}},$$

which satisfy the following Fritz–John conditions:

$$\begin{aligned} (2.5) \quad & \tilde{\alpha} \nabla_x f(x^*) + \sum_{i=1}^{n_i} \tilde{\nu}_i \nabla_x g_i(x^*) + \sum_{j=1}^{n_e} \tilde{\pi}_j \nabla_x h_j(x^*) \\ & + \sum_{k \in \bar{\mathcal{K}}} [\tilde{\mu}_{k,1} \nabla_x F_{k,1}(x^*) + \tilde{\mu}_{k,2} \nabla_x F_{k,2}(x^*)] + \sum_{k \in \mathcal{K}} \tilde{\eta}_{k, i(k)} \nabla_x F_{k, i(k)}(x^*) = 0, \end{aligned}$$

$$\begin{aligned} (2.6) \quad & g_i(x^*) \leq 0, \quad \tilde{\nu}_i \geq 0, \quad \tilde{\nu}_i g_i(x^*) = 0, \quad i = 1, 2, \dots, n_i, \\ & \tilde{\mu}_{k,1} \geq 0, \quad \tilde{\mu}_{k,2} \geq 0, \quad k \in \bar{\mathcal{K}}. \end{aligned}$$

Here $\tilde{\mu}$ is a vector that is indexed by elements of $\bar{\mathcal{D}}$, and $\tilde{\eta}$ is indexed by elements of $\mathcal{I}$.

2.3. Relations between the generalized Lagrange multiplier sets of (MPCC) and (RNLP). Take $\tilde{\lambda} = (\tilde{\alpha}, \tilde{\nu}, \tilde{\pi}, \tilde{\mu}, \tilde{\eta}) \in \Lambda^g_{\text{RNLP}}$. We construct from the generalized multiplier $\tilde{\lambda}$ of (RNLP) a generalized multiplier λ° of (MPCC). We define the following types of components of λ° .

1. Components that correspond to the objective function or the inequality constraints $g_i(x) \leq 0$ and equality constraints $h_j(x) = 0$,

$$(2.7) \quad \alpha^\circ = \tilde{\alpha}; \quad \nu^\circ = \tilde{\nu}; \quad \pi^\circ = \tilde{\pi}.$$

2. Components connected to the pairwise degenerate constraints. For these we have $k \in \bar{\mathcal{K}}$ and $(k, 1), (k, 2) \in \bar{\mathcal{D}}$ or $F_{k,1}(x^*) = F_{k,2}(x^*) = 0$. We define

$$(2.8) \quad \mu_{k,i}^\circ = \tilde{\mu}_{k,i}, \quad (k, i) \in \bar{\mathcal{D}}; \quad \eta_k^\circ = 0, \quad k \in \bar{\mathcal{K}}.$$

We have that

$$\nabla_x(F_{k,1}F_{k,2})(x^*) = 0,$$

and therefore

$$(2.9) \quad \begin{aligned} \tilde{\mu}_{k,1} \nabla_x F_{k,1}(x^*) + \tilde{\mu}_{k,2} \nabla_x F_{k,2}(x^*) &= \mu_{k,1}^\circ \nabla_x F_{k,1}(x^*) \\ &+ \mu_{k,2}^\circ \nabla_x F_{k,2}(x^*) + \eta_k^\circ \nabla_x (F_{k,1}F_{k,2})(x^*). \end{aligned}$$

3. Components connected to pairwise strictly complementary constraints. In this case we have $k \in \mathcal{K}$, $(k, i(k)) \in \mathcal{I}$, and $(k, \bar{i}(k)) \in \bar{\mathcal{I}}$. Therefore $F_{k,\bar{i}(k)}(x^*) < 0$, $F_{k,i(k)}(x^*) = 0$, and we thus define the multipliers

$$(2.10) \quad \begin{aligned} \mu_{k,i(k)}^\circ &= \max\{\tilde{\eta}_{k,i(k)}, 0\}, & (k, i(k)) \in \mathcal{I}, \\ \mu_{k,\bar{i}(k)}^\circ &= 0, & (k, \bar{i}(k)) \in \bar{\mathcal{I}}, \\ \eta_k^\circ &= \frac{1}{F_{k,\bar{i}(k)}(x^*)} \min\{\tilde{\eta}_{k,i(k)}, 0\}, & k \in \mathcal{K}. \end{aligned}$$

It is immediate from these definitions that $\mu_{k,i(k)}^\circ \geq 0$ and $\eta_k^\circ \geq 0$. Since, for fixed k , $\tilde{\eta}_{k,i(k)}$ is the only multiplier of (RNLP) involved in definition (2.10), we obtain that

$$(2.11) \quad \begin{aligned} \tilde{\eta}_{k,i(k)} \nabla_x F_{k,i(k)}(x^*) &= [\max\{\tilde{\eta}_{k,i(k)}, 0\} + \min\{\tilde{\eta}_{k,i(k)}, 0\}] \nabla_x F_{k,i(k)}(x^*) \\ &= \mu_{k,i(k)}^\circ \nabla_x F_{k,i(k)}(x^*) + \eta_k^\circ F_{k,\bar{i}(k)}(x^*) \nabla_x F_{k,i(k)}(x^*) \\ &= \mu_{k,i(k)}^\circ \nabla_x F_{k,i(k)}(x^*) + \mu_{k,\bar{i}(k)}^\circ \nabla_x F_{k,\bar{i}(k)}(x^*) \\ &\quad + \eta_k^\circ \nabla_x (F_{k,i(k)}F_{k,\bar{i}(k)})(x^*). \end{aligned}$$

After we compare the terms that, following (2.9) and (2.11), are equal in (2.5) and (2.3), we obtain that $\lambda^\circ = (\alpha^\circ, \nu^\circ, \pi^\circ, \mu^\circ, \eta^\circ)$ satisfies (2.3) as well as (2.4). By tracing the definition of λ° we also have that $\tilde{\lambda} \neq 0 \Rightarrow \lambda^\circ \neq 0$. Therefore λ° is a generalized Lagrange multiplier of (MPCC) or

$$(2.12) \quad \lambda^\circ = (\alpha^\circ, \nu^\circ, \pi^\circ, \mu^\circ, \eta^\circ) \in \Lambda^g_{\text{MPCC}},$$

where $\alpha^\circ = \tilde{\alpha}$ from (2.7).

THEOREM 2.2. *The set of Lagrange multipliers of (RNLP) is not empty if and only if the set of Lagrange multipliers of (MPCC) is not empty.*

Proof. If the Lagrange multiplier set of (RNLP) is not empty, we can choose $\tilde{\lambda} = (1, \tilde{\nu}, \tilde{\pi}, \tilde{\mu}, \tilde{\eta}) \in \Lambda_{1, \text{RNLP}}^g$. From (2.12) it follows that $\lambda^\circ = (1, \nu^\circ, \pi^\circ, \mu^\circ, \eta^\circ) \in \Lambda_{1, \text{MPCC}}^g$ is a generalized multiplier of (MPCC). From (1.10) it follows that the Lagrange multiplier set of MPCC is not empty. The proof of the reverse statement follows in a similar manner and is omitted. $\square$

On one hand, in our work we are interested in KKT points of (MPCC). On the other, existence of Lagrange multipliers for (RNLP) at x^* amounts to what is called strong stationarity in the literature on MPCCs [40]. Therefore our previous result says that x^* is a KKT point for (MPCC) if and only if it is strongly stationary.

COROLLARY 2.3. *Assume that (TNLP) satisfies (SMFCQ) at a solution x^* of (MPCC), i.e.,*

1. $\nabla_x F_{\overline{\mathcal{D}}}(x^*)$, $\nabla_x F_{\mathcal{I}}(x^*)$, and $\nabla_x h(x^*)$ are linearly independent.
2. There exists $p \neq 0$ such that $\nabla_x F_{\overline{\mathcal{D}}}(x^*)p = 0$, $\nabla_x F_{\mathcal{I}}(x^*)p = 0$, $\nabla_x h(x^*)p = 0$, $\nabla_x g_i(x^*)p < 0$, for $i \in \mathcal{A}(x^*)$.
3. The Lagrange multiplier set of (TNLP) at x^* has a unique element.

Then the Lagrange multiplier set of (MPCC) is not empty.

Proof. From [40, Theorem 2], since (TNLP) satisfies (SMFCQ) at x^* , the Lagrange multiplier set of (RNLP) is not empty. Following Theorem 2.2, we obtain that the Lagrange multiplier set of (MPCC) is not empty, which proves the claim. $\square$

The proof of Corollary 2.3 also follows immediately from [40, Theorem 4] after using the observation that, when (SMFCQ) holds, x^* is a B-stationary point of (MPCC) [40, section 2.1]. The conclusion of Corollary 2.3 does not hold in absence of (SMFCQ) [40, Example 3].

2.4. An alternative formulation. We also investigate the following equivalent formulation of (MPCC), where the complementarity constraints have been replaced by one constraint:

$$(2.13) \quad \begin{array}{ll} \min_x & f(x) \\ \text{subject to} & g_i(x) \leq 0, \quad i = 1, 2, \dots, n_i, \\ & h_j(x) = 0, \quad j = 1, 2, \dots, n_e, \\ & F_{k,1}(x) \leq 0, \quad k = 1, 2, \dots, n_c, \\ & F_{k,2}(x) \leq 0, \quad k = 1, 2, \dots, n_c, \\ & \sum_{k=1}^{n_c} F_{k,1}(x)F_{k,2}(x) \leq 0. \end{array}$$

At a feasible point of the above program, we must have that $\sum_{k=1}^{n_c} F_{k,1}(x)F_{k,2}(x) = 0$ and the equivalence between (2.13) and (MPCC) follows immediately. This formulation is of interest in computations because it has fewer constraints than (MPCC).

LEMMA 2.4. *If the Lagrange multiplier set of (MPCC) is not empty, there exists a generalized Lagrange multiplier $(1, \nu, \pi, \mu, \eta) \in \Lambda_{\text{MPCC}}^g$ such that $\eta_k = \eta_1$, $k = 2, 3, \dots, n_c$.*

Proof. The proof follows from straightforward algebra of the type used in subsection 2.2. $\square$

We now describe the generalized Lagrange multiplier set of the alternative formulation (2.13). We denote mathematical objects connected to (2.13) by the subscript MPCC1 . We write the Fritz–John conditions (1.2) for (2.13) at the point x^* , and we obtain

$$\begin{aligned}
(2.14) \quad & \alpha^\diamond \nabla_x f(x^*) + \sum_{i=1}^{n_i} \nu_i^\diamond \nabla_x g_i(x^*) + \sum_{j=1}^{n_e} \pi_j^\diamond \nabla_x h_j(x^*) \\
& + \sum_{k=1}^{n_c} \sum_{i=1}^2 \mu_{k,i}^\diamond \nabla_x F_{k,i}(x^*) + \eta_1^\diamond \sum_{k=1}^{n_c} \nabla_x (F_{k,1} F_{k,2})(x^*) = 0,
\end{aligned}$$

$$\begin{aligned}
(2.15) \quad & F_{k,i}(x^*) \leq 0, \quad \mu_{k,i}^\diamond \geq 0, \quad \mu_{k,i}^\diamond F_{k,i}(x^*) = 0, \quad k = 1, 2, \dots, n_c, \quad i = 1, 2, \\
& g_i(x^*) \leq 0, \quad \nu_i^\diamond \geq 0, \quad \nu_i^\diamond g_i(x^*) = 0, \quad i = 1, 2, \dots, n_i,
\end{aligned}$$

and $\eta_1^\diamond \geq 0 \in \mathcal{R}$. Note that from the fact that x^* is feasible for (2.13) we also have that $\sum_{k=1}^{n_c} F_{k,1}(x) F_{k,2}(x) \leq 0$ and $F_{ki} \leq 0$, for $k = 1, 2, \dots, n_c$ and $i = 1, 2$. Consequently, we have that $\sum_{k=1}^{n_c} F_{k,1}(x) F_{k,2}(x) = 0$ and thus the complementarity condition $\eta_1^\diamond \sum_{k=1}^{n_c} F_{k,1}(x) F_{k,2}(x) = 0$ is redundant. We therefore ignore it in the Fritz–John conditions (2.14), (2.15).

A generalized multiplier of (2.13) is thus

$$\lambda^\diamond = (\alpha^\diamond, \nu^\diamond, \pi^\diamond, \mu^\diamond, \eta_1^\diamond) \in \Lambda_{MPCC1}^q \subset \mathcal{R} \times \mathcal{R}^{n_i} \times \mathcal{R}^{n_e} \times \mathcal{R}^{2n_c} \times \mathcal{R},$$

where $\lambda^\diamond$ satisfies the Fritz–John conditions (2.14), (2.15).

THEOREM 2.5.

(i) *The formulation (2.13) has a nonempty Lagrange multiplier set if and only if (MPCC) has a nonempty Lagrange multiplier set.*

(ii) *If (MPCC) has a nonempty Lagrange multiplier set and if it satisfies the quadratic growth condition at x^* , then (2.13) has a nonempty Lagrange multipliers set and satisfies the quadratic growth condition at x^* .*

Proof. The proof of part (i) this result is based on Lemma 2.4, in conjunction with (2.3), (2.4) and (2.14), (2.15).

The proof of part (ii) follows from part (i), the fact that the quadratic growth condition holds at x^* [7, Theorem 3.113] applied to the L_1 merit function, and the fact that the following inequality holds for x in some neighborhood of x^* and for some $c_{FF} > 0$:

$$\left(\sum_{k=1}^{n_c} F_{k1}(x) F_{k2}(x) \right)^+ \geq \sum_{k=1}^{n_c} (F_{k1}(x) F_{k2}(x))^+ - c_{FF} (\|F_1^+(x)\|_1 + \|F_2^+(x)\|_1). \quad \square$$

Theorems 2.2 and 2.5 give sufficient conditions for (MPCC) and (2.13) to have a nonempty Lagrange multiplier set in spite of the fact that neither satisfy a constraint qualification at any point in the usual sense of nonlinear programming. In section 3 these conditions will imply that a relaxed version of either (MPCC) or (2.13) will have the same solution as (MPCC) and will satisfy (MFCQ), which makes either approachable by SQP algorithms.

3. The elastic mode. An important class of techniques for solving nonlinear programs (1.1) is SQP. The main step in an algorithm of this type is solving the quadratic program

$$\begin{aligned}
(3.1) \quad & \min_d \quad \nabla_x \tilde{f}(x) d + d^T \widetilde{W} d \\
& \text{subject to} \quad \tilde{g}_i(x) + \nabla_x \tilde{g}_i(x) d \leq 0, \quad i = 1, 2, \dots, m, \\
& \quad \quad \quad \tilde{h}_j(x) + \nabla_x \tilde{h}_j(x) d = 0, \quad j = 1, 2, \dots, r.
\end{aligned}$$

The matrix $\widetilde{W}$ can be the Hessian of the Lagrangian (1.1) at (x, λ) [17], where λ is a suitable approximation of a Lagrange multiplier or a positive definite matrix that approximates the Hessian of the Lagrangian on a certain subspace [15, 24, 34]. A trust-region type constraint may be added to (3.1) to enforce good global convergence properties [17]. The solution $\bar{d}$ of (3.1) is then used in conjunction with a merit function and/or line search to determine a new iterate. We give here only a brief description of SQP algorithms, since our interest is solely in showing how the difficulties regarding the potential infeasibility of (3.1), when applied to (MPCC), can be circumvented. For more details about SQP methods, see [15, 17, 24, 34].

If a nonlinear program satisfies (MFCQ) at x^* , then the quadratic program will be feasible in a neighborhood of x^* . If (MFCQ) does not hold at x^* , however, the possibility exists that (3.1) is infeasible, no matter how close to x^* [30, 36, 40]. This is an issue in the context of this paper because (MPCC) does not satisfy the (MFCQ) at a solution x^* .

In the case of infeasible subproblems some of the SQP algorithms initiate the *elastic mode* [24]. This consists of modifying the nonlinear program (1.1) by relaxing the constraints and adding a penalty term to the objective function.

When the elastic mode is implemented, only the nonlinear constraints are relaxed [24]. To represent this situation in our approach, we assume that $\tilde{g}_i(x)$ for $i = 1, 2, \dots, l_i$, and $\tilde{h}_j(x)$, for $j = 1, 2, \dots, l_e$, are linear.

For these constraints, we assume the following:

[B1] The set $\mathcal{F}_l$ is feasible, where

$$\mathcal{F}_l = \{x \mid \tilde{g}_i(x) \leq 0, i = 1, 2, \dots, l_i, \tilde{h}_j(x) = 0, j = 1, 2, \dots, l_e\}.$$

[B2] The preceding representation of $\mathcal{F}_l$ is minimal: $\nabla_x \tilde{h}_j(x)$ are linearly independent, $j = 1, 2, \dots, l_e$, and $\exists d$ such that $\nabla_x \tilde{h}_j(x)d = 0$, $\nabla_x \tilde{g}_i(x)d < 0$.

None of these assumptions induces any loss of generality. Indeed, if $\mathcal{F}_l = \emptyset$, then the original nonlinear program (1.1) is infeasible. Most software for nonlinear programming starts with an analysis of the linear constraints and the infeasibility of the problem, which is the correct outcome when $\mathcal{F}_l = \emptyset$ is immediately detected. Clearly, the interesting situation is when $\mathcal{F}_l$ is feasible, which is our assumption [B1].

If the set $\mathcal{F}_l$ is polyhedral and nonempty, it must have a minimal representation [43, 11]. In addition, this representation can be computed by solving only one linear program [20]. The methods we use in this work are of the sequential quadratic programming type, where all constraints of the nonlinear program are linearized, and the nonlinear constraints are perhaps relaxed. Since the set $\mathcal{F}_l$ is invariant under linearization, any of its representations will produce the same quadratic program subproblems. As long as we do not involve the Lagrange multipliers of the constraints defining $\mathcal{F}_l$, that are not invariant to a change in representation, assumption [B2] does not result in any loss of generality as long as assumption [B1] holds.

Depending on the type of relaxation, we can have either an L_1 or an L_∞ approach. Our results are related to the situation in which the merit functions $\psi_1(x)$ and $\psi_\infty(x)$ are exact for the nonlinear program (1.1).

TABLE 3.1
The model algorithm.

Define $\tilde{c}_\infty > 0$, x^0 , $k = 0$, $\sigma \in (0, \frac{1}{2})$, $\tau \in (0, 1)$, $s > 0$.

QP Find the solution $d = d^k$ of the quadratic program.

$$\begin{aligned} & \text{minimize}_d \quad \frac{1}{2} d^T d + \nabla_x \tilde{f}(x^k) d \\ & \text{subject to} \quad \tilde{h}(x^k) + \nabla_x \tilde{h}(x^k) d = 0 \\ & \quad \quad \tilde{g}(x^k) + \nabla_x \tilde{g}(x^k) d \leq 0. \end{aligned}$$

Find the smallest integer $m = m^k$ that satisfies

$$\psi_\infty(x^k + \tau^m s d^k) - \psi_\infty(x^k) \geq \sigma \tau^m s d^{k^T} d^k.$$

Define $x^{k+1} = x^k + \tau^{m_k} s d^k$, and $k = k + 1$.

Go to **QP**.

Here we consider the case in which the added penalty term is of the L_∞ [3] type:

$$(3.2) \quad \begin{aligned} & \min_{x, \zeta} \quad \tilde{f}(x) + c_\infty \zeta \\ & \text{subject to} \quad \tilde{g}_i(x) \leq 0, \quad i = 1, 2, \dots, l_i, \\ & \quad \quad \tilde{g}_i(x) \leq \zeta, \quad i = l_i + 1, \dots, m, \\ & \quad \quad \tilde{h}_j(x) = 0, \quad j = 1, 2, \dots, l_e, \\ & \quad \quad -\zeta \leq \tilde{h}_j(x) \leq \zeta, \quad j = l_e + 1, \dots, r, \\ & \quad \quad \zeta \geq 0. \end{aligned}$$

All the constraints are now inequality constraints. A quadratic program analogous to (3.1) is constructed for (3.2), which, since [B1] and [B2] hold, now satisfies (MFCQ) at any feasible point. A feasible point of (3.2) can be immediately obtained by choosing ζ to be sufficiently large.

We make specific claims about one algorithm, presented in Table 3.1. The algorithm, including the line-search rule, is presented in [3, 4]. The algorithm is not necessarily practical, but it serves to show that rates of convergence results can be obtained under very general assumptions. We now define the algorithm for the general nonlinear program (1.1), though we later applied it to (3.2).

For fixed penalty parameter c_∞ , the problem (3.2) can be approached by the above SQP algorithms without resulting in an infeasible QP, since the linearization of the problem (3.2) is always feasible if [B1] and [B2] hold. If for a solution of (3.2) we have that $\zeta = 0$, then the x component of the solution of (3.2) is also a solution of the original, unrelaxed nonlinear program (1.1).

The possibility exists, however, that c_∞ may have to be increased indefinitely before a solution of (1.1) is obtained. In the following theorem we discuss sufficient conditions that ensure that the elastic mode relaxation (3.2) has x^* as a component of the solution for sufficiently large but finite penalty parameter.

THEOREM 3.1. *Assume that assumptions [B1] and [B2] hold for (1.1) and that, at a solution x^* of (1.1), we have that*

- the Lagrange multiplier set of (1.1) is not empty,
- the quadratic growth condition (1.16) is satisfied at x^* , and
- the data of (1.1) are twice continuously differentiable.

Then, for sufficiently large but finite value of the penalty parameter c_∞ , we have that

1. $(x^*, 0)$ is a local minimum of (3.2) at which both (MFCQ) and the quadratic growth condition (1.16) are satisfied.
2. $(x^*, 0)$ is an isolated stationary point of (3.2).

3. If the algorithm in Table 3.1 is initialized sufficiently close to $(x^*, 0)$ with a sufficiently large penalty parameter $\tilde{c}_\infty$, then the sequence x^k of iterates converges R -linearly to x^* .

Proof. We define the fully relaxed nonlinear program

$$(3.3) \quad \begin{array}{ll} \min_{x, \zeta} & \tilde{f}(x) + c_\infty \zeta \\ \text{subject to} & \tilde{g}_i(x) \leq \zeta, \quad i = 1, 2, \dots, m, \\ & -\zeta \leq \tilde{h}_j(x) \leq \zeta, \quad j = 1, 2, \dots, r, \\ & \zeta \geq 0. \end{array}$$

If (x, ζ) is a feasible point of (3.3), it immediately follows from the definition (1.14) of the L_∞ penalty function, $\tilde{P}_\infty(x)$, that $\zeta \geq \tilde{P}_\infty(x)$. From (1.19), under the assumptions stated in this theorem, we have that there exists $\tilde{c}_\infty > 0$ such that the penalty function $\psi_\infty(x)$ satisfies a quadratic growth condition at x^* . Choose now

$$c_\infty = \tilde{c}_\infty + 1.$$

Using (1.19), we obtain that, in a sufficiently small neighborhood of x^* , we must have that

$$\tilde{f}(x) + \tilde{c}_\infty \zeta \geq \tilde{f}(x) + \tilde{c}_\infty \tilde{P}_\infty(x) \geq \sigma_1 \|x - x^*\|^2.$$

Whenever $\zeta \leq \frac{1}{\sigma_1}$, we will have that $\sigma_1 \zeta^2 \leq \zeta$. Therefore, in a sufficiently small neighborhood of $(x^*, 0)$, for all (x, ζ) feasible, we will have that

$$\tilde{f}(x) + c_\infty \zeta = \tilde{f}(x) + \tilde{c}_\infty \zeta + \zeta \geq \sigma_1 (\|x - x^*\|^2 + \zeta^2).$$

Therefore, for our choice of c_∞ we have that (3.3) satisfies the quadratic growth condition for feasible points (x, ζ) . Since any feasible point of (3.2) is feasible for (3.3), it follows that (3.2) also satisfies quadratic growth at $(x^*, 0)$ for every feasible point (x, ζ) . Since (3.2) clearly satisfies (MFCQ) everywhere from assumption [B2], this is equivalent to the quadratic growth condition (1.16) holding for all (x, ζ) in a neighborhood of $(x^*, 0)$ [6, 7]. The proof of part 1 of the theorem is complete.

From the conclusion of part 1 we have that, since (MFCQ) and the quadratic growth condition hold for (3.2) at $(x^*, 0)$, this point must be isolated stationary points of the respective nonlinear programs [1]. This concludes the proof of part 2.

Part 3 immediately follows from [1] since (3.2) satisfies the quadratic growth condition and (MFCQ) at $(x^*, 0)$. Note that $\tilde{c}_\infty$ that enters the definition of ψ_∞ used in the algorithm is not the same as the one in the proof of part 1. It can be shown that now we need $\tilde{c}_\infty \geq c_\infty$, once the latter is chosen. $\square$

Discussion.

- The same conclusions hold for an L_1 elastic mode, such as the one in SNOPT [24].

- Determining that a solution point is an isolated stationary point is an important issue in nonlinear programming [39, 14, 7]. In practical terms, it means that a nonlinear programming algorithm with global convergence safeguards that does not leave a neighborhood of the solution point x^* will in fact converge to x^* . Examples of such algorithms are provided in [8, 9, 10, 13, 29, 17].

- A difficulty with the definition of the algorithm is that the successful completion of the algorithm depends on the choice of the parameters c_∞ , $\tilde{c}_\infty$, c_1 , that need to be sufficiently large but finite. Note that here c_∞ and $\tilde{c}_\infty$ have different purposes:

c_∞ is needed to enforce $\zeta = 0$ at a solution of (3.2), whereas $\tilde{c}_\infty$ is the parameter of the merit function $\tilde{\psi}_\infty(x)$ when the preceding algorithm is applied to (3.2). The two parameters are related, but we need at least $\tilde{c}_\infty > c_\infty$ if we follow the proof of linear convergence from [1]. If (MFCQ) does not hold, the usual update [2, 4], that depends on the Lagrange multipliers, for the value of the penalty parameter cannot be used to adapt c_∞ , if the original value is insufficient to result in $\zeta = 0$. An adaptive elastic mode can be implemented where if ζ is not sufficiently small; then the penalty parameters are increased [24]. In section 4 we show that such an update works under certain conditions.

- When applied to (MPCC), Theorem 3.1, part 1 is close in aim to [40, Theorem 8]. Note, however, that the cited result is rather an analysis tool since it involves a nonlinear program whose setup requires the knowledge of the various active sets at the solution. In particular, a nonlinear penalty term appears in the objective function and involves only the portion of the complementarity constraints corresponding to the complementarity constraints with $k \in \bar{\mathcal{K}}(x^*)$. Such information is not available in an algorithmic framework and is not included in our approach.

We now apply Theorem 3.1 for the case of interest in this work, MPCC. The following corollary is a simple restatement of Theorem 3.1 for (MPCC).

COROLLARY 3.2. *Assume that (MPCC) satisfies the following conditions at a solution x^* :*

- *The Lagrange multiplier set of (MPCC) is not empty. From Corollary 2.3, (SMFCQ) holding for (TNLP) is a sufficient condition for this assumption to hold.*
- *The quadratic growth condition (1.16) is satisfied at x^* .*
- *The data of (MPCC) are twice continuously differentiable.*

Then the conclusions of Theorem 3.1 hold for (MPCC) and for (2.13).

Proof. From Theorem 3.1 the conclusion immediately applies for (MPCC). For (2.13) we apply Theorem 2.5 followed by Theorem 3.1 to obtain the conclusion. $\square$

Consequently, when started sufficiently close to a solution and with a sufficiently large penalty parameter, the algorithm will converge to that solution of (MPCC) or (2.13) with a sufficiently large but finite c_∞ and $\tilde{c}_\infty$ as soon as (MPCC) satisfies the quadratic growth condition and has a nonempty Lagrange multiplier set at a solution x^* . Since (SMFCQ) is a generic condition for (MPCC) and holds with probability 1 for instances of problems in the MPCC class [40] and the quadratic growth condition is the weakest second-order sufficient condition, this convergence property is expected to hold with probability 1.

3.1. Numerical experiments. We conducted some numerical experiments on MPCCs from the collection MacMPEC of Sven Leyffer and used SNOPT [24], an SQP package that implements an adaptive L_1 elastic mode approach.

We considered three types of problems, all of which appear in [36].

1. Stackelberg games [36, section 12.1], the **gnash** problems.
2. Generalized Nash complementarity points [36, section 12.2], the **gne** problem, an instance of the problem 12.34 in [36].
3. Optimum packaging problem. With the two-dimensional elliptic operator, discretized on a grid of 8×8 , 16×16 , and 32×32 elements, which are the problems **pack-rig** followed by the discretization index in our table.

With the exception of **gne**, all the problems have the complementarity constraints lumped together as one inequality, as in the formulation (2.13).

In the tables showing the results, we indicate the number of variables, constraints, and complementarity constraints (Var-Con-CC in the first column), the final value of

TABLE 3.2
Results obtained with SNOPT.

Problem	Var-Con-CC	Value	Status	Feval	Elastic, Why
gnash14	21-13-1	-0.17904	Optimal	27	Yes, Inf
gnash15	21-13-1	-354.699	Optimal	12	None
gnash16	21-13-1	-241.441	Optimal	7	None
gnash17	21-13-1	-90.7491	Optimal	9	None
gne	16-17-10	0	Optimal	10	Yes, Inf
pack-rig1-8	89-76-1	0.721818	Optimal	15	None
pack-rig1-16	401-326-1	0.742102	Optimal	21	None
pack-rig1-32	1697-1354-1	0.751564	Optimal	19	None

the objective function, the number of function evaluations and the final status of the run. SNOPT was run through NEOS server [35] at Argonne National Laboratory. We also indicate if the elastic mode was started and the reason. In both examples for which the elastic mode was started, this happened because of an infeasible subproblem, a fact that we indicated in the *Elastic, Why* column of Table 3.2 by an “Inf” symbol.

SNOPT solved all the problems presented in a reasonable number of iterations. The fact that the elastic mode was initiated for SNOPT shows that the use of the elastic mode considerably increases the robustness of sequential quadratic programs, since otherwise SNOPT would have failed with an *Infeasible* diagnostic.

The fact that the MPCC problem class is substantially more difficult for nonlinear program solvers, especially interior point solvers, was validated by an extensive computational investigation [16]. The elastic mode approach, like the one implemented by SNOPT, is guaranteed to succeed for a finite penalty parameter under the conditions discussed in this paper.

4. A superlinearly convergent algorithm for MPCC. In the following, we present a superlinear convergence result for a special but widely encountered type of algorithm that uses exact second derivatives. Here we extend the work in [18], in that we relax one assumption that was critical to the convergence proof: that the iterates satisfy either the complementarity constraints or the feasibility conditions for the unrelaxed quadratic program. We will show that in certain distinguished cases, an adaptive elastic mode approach can be used to induce either feasibility or complementarity for all iterates from which superlinear convergence follows from [18].

We assume that the complementary variables are the last components of the unknown vector x . This is not a restrictive assumption: any MPCC can be recast in such a form by using slack variables [18].

Consider the MPCC that satisfies this assumption for the (2.13) form of (MPCC):

$$\begin{aligned}
 & \text{minimize} && f(x) \\
 & \text{subject to} && g_i(x) \leq 0, \quad i = 1, 2, \dots, n_i, \\
 & && h_j(x) = 0, \quad j = 1, 2, \dots, n_j, \\
 (4.1) \quad & \text{(MPCCS)} && x_{k1} \leq 0, \quad k = 1, 2, \dots, n_c, \\
 & && x_{k2} \leq 0, \quad k = 1, 2, \dots, n_c, \\
 & && \sum_{k=1}^{n_c} x_{1k} x_{2k} \leq 0.
 \end{aligned}$$

To simplify the subsequent notation, we assume that we are interested in local convergence to a solution x^* that satisfies

$$(4.2) \quad \begin{aligned} x_{k1}^* &= 0, & k &= 1, 2, \dots, n_c, & x_{k2}^* &= 0, & k &= 1, 2, \dots, n_d, \\ x_{k2}^* &< 0, & k &= n_{d+1}, \dots, n_c. \end{aligned}$$

We also denote by $x_1 = (x_{11}, x_{21}, \dots, x_{n_c1})$ and by $x_2 = (x_{12}, x_{22}, \dots, x_{n_c2})$. With this notation, the complementarity constraint becomes $x_1^T x_2 = 0$, or, equivalently over the feasible set of (MPCCS), $x_1^T x_2 \leq 0$.

Note that here we use a more specialized notation convention for the various index sets, same as in Fletcher et al. [18], that is related to the objects defined in (1.22)–(1.26), for which the notation is the same as in [40].

To prove our convergence results, we will invoke stronger conditions than in our preceding sections. One such condition is MPEC-LICQ. We say that (MPCCS) satisfies MPEC-LICQ at x^* , if the associated relaxed nonlinear program satisfies the linear independence constraint qualification (LICQ) at x^* . Specifically, the condition is expressed as

$$\text{MPEC-LICQ:} \quad \begin{aligned} &\nabla_x g_i(x^*)|_{i \in \mathcal{A}(x^*)}, \quad \nabla_x h_j(x^*)|_{j=1,2,\dots,n_e}, \quad e_{k1}|_{k=1,2,\dots,n_c}, \\ &e_{k2}|_{k=1,2,\dots,n_d}, \quad \text{are linearly independent.} \end{aligned}$$

As opposed to the preceding sections, where a symbol e was used to denote the vector of all ones, in this section we denote by $e_* \in \mathcal{R}^n$ a vector that has zeros everywhere, except in the $*$ position, where it has a 1. We also denote by $\mathcal{A}(x^*)$ the set of inequality constraints that is active:

$$\mathcal{A}(x^*) = \{i \mid i = 1, 2, \dots, n_i, \quad g_i(x^*) = 0\}.$$

For an arbitrary point x , we denote by

$$\mathcal{A}(x) = \{i \mid i = 1, 2, \dots, n_i, \quad g_i(x) \geq 0\}.$$

The associated (RNLP) is

$$(4.3) \quad \begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && g_i(x) \leq 0, \quad i = 1, 2, \dots, n_i, \\ &(\text{RNLPs}) && h_j(x) = 0, \quad j = 1, 2, \dots, n_j, \\ & && x_{k1} \leq 0, \quad k = 1, 2, \dots, n_d, \\ & && x_{k2} \leq 0, \quad k = 1, 2, \dots, n_d, \\ & && x_{k1} = 0, \quad k = n_{d+1}, \dots, n_c, \end{aligned}$$

and, if x^* is a local solution of (MPCCS) and MPEC-LICQ holds at x^* , then x^* is a KKT point of (MPCCS) [40] as well as a local solution of (RNLPs). If MPEC-LICQ holds at x^* , we have that LICQ holds for (RNLPs) at x^* , and (RNLPs) has a unique Lagrange multiplier at the solution $\tilde{\pi} = (\tilde{\nu}, \tilde{\pi}, \tilde{\mu}, \tilde{\eta})$, that satisfies

$$\begin{aligned} &\tilde{\nu} \geq 0, \quad \tilde{\mu}_{k1} \geq 0, \quad \tilde{\mu}_{k2} \geq 0, \quad k = 1, 2, \dots, n_d \\ &-\nabla_x f(x^*) = \sum_{i \in \mathcal{A}(x^*)} \tilde{\nu}_i \nabla_x g_i(x^*) + \sum_{j=1}^{n_e} \nabla_x h_j(x^*) \tilde{\pi}_j \\ &\quad + \sum_{k=1}^{n_d} (\tilde{\mu}_{k1} e_{k1} + \tilde{\mu}_{k2} e_{k2}) + \sum_{k=n_{d+1}}^{n_c} \tilde{\mu}_{k1} e_{k1}. \end{aligned}$$

Here we use the natural implicit convention that $\tilde{\nu}_i = 0$ for $i \notin \mathcal{A}(x^*)$, which we will use at other points since this way we can treat the multiplier as a vector that satisfies the complementarity constraints corresponding to the optimality conditions. However, at times we will refer to $\tilde{\nu}$ or similar quantities as to a vector with n_i components, especially in a solution stability context.

When the MPEC-LICQ assumption holds, we can define strong second-order sufficient conditions,

$$s^T \nabla_{xx}^2 \mathcal{L}^* s > 0, \forall s \neq 0, s \in \mathcal{C}_{RNLPs}, \text{ where} \\ \text{(MPEC-SOSC)} \quad \nabla_{xx}^2 \mathcal{L}^* \text{ is the Hessian of the Lagrangian evaluated at } (x^*, \tilde{\nu}, \\ \tilde{\pi}, \tilde{\mu}, \tilde{\eta}) \text{ and } \mathcal{C}_{RNLPs} \text{ is the critical cone of (RNLPs) at } x^*.$$

In the rest of this work we invoke the following assumptions:

- [A1] f, g, h are twice continuously differentiable.
- [A2] (MPCCS) satisfies MPEC-LICQ at the solution x^* .
- [A3] (MPCCS) satisfies MPEC-SOSC at the solution x^* .
- [A4] We have that $\tilde{\mu}_{k1} > 0$ for $k = 1, 2, \dots, n_d$.

Note that assumption [A4] is equivalent to the following assumption:

- [A4b] Either $\tilde{\mu}_{k1} > 0$ or $\tilde{\mu}_{k2} > 0$, for $k = 1, 2, \dots, n_d$.

All one has to do to recover assumption [A4] is to relabel appropriately the variables that appear in complementary pairs. The form of assumption [A4] that we consider here will allow us to simplify the notation considerably in our analysis.

At the point in our analysis where we invoke results from [18], we will replace assumption [A4] by a stronger assumption, as well as use an assumption about the solution of the quadratic programs.

- [A4a] $\tilde{\nu}_i > 0, i \in \mathcal{A}(x^*), \tilde{\pi}_j \neq 0, j = 1, 2, \dots, n_e$, and $\tilde{\mu}_{k1} > 0, \tilde{\mu}_{k2} > 0$ for $k = 1, 2, \dots, n_d$.

- [A5] When a QP is solved, the QP solver picks a linearly independent basis.

Assumptions [A1]–[A3] and [A5] are identical to the assumptions with the same name used in [18]. Assumption [A4a] is identical to Assumption [A4] from [18].

Note, however, that at no point in this work will we invoke either Assumption [A6] (that the current point satisfies the complementarity constraint $x_1^T x_2 = 0$) or Assumption [A7] (that all subproblems are feasible) from the same reference. As we later show, for suitable values of the penalty parameter, the elastic mode approach that we present here will induce one of the situations covered by these two assumptions, without the need to assume it from the outset.

We now explore in more details the relationship between the Lagrange multipliers and the second-order conditions of (RNLPs), (MPCCS), and of an elastic mode relaxation of (MPCCS).

In the following analysis, an important role will be played by the fact that we will assume that the elastic mode implementation enforces the linear constraints exactly. To that end, we assume that

$$g_i(x), \quad i = 1, 2, \dots, l_i, \quad h_j(x), \quad j = 1, 2, \dots, l_e, \text{ are linear functions.}$$

This is not a restrictive assumption: it is only a bookkeeping convention, since we allow l_i and l_e to be 0. Also note, following from our assumption [A2], that MPEC-LICQ holds at x^* , that we have that the assumptions [B1] and [B2] that were used for the elastic mode defined in section 3 will automatically hold here.

We now present an L_∞ elastic mode approach. An L_1 approach with similar properties can be defined, exactly as used in SNOPT [24]. We use the L_∞ approach because of notational convenience: only one extra variable is needed. In this case, the relaxed MPCC becomes

$$\begin{aligned}
 & \text{minimize} && f(x) + c_\infty \zeta \\
 & \text{subject to} && g_i(x) \leq 0, \quad i = 1, 2, \dots, l_i, \\
 & && g_i(x) \leq \zeta, \quad i = l_i + 1, \dots, n_i, \\
 & && h_j(x) = 0, \quad j = 1, 2, \dots, l_e, \\
 \text{(MPCC(c))} & && -\zeta \leq h_j(x) \leq \zeta, \quad j = l_e + 1, \dots, n_e, \\
 & && x_{k1} \leq 0, \quad k = 1, 2, \dots, n_c, \\
 & && x_{k2} \leq 0, \quad k = 1, 2, \dots, n_c, \\
 & && \sum_{k=1}^{n_c} x_{k1} x_{k2} \leq \zeta, \\
 & && \zeta \geq 0.
 \end{aligned}$$

When MPEC-LICQ holds, we can say more about the relationship between the multipliers of (RNLPS), (MPCCS), and (MPCC(c)) than we did in section 2. The Lagrange multiplier set of (MPCCS) is not empty at x^* , following [40]. Therefore, there must exist ν_i , $i \in \mathcal{A}(x^*)$; π_j , $j = 1, 2, \dots, n_e$; μ_{k1} , for $k = 1, 2, \dots, n_c$; and μ_{k2} for $k = 1, 2, \dots, n_d$; and $\eta \in \mathcal{R}$, that satisfy the KKT conditions for (MPCCS),

$$\begin{aligned}
 -\nabla_x f(x^*) &= \sum_{i \in \mathcal{A}(x^*)} \nu_i \nabla_x g_i(x^*) + \sum_{j=1}^{n_e} \nabla_x h_j(x^*) \pi_j \\
 &+ \sum_{k=1}^{n_d} (\mu_{k1} e_{k1} + \mu_{k2} e_{k2}) + \sum_{k=n_d+1}^{n_c} \mu_{k1} e_{k1} + \eta \sum_{k=n_d+1}^{n_c} x_{k2}^* e_{k1},
 \end{aligned}$$

as well as the inequality constraints $\nu_i \geq 0$, $i \in \mathcal{A}(x^*)$, $\mu_{k1} \geq 0$, $k = 1, 2, \dots, n_c$, $\mu_{k2} \geq 0$, $k = 1, 2, \dots, n_d$, and $\eta \geq 0$. Here η is the Lagrange multiplier of the complementarity constraint of (MPCCS) $x_1^T x_2 \leq 0$.

One immediate consequence is that (MPCCS) has a Lagrange multiplier that is minimal (in terms of the 1 norm) [18]. We call that multiplier the *fundamental multiplier*. Comparing the algebraic expression of its components in terms of the components of the multiplier of (RNLPS) we obtain

$$\begin{aligned}
 \nu^* &= \tilde{\nu}, \\
 \pi^* &= \tilde{\pi}, \\
 \mu_{k1}^* &= \tilde{\mu}_{k1} && \geq 0, \quad k = 1, 2, \dots, n_d, \\
 \mu_{k2}^* &= \tilde{\mu}_{k2} && \geq 0, \quad k = 1, 2, \dots, n_d, \\
 \mu_{k1}^* &= \tilde{\mu}_{k1} - \eta^* x_{k2}^* && \geq 0, \quad k = n_d + 1, \dots, n_c, \\
 \eta^* &= \max \left\{ 0, \max_{k=n_d+1, \dots, n_c} \left\{ \frac{\tilde{\mu}_{k1}}{x_{k2}^*} \right\} \right\} \geq 0.
 \end{aligned}
 \tag{FMC'}$$

Since MPEC-LICQ holds, it is immediate that (MPCCS) has a degeneracy of order 1 and thus its Lagrange multiplier set will have dimension at most 1. Therefore

any multiplier of (MPCCS) must satisfy

$$(4.4) \quad \begin{aligned} \nu &= \nu^*, \\ \pi &= \pi^*, \\ \mu_{k1} &= \tilde{\mu}_{k1}^* \geq 0, & k = 1, 2, \dots, n_d, \\ \mu_{k2} &= \tilde{\mu}_{k2}^* \geq 0, & k = 1, 2, \dots, n_d, \\ \mu_{k1} &= \tilde{\mu}_{k1}^* - ax_{k2} \geq 0, & k = n_d + 1, \dots, n_c, \\ \eta &= \eta^* + a \geq 0, \end{aligned}$$

where $a \geq 0$.

We now write the KKT conditions for the relaxed problem (MPCC(c)) at $(x^*, 0)$, assuming that $(x^*, 0)$ is a stationary point of (MPCC(c)), an assumption for which we will later determine sufficient conditions for it to hold. We have that, for a Lagrange multiplier of MPCC(c) with components $\hat{\nu}_i$, $i \in \mathcal{A}(x^*)$; $\hat{\pi}_j$, for $j = 1, 2, \dots, l_e$; $\hat{\pi}_j^+$, associated with the inequality $h_j(x) \leq \zeta$ and $\hat{\pi}_j^-$, associated with the inequality $-h_j(x) \leq \zeta$, $j = l_e + 1, \dots, n_e$; $\hat{\mu}_{k1}$ for $k = 1, 2, \dots, n_c$; $\hat{\mu}_{k2}$ for $k = 1, 2, \dots, n_d$; and $\hat{\eta}$, associated with the complementarity constraint $x_1^T x_2 \leq \zeta$; and $\hat{\theta} \geq 0$, associated with the inequality $\zeta \geq 0$, we can write the KKT conditions:

$$\begin{aligned} -\nabla_x f(x^*) &= \sum_{i \in \mathcal{A}(x^*)} \hat{\nu}_i \nabla_x g_i(x^*) + \sum_{j=1}^{l_e} \nabla_x h_j(x^*) \hat{\pi}_j \\ &\quad + \sum_{j=l_e+1}^{n_e} \nabla_x h_j(x^*) (\hat{\pi}_j^+ - \hat{\pi}_j^-) + \sum_{k=1}^{n_d} (\hat{\mu}_{k1} e_{k1} + \hat{\mu}_{k2} e_{k2}) \\ &\quad + \sum_{k=n_d+1}^{n_c} \hat{\mu}_{k1} e_{k1} + \hat{\eta} \sum_{k=n_d+1}^{n_c} (x_{k2}^* e_{k1}), \\ c_\infty &= \sum_{i=l_i+1, \dots, n_i, i \in \mathcal{A}(x^*)} \hat{\nu}_i + \sum_{j=l_e+1}^{n_e} (\hat{\pi}_j^+ + \hat{\pi}_j^-) + \hat{\eta} + \hat{\theta}. \end{aligned}$$

If MPEC-LICQ holds, and $(x^*, 0)$ is a stationary point of (MPCC(c)), then using (4.4) we obtained that any Lagrange multiplier of (MPCC(c)) must satisfy, in terms of the components of the fundamental multiplier, the following relations:

$$(4.5) \quad \begin{aligned} \hat{\nu}_i &= \nu_i^* \geq 0, & i \in \mathcal{A}(x^*), \\ \hat{\pi}_j &= \pi_j^*, & j = 1, 2, \dots, l_e, \\ \hat{\pi}_j^+ &= \max\{\pi_j^*, 0\} + f_j \geq 0, & j = l_e + 1, \dots, n_e, \\ \hat{\pi}_j^- &= \max\{-\pi_j^*, 0\} + f_j \geq 0, & j = l_e + 1, \dots, n_e, \\ \hat{\mu}_{k1} &= \mu_{k1}^* \geq 0, & k = 1, 2, \dots, n_d, \\ \hat{\mu}_{k2} &= \mu_{k2}^* \geq 0, & k = 1, 2, \dots, n_d, \\ \hat{\mu}_{k1} &= \mu_{k1}^* - ax_{k2} \geq 0, & k = n_d + 1, \dots, n_c, \\ \hat{\eta} &= \eta^* + a \geq 0, \\ \hat{\theta} &= c_\infty - \sum_{i=l_i+1, i \in \mathcal{A}(x^*)}^{n_i} \hat{\nu}_i^* \\ &\quad - \sum_{j=l_e+1}^{n_e} (\hat{\pi}_j^+ + \hat{\pi}_j^-) - \hat{\eta} \geq 0, \end{aligned}$$

where $a \geq 0$, and $f_j \geq 0$, for $j = l_e + 1, \dots, n_e$. Here $\hat{\theta}$ is the Lagrange multiplier of the constraint $\zeta \geq 0$ and the requirement $\hat{\theta} \geq 0$ results in a condition on the

value that c_∞ must take for $(x^*, 0)$ to be a stationary point of (MPCC(c)). We do not add the complementarity constraints that appear when applying the first-order optimality condition, since these are automatically satisfied by our choice of active set and multiplier components.

Rewriting $\hat{\theta}$ in terms of the fundamental multiplier (FMC), we obtain that

$$(4.6) \quad \hat{\theta} = c_\infty - \sum_{i=l_e+1, i \in \mathcal{A}(x^*)}^{n_i} \nu_i^* - \sum_{j=l_e+1}^{n_e} (|\pi_j^*| + 2f_j) - \eta^* - a \geq 0.$$

We denote by ν_0 the following quantity, that is also defined in terms of the components of the fundamental multiplier (FMC),

$$(4.7) \quad \nu_0 = \sum_{i=l_e+1, i \in \mathcal{A}(x^*)}^{n_i} \nu_i^* + \sum_{j=l_e+1}^{n_e} |\pi_j^*| + \eta^*.$$

We have the following lemma.

LEMMA 4.1. *If $c_\infty \geq \nu_0$, where ν_0 is given by (4.7), then*

1. *the point $(x^*, 0)$ is a stationary point of (MPCC(c));*
2. *the set of Lagrange multipliers of (MPCC(c)) is not empty and is defined by (4.5), where $a \geq 0$ and $f_j \geq 0$, $j = l_e + 1, \dots, n_e$, also satisfy*

$$(4.8) \quad c_\infty - \nu_0 \geq \sum_{j=l_e+1}^{n_e} (2f_j) + a.$$

Proof. Since the values of the fundamental multiplier are fixed, and since $a \geq 0$ and $f_j \geq 0$, $j = l_e + 1, \dots, n_e$, the necessary and sufficient condition for $(x^*, 0)$ to be a stationary point of (MPCC(c)) is that $\hat{\theta} \geq 0$. From (4.6) and (4.7) this is equivalent to

$$c_\infty \geq \nu_0,$$

which proves the first part of the claim.

By inspecting (4.5) we see that the set of Lagrange multipliers of (MPCC(c)) is defined by $f_j \geq 0$, $j = l_e + 1, \dots, n_e$, and $a \geq 0$ that satisfy the inequality $\hat{\theta} \geq 0$, that is,

$$c_\infty - \nu_0 \geq \sum_{j=l_e+1}^{n_e} (2f_j) + a.$$

The proof is complete. $\square$

An immediate consequence is that, if $c_\infty = \nu_0$, then the program (MPCC(c)) has a unique multiplier! This is formally stated in the next result.

4.1. Second-order conditions.

LEMMA 4.2. *Assume that (MPCCS) satisfies MPEC-LICQ and MPEC-SOSC. Assume that c_∞ satisfies $c_\infty \geq \nu_0$. Then (MPCC(c)) satisfies (MFCQ) and (RSOSC) at $(x^*, 0)$. In addition, if $c_\infty = \nu_0$, then the Lagrange multiplier set of (MPCC(c)) at $(x^*, 0)$ has a unique element.*

Proof. Consider one multiplier of (MPCC(c)), whose components are $\hat{\nu}_i \geq 0$ for $i \in \mathcal{A}(x^*)$; $\hat{\pi}_j$ for $j = 1, 2, \dots, l_e$; $\hat{\pi}_j^+ \geq 0$ and $\hat{\pi}_j^- \geq 0$ for $j = l_e + 1, \dots, n_e$; $\hat{\mu}_{k1} \geq 0$ and $\hat{\mu}_{k2} \geq 0$ for $k = 1, 2, \dots, n_d$; $\hat{\mu}_{k1} \geq 0$ for $k = 1, 2, \dots, n_d$; $\hat{\eta} \geq 0$ and $\hat{\theta} \geq 0$.

The Hessian of the Lagrangian at x^* corresponding to this Lagrange multiplier, accounting for only the nonlinear terms, and using (FMC) and (4.5), is the following:

$$\begin{aligned}
\nabla_{xx}^2 \mathcal{L}_{MPCC(c)}(x^*, 0) &= \nabla_{xx}^2 f(x^*) + \sum_{j=l_e+1}^{n_e} (\hat{\pi}_j^+ \nabla_{xx}^2 h_j(x^*) - \hat{\pi}_j^- \nabla_{xx}^2 h_j(x^*)) \\
&\quad + \sum_{i=l_i+1, i \in \mathcal{A}(x^*)}^{n_i} \hat{\nu}_i \nabla_{xx}^2 g(x^*) + \hat{\eta} \sum_{k=1}^{n_c} (e_{k1}^T e_{k2} + e_{k2}^T e_{k1}) \\
&= \nabla_{xx}^2 f(x^*) + \sum_{j=l_e+1}^{n_e} \pi_j^* \nabla_{xx}^2 h_j(x^*) \\
&\quad + \sum_{i=l_i+1, i \in \mathcal{A}(x^*)}^{n_i} \nu_i^* \nabla_{xx}^2 g(x^*) + \hat{\eta} \sum_{k=1}^{n_c} (e_{k1}^T e_{k2} + e_{k2}^T e_{k1}) \\
&= \nabla_{xx}^2 \mathcal{L}_{RNLPs}(x^*) + \hat{\eta} \sum_{k=1}^{n_c} (e_{k1}^T e_{k2} + e_{k2}^T e_{k1}).
\end{aligned}$$

The Hessian of the Lagrangian of (MPCC(c)) should also be computed with respect to ζ , but, since the contribution of ζ is linear both in the constraints and in the objective function of (MPCC(c)), it follows that the Hessian of the Lagrangian is

$$\nabla_{(x,\zeta)(x,\zeta)}^2 \mathcal{L}_{MPCC(c)}(x^*, 0) = \begin{pmatrix} \nabla_{xx}^2 \mathcal{L}_{MPCC(c)}(x^*, 0) & 0_{n \times 1} \\ 0_{1 \times n} & 0 \end{pmatrix}.$$

Note that from Lemma 2.1 we have that $\mathcal{C}_{MPCCS} = \mathcal{C}_{RNLPs}$ at x^* , where by $\mathcal{C}$ we denote the critical cone of the respective nonlinear program. When $c_\infty \geq \nu_0$, since x^* is a stationary point for (MPCC) and $(x^*, 0)$ is a stationary point for (MPCC(c)), we will have that $\mathcal{C}_{MPCC(c)} = \mathcal{C}_{MPCCS} \oplus \{0\}$. Let now $(u, 0) \in \mathcal{C}_{MPCC(c)}$ and thus $u \in \mathcal{C}_{RNLPs}$. We have that

$$\begin{aligned}
&(u, 0)^T \nabla_{(x,\zeta)(x,\zeta)}^2 \mathcal{L}_{MPCC(c)}(u, 0) \\
&= u^T \mathcal{L}_{MPCC(c)} u = u^T \mathcal{L}_{RNLPs} u + 2\hat{\eta} \sum_{k=1}^{n_c} (e_{k1}^T u)(e_{k2}^T u) \geq u^T \mathcal{L}_{RNLPs} u > 0,
\end{aligned}$$

where the last two inequalities follow, respectively, from the fact that on the critical cone of (RNLPs) we have $(e_{k1}^T u)(e_{k2}^T u) \geq 0$, for $k = 1, 2, \dots, n_d$, and $(e_{k1}^T u)(e_{k2}^T u) = 0$, for $k = n_d + 1, \dots, n_c$, and $\hat{\eta} \geq 0$, and, respectively, from the MPEC-SOSC assumption. Since we obtained, for any multiplier of MPCC(c), that

$$\forall \hat{u} \in \mathcal{C}_{MPCC(c)}, \hat{u} \neq 0 \Rightarrow \hat{u}^T \nabla_{(x,\zeta)(x,\zeta)}^2 \mathcal{L}_{MPCC(c)}(x^*, 0) \hat{u} > 0,$$

this means that Robinson's condition (RSOSC) holds for (MPCC(c)) at $(x^*, 0)$. Since (MFCQ) clearly holds at $(x^*, 0)$, the conclusion of the lemma follows. The uniqueness of the multiplier for the case $c_\infty = \nu_0$ follows from (4.8). $\square$

4.2. The algorithm. We now consider the effect of applying SQP to MPCC in either the original or the relaxed form. As matrix W of the QP we use the Hessian of the Lagrangian, with the Lagrange multipliers computed at the previous step.

We define the following matrix $\widehat{W}$ at the point x :

$$\widehat{W} = \left\{ \begin{array}{l} \nabla_{xx}^2 f(x) + \sum_{i=l_i+1}^{n_i} \hat{\nu}_i \nabla_{xx}^2 g_i(x) \\ \quad + \sum_{j=l_e+1}^{n_e} (\hat{\pi}_j^+ - \hat{\pi}_j^-) \nabla_{xx}^2 h_j(x) \\ \quad + \hat{\eta} \sum_{k=1}^{n_c} (e_{k1}^T e_{k2} + e_{k2}^T e_{k1}), \\ \text{or, alternatively,} \\ \nabla_{xx}^2 f(x) + \sum_{i=l_i+1}^{n_i} \nu_i \nabla_{xx}^2 g_i(x) \\ \quad + \sum_{j=l_e+1}^{n_e} \pi_j \nabla_{xx}^2 h_j(x) \\ \quad + \eta \sum_{k=1}^{n_c} (e_{k1}^T e_{k2} + e_{k2}^T e_{k1}), \end{array} \right.$$

where the first branch is followed if the quadratic program associated with the relaxation (QPMC(c)) was solved at the preceding iteration, and the second branch is followed if the quadratic program associated with the unrelaxed problem ((QPX), to be defined later) was solved at the preceding iteration.

Consider now the following quadratic program associated to (MPCC(c)) at the point (x, ζ) :

$$(\text{QPMC(c)}) \left\{ \begin{array}{l} \min_{d, d_\zeta} \quad \frac{1}{2} d^T \widehat{W} d + \nabla_x f(x) d + c_\infty (\zeta + d_\zeta), \\ \quad \quad \quad g_i(x) + \nabla_x g_i(x) d \leq 0, \quad i = 1, 2, \dots, l_i, \\ \quad \quad \quad g_i(x) + \nabla_x g_i(x) d \leq \zeta + d_\zeta, \quad i = l_i + 1, \dots, n_i, \\ \quad \quad \quad h_j(x) + \nabla_x h_j(x) d = 0, \quad j = 1, 2, \dots, l_e, \\ \quad \quad \quad h_j(x) + \nabla_x h_j(x_k) d \leq \zeta + d_\zeta, \quad j = l_e + 1, \dots, n_e, \\ \quad \quad \quad -h_j(x) - \nabla_x h_j(x_k) d \leq \zeta + d_\zeta, \quad j = l_e + 1, \dots, n_e, \\ \text{subject} \quad \quad \quad x_{k1} + d_{k1} \leq 0, \quad k = 1, 2, \dots, n_c, \\ \text{to} \quad \quad \quad \quad \quad x_{k2} + d_{k2} \leq 0, \quad k = 1, 2, \dots, n_c, \\ \quad \quad \quad x_1^T d_2 + x_2^T d_1 + x_1^T x_2 \leq \zeta + d_\zeta, \\ \quad \quad \quad \quad \quad \zeta + d_\zeta \geq 0. \end{array} \right.$$

To obtain results for the application of the algorithm to (MPCCS), we state and prove the following stability result for general nonlinear programming.

LEMMA 4.3. *Consider the nonlinear program*

$$\begin{array}{ll} \min_x & \tilde{f}(x) \\ \text{subject to} & \tilde{h}_j(x) = 0, \quad j = 1, 2, \dots, r, \\ & \tilde{g}_i(x) \leq 0, \quad i = 1, 2, \dots, m. \end{array}$$

Assume that it satisfies (MFCQ) and (RSOSC) at a solution x^* . We denote the compact Lagrange multiplier set of this nonlinear program by $\Lambda(x^*)$.

Consider the quadratic program

$$\begin{aligned} \min_d \quad & \nabla_x \tilde{f}(x)d + \frac{1}{2}d^T \tilde{W}d \\ \text{subject to} \quad & \tilde{g}_i(x) + \nabla_x \tilde{g}_i(x)d \leq 0, \quad i = 1, 2, \dots, m, \\ & \tilde{h}_j(x) + \nabla_x \tilde{h}_j(x)d = 0, \quad j = 1, 2, \dots, r, \end{aligned}$$

where $\tilde{W} = \nabla_{xx}^2 \tilde{f} + \sum_{j=1}^{n_e} \pi_j \nabla_{xx}^2 \tilde{h}_j(x) + \sum_{i=1}^{n_i} \nu_i \nabla_{xx}^2 \tilde{g}_i(x)$. Let d be a solution of this quadratic program, and ν^+ and π^+ its Lagrange multipliers.

Then, there exist $\epsilon > 0$, $c_1 > 0$, $c_2 > 0$ such that $\text{dist}((\nu, \pi); \Lambda(x^*)) \leq \epsilon$ and $\text{dist}(x; x^*) \leq \epsilon$ imply that

- (i) $\frac{1}{c_1} \|x - x^*\| \leq \|d\| \leq c_2 \|x - x^*\|$,
- (ii) $\text{dist}((\nu^+, \pi^+); \Lambda(x^*)) \leq c_2 \|x - x^*\|$.

Here $\text{dist}(\cdot, \cdot)$ denotes the distance between two sets.

Proof. The rightmost inequalities in both parts (i) and (ii) are a consequence of [45, Theorem A1]. In that reference, the inequality-only case is treated, but the conclusion can be immediately extended for the case where there are also equality constraints, that are linearly independent and that, together with the inequality constraints, satisfy (MFCQ). Since (MFCQ) holds, we have that

$$\begin{aligned} \tilde{g}(x+d) &\leq O(\|d\|^2) \leq O(\|x - x^*\|^2), \\ \tilde{h}(x+d) &\leq O(\|d\|^2) \leq O(\|x - x^*\|^2). \end{aligned}$$

This implies that the size of the infeasibility, $\mathcal{I}(x)$, where

$$\mathcal{I}(x) = \max \left\{ \max_{i=1,2,\dots,n_i} \{\tilde{g}_i^+(x)\}, \max_{j=1,2,\dots,n_e} \{|\tilde{h}_j(x)|\} \right\}$$

satisfies $\mathcal{I}(x) = O(\|d\|)$. Using the result [45, Theorem A1], as well as $\nu^+ \geq 0$ and the optimality conditions for the QP, we obtain that, since $Wd = -\nabla_x \tilde{f}^T(x) - \nabla_x \tilde{g}^T(x)\nu^+ - \nabla_x \tilde{h}^T(x)\pi^+$, we have that

$$\begin{aligned} &\Omega(\|x - x^*\| + \text{dist}((\nu^+, \pi^+); \Lambda(x^*))) \\ &= \max\{\|\nabla_x \tilde{f}(x) + \nabla_x \tilde{g}(x)\nu^+ + \nabla_x \tilde{h}(x)\pi^+\|, \mathcal{I}(x)\} = O(\|d\|). \end{aligned}$$

In turn, this implies that there exists a $c_1 > 0$ such that $\|x - x^*\| \leq c_1 \|d\|$, which completes the proof. $\square$

In the following, we will characterize the behavior of the solution of (QPMC(c)). Towards that end, we consider the following multipliers. The multiplier

$$\hat{\Pi} = (\hat{\nu}, \hat{\pi}, \hat{\pi}^+, \hat{\pi}^-, \hat{\mu}_1, \hat{\mu}_2, \hat{\nu}, \hat{\theta})$$

represents an approximation to the multiplier of the quadratic program (QPMC(c)). The multiplier

$$\hat{\Pi}^+ = (\hat{\nu}^+, \hat{\pi}^+, \hat{\pi}^{++}, \hat{\pi}^{-+}, \hat{\mu}_1^+, \hat{\mu}_2^+, \hat{\nu}^+, \hat{\theta}^+)$$

represents the multiplier of the quadratic program (QPMC(c)) computed at (x, ζ) , with the approximation to the Hessian matrix, $\hat{W}$, computed with $\hat{\Pi}$. Finally, the multiplier

$$\hat{\Pi}^* = (\hat{\nu}^*, \hat{\pi}^*, \hat{\pi}^{+*}, \hat{\pi}^{-*}, \hat{\mu}_1^*, \hat{\mu}_2^*, \hat{\nu}^*, \hat{\theta}^*)$$

is a Lagrange multiplier of (MPCCS) at $(x^*, 0)$.

LEMMA 4.4. Consider the quadratic program (QPMC(c)), whose solution we denote by (d, d_ζ) . Assume that $c_\infty - \nu_0 = \gamma_0 > 0$. There exists $\epsilon > 0$ such that, if

$$\text{dist}(\widehat{\Pi}, \Lambda_{MPCC(c)}) < \epsilon \quad \text{and} \quad \text{dist}((x, \zeta), (x^*, 0)) < \epsilon,$$

then the following hold.

- (i) If $\zeta + d_\zeta > 0$, then $x + d$ satisfies $x_{k1} + d_{k1} = 0$, for $k = 1, 2, \dots, n_c$.
- (ii) If $x_{k1} = 0$, for $k = 1, 2, \dots, n_c$, then $\zeta + d_\zeta = 0$, and $x + d$ also satisfies $x_{k1} + d_{k1} = 0$, for $k = 1, 2, \dots, n_c$.

Note. A point that satisfies $x_{k1} = 0$, for $k = 1, 2, \dots, n_c$, also satisfies the complementarity constraints $x_1^T x_2 = 0$.

Proof of part (i). Assume that $\zeta + d_\zeta > 0$. Using Lemmas 4.2 and 4.3, it follows from the expression of the Lagrange multipliers of (MPCC(c)) at $(x^*, 0)$, (4.5), that we must have that

$$(4.9) \quad \widehat{\mu}_{k1}^+ > 0 \quad k = 1, 2, \dots, n_d.$$

The last result follows from assumption [A4].

As a result we obtain from the optimality conditions of (QPMC(c)) that

$$(4.10) \quad x_{k1} + d_{k1} = 0, \quad k = 1, 2, \dots, n_d.$$

In addition, since $\zeta + d_\zeta > 0$, not both bound constraints on $h_j(x) + \nabla_x h_j(x)d$ in (QPMC(c)) can be active at the same time. Using the optimality conditions of (QPMC(c)), this implies that $\widehat{\pi}_j^{++} \widehat{\pi}_j^{-+} = 0$ for $j = n_{l_e+1}, \dots, n_e$. Assume that $\widehat{\pi}_j^{-+} = 0$ for some $n_{l_e+1} \leq j \leq n_e$.

Let now $\widehat{\Pi}^*$ be a Lagrange multiplier of (MPCC(c)) that satisfies $\|\widehat{\Pi}^* - \widehat{\Pi}\| \leq c_2(\|x - x^*\| + |\zeta - \zeta^*|)$, where $\zeta^* = 0$. Such a multiplier must exist by Lemmas 4.2 and 4.3. It then follows that $0 \leq \widehat{\pi}_j^{*-} \leq c_2(\|x - x^*\| + |\zeta|)$. Define now $f_j^* = \widehat{\pi}_j^{*-} - \max\{-\pi_j^*, 0\}$. From the preceding two equations we must have that $f_j^* \leq c_2(\|x - x^*\| + \zeta)$, whereas from (4.5) we must have that $f_j^* \geq 0$.

Applying this argument for any $n_{l_e+1} \leq j \leq n_e$, we obtain that

$$\sum_{j=n_{l_e+1}}^{n_e} f_j^* \leq c_2 n_e (\|x - x^*\| + \zeta).$$

Finally, $\zeta + d_\zeta > 0$ implies, from the optimality conditions of (QPMC(c)), that $\widehat{\theta}^+ = 0$, and thus $0 \leq \widehat{\theta}^* \leq c_2(\|x - x^*\| + \zeta)$.

Define now $a^* = \widehat{\eta}^* - \eta^*$. Using (4.5) and the fact that $c_\infty - \nu_0 = \gamma_0 > 0$, we obtain that

$$a^* = \gamma_0 - \sum_{j=n_{l_e+1}}^{n_e} 2f_j^* - \widehat{\theta}^* \geq \gamma_0 - (2n_e + 1)c_2(\|x - x^*\| + \zeta).$$

It therefore follows that, since $\gamma_0 > 0$, if we choose $\epsilon > 0$ sufficiently small we have that $a^* \geq \frac{\gamma_0}{2}$. Using again (4.5), Lemmas 4.2 and 4.3, as well as our definition of $\widehat{\Pi}^*$, we obtain that

$$\widehat{\mu}_{k1}^+ > 0, \quad k = n_d + 1, \dots, n_c.$$

From the optimality conditions of (QPMC(c)) and (4.9), this implies that

$$x_{k1} + d_{k1} = 0, \quad k = 1, 2, \dots, n_c.$$

Thus the proof of the first part of the result is complete.

Proof of part (ii). Assume now that x satisfies $x_{k1} = 0$, $k = 1, 2, \dots, n_c$. We define

$$W = \widehat{W} - \widehat{\eta} \sum_{k=1}^{n_c} (e_{k1}^T e_{k2} + e_{k2}^T e_{k1}).$$

Consider the quadratic program

$$(QPR) \begin{cases} \min_d & \frac{1}{2} d^T W d + \nabla_x f(x) d, \\ \text{subject to} & \begin{aligned} g_i(x) + \nabla_x g_i(x) d &\leq 0, & i = 1, 2, \dots, n_i, \\ h_j(x) + \nabla_x h_j(x) d &= 0, & j = 1, 2, \dots, n_e, \\ d_{k1} &= 0, & k = 1, 2, \dots, n_c, \\ x_{k2} + d_{k2} &\leq 0, & k = 1, 2, \dots, n_c. \end{aligned} \end{cases}$$

Using assumption [A2], it is immediate that (QPR) satisfies LICQ in a neighborhood of $x = x^*$ and $d = 0$. Using Lemmas 4.2 and 4.3 it follows that as $x \rightarrow x^*$, (QPR) has a unique solution in the neighborhood of $d = 0$, and the unique corresponding Lagrange multiplier $\tilde{\Pi}_R$ of (QPR) approaches the unique Lagrange multiplier $\tilde{\Pi}$ of (RNLPs) at $x = x^*$.

Also, since we have that $d^T \widehat{W} d = d^T W d$ for any d in the feasible set of (QPR), it follows that the quadratic program that is obtained by substituting W by $\widehat{W}$ has the same solution and the same multiplier.

It then follows that d is also a stationary point of the following quadratic program:

$$(QPX) \begin{cases} \min_d & \frac{1}{2} d^T \widehat{W} d + \nabla_x f(x) d \\ \text{to} & \begin{aligned} g_i(x) + \nabla_x g_i(x) d &\leq 0 & i = 1, 2, \dots, n_i, \\ h_j(x) + \nabla_x h_j(x) d &= 0 & j = 1, 2, \dots, n_e, \\ x_{k1} + d_{k1} &\leq 0 & k = 1, 2, \dots, n_c, \\ x_{k2} + d_{k2} &\leq 0 & k = 1, 2, \dots, n_c, \end{aligned} \\ \text{subject to} & \begin{aligned} x_1^T d_2 + x_2^T d_1 + x_1^T x_2 &\leq 0 \end{aligned} \end{cases}$$

with a multiplier Π that is constructed from the multiplier $\tilde{\Pi}_R$ by the same procedure that we used to construct the fundamental multiplier (FMC) from $\tilde{\Pi}$. It therefore follows that $\Pi = (\nu, \pi, \mu_{k1}, \mu_{k2}, \eta)$ approaches the fundamental multiplier defined in (FMC) as x approaches x^* .

It then follows that we can construct from Π a Lagrange multiplier $\widehat{\Pi}$ for (QPMC(c)) by the same relation that we used in (4.5) to construct all Lagrange multipliers of (MPCC(c)) starting from the fundamental multiplier of (MPCCS), where in the same procedure we take $a = 0$.

The only relation in (4.5) that is not immediately true is $\widehat{\theta} \geq 0$. However, since we start from a multiplier that is close to the fundamental multiplier, since we take $a = 0$, and since we assume that $c_\infty - \nu_0 = \gamma_0 > 0$, it is clear that $\widehat{\theta}$ approaches γ_0 and is therefore positive when $\epsilon > 0$ is sufficiently small.

It then follows that we can choose $d_\zeta = -\zeta$ and then (d, d_ζ) is a stationary point of (QPMC(c)). However, from Lemma 4.2, it follows from [39] that the solution of

(QPMC(c)) in a neighborhood of $(0, 0)$ is local unique. Therefore (d, d_ζ) is a solution of (QPMC(c)). Since, following the definition of (QPR), for that solution we will have that $x_{k1} + d_{k1} = 0$, $k = 1, 2, \dots, n_c$ as well as $\zeta + d_\zeta = 0$, the conclusion follows. $\square$

4.3. A superlinearly convergent algorithm. We now state our algorithm for general nonlinear programming and then show that, under certain conditions, it will converge superlinearly when applied to (MPCCS). The problem to be solved is the following:

$$(4.11) \quad \begin{aligned} & \min_x \quad \tilde{f}(x) \\ & \text{subject to} \quad \tilde{g}_i(x) \leq 0, \quad i = 1, 2, \dots, m, \\ & \quad \quad \quad \tilde{h}_j(x) = 0, \quad j = 1, 2, \dots, r. \end{aligned}$$

We assume that $\tilde{g}_i(x)$, $i = 1, 2, \dots, l_i$, and $\tilde{h}_j(x)$, $j = 1, 2, \dots, l_e$, are linear constraints that do not get relaxed in an elastic mode approach.

In the algorithm that we describe we use one of the two following quadratic programs:

$$\begin{aligned} & \min_d \quad \nabla_x \tilde{f}(x)d + \frac{1}{2}d^T \tilde{W}d \\ (QP) \quad & \text{subject to} \quad \tilde{g}_i(x)d + \nabla_x \tilde{g}_i(x)^T d \leq 0, \quad i = 1, 2, \dots, m, \\ & \quad \quad \quad \tilde{h}_j(x)d + \nabla_x \tilde{h}_j(x)^T d = 0, \quad j = 1, 2, \dots, r, \\ \\ & \min_{d, d_\zeta} \quad \nabla_x \tilde{f}(x)d + \frac{1}{2}d^T \tilde{W}d + c_\infty(\zeta + d_\zeta) \\ (QPC) \quad & \text{subject to} \quad \tilde{g}_i(x)d + \nabla_x \tilde{g}_i(x)d \leq 0, \quad i = 1, 2, \dots, l_i, \\ & \quad \quad \quad \tilde{g}_i(x)d + \nabla_x \tilde{g}_i(x)d \leq \zeta + d_\zeta, \quad i = l_i + 1, \dots, m, \\ & \quad \quad \quad \tilde{h}_j(x)^T d + \nabla_x \tilde{h}_j(x)d = 0, \quad j = 1, 2, \dots, l_e, \\ & \quad \quad \quad -\zeta - d_\zeta \leq \tilde{h}_j(x)^T d + \nabla_x \tilde{h}_j(x)d \leq \zeta + d_\zeta, \quad j = l_e + 1, \dots, r, \\ & \quad \quad \quad \zeta + d_\zeta \geq 0. \end{aligned}$$

When (QP) is solved, we obtain a direction d and Lagrange multipliers ν_i , $i = 1, 2, \dots, n_i$ and π_j , for $j = 1, 2, \dots, n_e$. When (QPC) is solved, we obtain Lagrange multipliers ν_i , $i = 1, 2, \dots, n_i$; π_j^+ , $j = 1, 2, \dots, l_i$; π_j^- , $j = l_e + 1, \dots, n_j$; and θ . We define the matrix $\tilde{W}$ to be used to in the next quadratic program, which is either (QP) or (QPC) as follows:

$$\tilde{W} = \begin{cases} \nabla_{xx}^2 \tilde{f}(x) + \sum_{i=l_i+1}^r \nu_i \nabla_{xx}^2 \tilde{g}_i(x) + \sum_{j=l_e+1}^m \pi_j \nabla_{xx}^2 \tilde{h}_j(x), \\ \quad \text{if (QP) was last solved,} \\ \nabla_{xx}^2 \tilde{f}(x) + \sum_{i=l_i+1}^r \nu_i \nabla_{xx}^2 \tilde{g}_i(x) + \sum_{j=l_e+1}^m (\pi_j^+ - \pi_j^-) \nabla_{xx}^2 \tilde{h}_j(x), \\ \quad \text{if (QPC) was last solved.} \end{cases}$$

We now define our algorithm in Table 4.1. The algorithm depends upon the parameters c_μ , $c_\gamma > 1$, c_0 .

The (QP) subproblem of this algorithm is the same subproblem as in the algorithm by [19], minus the trust-region constraint that is imposed for globalization. The (QPC) subproblem is the natural extension of (QP) when using the elastic mode. The elastic mode strategy used here is identical to the one used in SNOPT [24], except that we use here the L_∞ function.

TABLE 4.1
The elastic mode algorithm.

$x^0 = x$, $c_\infty = c_0$, $q = 0$.

NLP: Solve (QP).

If $\sum_{i=l_i+1}^m \nu_i^q + \sum_{j=l_e+1}^r |\pi_j^q| \leq c_\mu$ and (QP) is feasible
 $x^{q+1} = x^q + d^q$, $q = q + 1$, return to **NLP**.

Else

NLPC: solve (QPC).

$x^{q+1} = x^q + d^q$, $\zeta^{q+1} = \zeta^q + \delta_\zeta^q$, $q = q + 1$.

If $\sqrt{\|d^q\| + \|\delta_\zeta^q\|} \leq \zeta^q$,

$c_\infty = c_\infty c_\gamma$, $q = q + 1$ return to **NLPC**.

End If

End If

Note that once the subproblem (QPC) is solved the algorithm never solves the problem (QP) again. So we either solve (QP) till convergence or (QPC) till convergence.

We now analyze the effect of applying this algorithm to (MPCC). For this case, the (QP) subproblem becomes (QPX), whereas the (QPC) subproblem becomes (QPMC(c)).

THEOREM 4.5. Assume that assumptions [A1]–[A3], [A4a], and [A5] hold near a solution x^* of (MPCCS). Assume that the point x^{q_0} is sufficiently close to x^* and either

(i) the elastic mode is never invoked, and the algorithm uses at q_0 , for the purpose of constructing the matrix $\bar{W}$, an estimate of the Lagrange multiplier that is sufficiently close to a multiplier of (MPCCS);

(ii) the elastic mode is invoked at q_0 with $c_\infty \geq \nu_0$ and at all subsequent iterates, and the algorithm uses at q_0 , for the purpose of constructing the matrix $\bar{W}$, an estimate of the Lagrange multiplier that is sufficiently close to a multiplier of (MPCC(c)), and a ζ^{q_0} that is sufficiently close to 0.

Then x^q converges to x^* superlinearly in case (i) and (x^q, ζ^q) converges to $(x^*, 0)$ superlinearly in case (ii).

Proof. For case (i) to hold, we must have that the test involving the 1 norm of the nonlinear multipliers is always satisfied. Therefore, for all iteration indices q , we have that the following inequality holds:

$$\sum_{i=l_i+1}^{n_i} \nu_i^q + \sum_{j=l_e+1}^{n_e} |\pi_j^q| + \eta^q \leq c_\mu,$$

where the left-hand side is composed of multipliers of (QPX). Since we are sufficiently close to x^* , this must imply that $c_\mu \geq \nu_0$. In the latter case, we will have that a solution of (QPX) can be completed to a solution of (QPMC(c)) with $\zeta = 0$ and $\zeta + \delta_\zeta = 0$, and $c_\infty = c_\mu \geq \nu_0$. We therefore treat the case (i) as a special case of case (ii) to which we now confine our attention. We have the following cases.

• **Case 1** ($c_\infty = \nu_0$). In this case, following Lemma 4.2, (MPCC(c)) has a unique multiplier, and the result from [5] applies to give superlinear convergence of (x^q, ζ^q) to $(x^*, 0)$.

• Case 2 ($c_\infty > \nu_0$ and $\zeta^q + \delta\zeta^q = 0$ for all $q \geq q_0$). Then the solution obtained of (QPMC(c)) is also a solution of (QPX) which is feasible for all $x = x^q$. The result claimed follows from [18].

• Case 3 ($c_\infty > \nu_0$ and $\zeta^{q_0} + \delta\zeta^{q_0} > 0$). Then using Lemma 4.4, part (i) it follows that the point $x^{q_0+1} = x^{q_0} + d^{q_0}$, where d^{q_0} are the components of the solution of the quadratic program (QPMC(c)) corresponding to x at (x^q, ζ^q) , satisfies the complementarity constraints and that $x_{k1}^{q_0+1} = 0$, $k = 1, 2, \dots, n_c$. Now using Lemma 4.4, part (ii), for all $q \geq q_0 + 1$, we obtain that x^q satisfies $x_{k1}^q = 0$, $k = 1, 2, \dots, n_c$ for $q \geq q_0 + 1$, and $\zeta^q = 0$ for $q \geq q_0 + 2$. The superlinear convergence result again follows from [18].

The proof is complete. $\square$

We have the following observations:

1. We do not prove here the more desirable result that if x^q is sufficiently close to x^* we obtain superlinear convergence to x^* or of (x^q, ζ^q) to $(x^*, 0)$. The difficulty is that, unless we have an estimate for ν_0 , the subsequent iterates may find themselves far away from x^* once the elastic mode is entered. In effect, if the penalty parameter c_∞ is too small, the iterates may even be unbounded, even if the objective function is bounded on the feasible set. Such an adverse outcome can be prevented only by a global convergence result that will be the result of future research.

2. From Lemmas 4.2 and 4.3 we have that $\|\zeta^q\| = O(\|d^q\| + |d_\zeta^q|)$, so the c_∞ update rule from Table 4.1 will not be triggered, for x^q sufficiently close to x^* and for $c_\infty > \nu_0$. *Therefore, the update rule does not interfere with superlinear convergence.* On the other hand, it is also clear that, if $c_\infty < \nu_0$, then $\zeta^q > 0$ and the rule will eventually be triggered assuming that the iterates approach a stationary point. The test we use here is important because we do not spend an infinite amount of steps with an inappropriate c_∞ . The complete study of an appropriate rule should also involve global convergence issues, since one possible outcome of the penalty parameter adaptation rule is to obtain unbounded iterates.

3. The appropriate value of c_μ is ν_0 (or a slightly larger value), which of course cannot be determined unless we specifically use the MPCC structure. A general NLP approach cannot be guaranteed to succeed in determining the appropriate value for c_∞ through c_μ by looking at the multipliers of (QPMC(c)) alone, since the Lagrange multiplier set of (MPCCS) is unbounded, whereas such rules are based on the assumption of boundedness of the Lagrange multiplier set at the solution x^* [2]. Whether c_μ can be adaptively defined is a matter for future research. For NLP implementations, a user-defined value for c_μ is considered to be an acceptable approach [24].

4. Note that the only meaningful difference between the proofs outlined here and the ones in [18], once Lemma 4.4 has been established, is the one involved in the case where $c_\infty = \nu_0$, for which the well-known result from [5] has been invoked. Our contribution to this class of superlinear convergent results has been essentially to show that, for the case $c_\infty > \nu_0$, the elastic mode approach, as presented in Table 4.1, will force the algorithm to choose points that are either always feasible or always satisfy the complementarity constraints and in so doing we do not have to a priori assume that this holds, as it was done in [18].

A similar result holds for the case where we use only assumption [A4], rather than its stronger version [A4a], which is used in [18]. The only weakening we have to do of the result is that we have to assume convergence of the sequences from the outset, rather than assuming only that the sequences find themselves in a sufficiently small neighborhood of the respective limit points. Following techniques from [45, 44],

this assumption could probably be dropped at the cost of substantially complicating the analysis.

THEOREM 4.6. *Assume that assumptions [A1]–[A4] hold near a solution x^* of (MPCCS). Assume that either*

(i) *the elastic mode is never invoked, and the algorithm uses at q_0 , for the purpose of constructing the matrix $\tilde{W}$, an estimate of the Lagrange multiplier that is sufficiently close to a multiplier of (MPCCS) and that $x^q \rightarrow x^*$;*

(ii) *the elastic mode is invoked at q_0 with $c_\infty \geq \nu_0$ and at all subsequent iterates, and the algorithm uses at q_0 , for the purpose of constructing the matrix $\tilde{W}$, an estimate of the Lagrange multiplier that is sufficiently close to a multiplier of (MPCC(c)), and a ζ^{q_0} that is sufficiently close to 0, and that $(x^q, \zeta^q) \rightarrow (x^*, 0)$.*

Then x^q converges to x^* superlinearly in case (i) and (x^q, ζ^q) converges to $(x^*, 0)$ superlinearly in case (ii).

Proof. Much as in the proof of Theorem 4.5, the proof of the case (i) can be reduced to the proof of the case (ii). Also, the conclusion of the case (ii), where $c_\infty = \nu_0$, follows by the same argument as in Theorem 4.5. So we concentrate on proving (ii) for the case $c_\infty - \nu_0 = \gamma_0 > 0$.

We will complete the proof by considering the following cases.

Case 1. For all iterations, we have that $\zeta^q + d_\zeta^q = 0$, but the relation $x_{k1}^q = 0$, $k = 1, 2, \dots, n_c$, does not hold at any iteration q .

Then on one hand, from assumption [A4] and Lemmas 4.2 and 4.3 we must have that $\hat{\mu}_{k1}^{q+1} > 0$ for $k = 1, 2, \dots, n_d$. Assume that, on the other hand, we have that $\hat{\mu}_{k1}^{q+1} > 0$ for some iteration index q and all k , $n_d + 1 \leq k \leq n_c$. Then, from the optimality conditions of (QPMC(c)) we must have that $x_{k1}^{q+1} = 0$, $k = 1, 2, \dots, n_c$. This is a contradiction with the assumption at the beginning of this case.

So there must be a k such that $\hat{\mu}_{k1}^{q+1} = 0$. Without loss of generality for the following argument, we assume that

$$(4.12) \quad \hat{\mu}_{n_c,1}^{q+1} = 0.$$

Now take j , $n_i + 1 \leq j \leq n_i$. Since we have that $\zeta^q + d_\zeta^q = 0$, it follows that the double inequality constraint $-\zeta - d_\zeta \leq h_j + \nabla_x h_j(x)d \leq \zeta + d_\zeta$ of (QPMC(c)) will have both sides active, with corresponding Lagrange multipliers $\hat{\pi}_j^{-q+1}$ and $\hat{\pi}_j^{+q+1}$. An immediate consequence of assumption [A2] is that $\hat{\pi}_j^{-q+1} - \min\{\hat{\pi}_j^{-q+1}, \hat{\pi}_j^{+q+1}\}$ and $\hat{\pi}_j^{+q+1} - \min\{\hat{\pi}_j^{-q+1}, \hat{\pi}_j^{+q+1}\}$ are also Lagrange multipliers of the same respective constraints. Note that the expression of $\widehat{W}$ at the next iteration $q+1$ does not change when using the new multipliers. We can therefore assume without loss of generality that

$$(4.13) \quad \hat{\pi}_j^{-q+1} \hat{\pi}_j^{+q+1} = 0.$$

Let now $\widehat{\Pi}^*$ be the Lagrange multiplier of (MPCC(c)) that is closest to

$$\widehat{\Pi}^{q+1} = (\hat{\nu}^{q+1}, \hat{\pi}^{q+1}, \hat{\pi}^{-q+1}, \hat{\pi}^{+q+1}, \hat{\mu}_1^{+q+1}, \hat{\mu}_2^{+q+1}, \hat{\eta}^{q+1}, \hat{\theta}^{q+1}).$$

Define now f_j^* , $j = n_d + 1, \dots, n_c$, and a^* by subtracting the respective components of the fundamental multiplier from the components of $\widehat{\Pi}^*$, as in (4.5).

It then follows from Lemmas 4.2 and 4.3 that

$$|\hat{\mu}_{n_c,1}^{q+1} - \hat{\mu}_{n_c,1}^*| \leq c_2(\|x^q - x^*\| + \zeta^q).$$

From (4.12) and (4.5) it then follows that

$$0 \leq -a^* x_{n_c 2}^* \leq \hat{\mu}_{n_c, 1}^* \leq c_2(\|x^q - x^*\| + \zeta^q),$$

which in turn implies that

$$0 \leq a^* \leq \frac{c_2}{-x_{n_c 2}^*}(\|x^q - x^*\| + \zeta^q).$$

By a similar rationale, we obtain that (4.13) implies that, for any j such that $n_{l_e} + 1 \leq j \leq n_c$, we have that

$$f_j^* \leq c_2(\|x^q - x^*\| + \zeta^q).$$

Now using (4.5) we get that

$$\hat{\Pi}^{q+1} \rightarrow (\Pi^*, \gamma_0),$$

where Π^* is the fundamental multiplier whose components are defined in (FMC).

Since the sequence of Lagrange multipliers is convergent, it follows from Lemma 4.2 by an argument similar to the one in [5], that is also discussed in [45, 44], that the sequence (x^q, ζ^q) is superlinearly convergent to (x^*, ζ^*) . The proof of this case is complete.

Case 2. We have that $\zeta^q + d_\zeta^q = 0$, as well as $x_{k1}^q = 0$, for $1 \leq k \leq n_c$ at iteration $q = q^0$. Then, from Lemma 4.4 we have that $\zeta^q + d_\zeta^q = 0$, as well as $x_{k1}^q = 0$, for $1 \leq k \leq n_c$ for all iterations $q \geq q_0 + 1$. It then follows that the solution of (QMPC(c)) in the x component coincides with the solution of (QPR), since on the feasible set of (QPR) we have that $d^T W d = d^T \tilde{W} d$ with the notations of Lemma 4.4. Subsequently, the steps produced by the algorithm in the x component are identical to the steps produced by the algorithm applied to the nonlinear program

$$\begin{aligned} & \text{minimize} && f(x) \\ & \text{subject to} && g_i(x) \leq 0, \quad i = 1, 2, \dots, n_i, \\ & && h_j(x) = 0, \quad j = 1, 2, \dots, n_j, \\ & && x_{k1} = 0, \quad k = 1, 2, \dots, n_c, \\ & && x_{k2} \leq 0, \quad k = 1, 2, \dots, n_c. \end{aligned}$$

Since, from assumptions [A2] and [A3] this program satisfies LICQ and SOSC at x^* , the fact that x^q converges superlinearly to x^* follows from [12, 38]. The result follows since $\zeta^q = 0$.

Case 3. We have that $\zeta^q + d_\zeta^q > 0$, at iteration $q = q_0$. It then follows from Lemma 4.4 that $x_{k,1}^{q+1} = 0$, for $1 \leq k \leq n_c$. Using Lemma 4.4 again, we obtain that $\zeta^q + d_\zeta^q = 0$, as well as $x_{k1}^q = 0$, for $1 \leq k \leq n_c$ for all iterations $q \geq q_0 + 2$. The conclusion follows from the preceding case. The proof is then complete. $\square$

Acknowledgments. Thanks to Stefan Scholtes for pointing out that the proof of Corollary 2.3 also stems from [40, Theorem 4]. Thanks to Alexander Shapiro for pointing out [5]. I am particularly grateful to my colleague Sven Leyffer for providing the AMPL examples and guidance for the numerical results and to Steve Wright for pointing out the result [45, Theorem A1]. Finally, I acknowledge the detailed comments from the referees and the editor, Danny Ralph, that have prompted the development of section 4, and have markedly improved the quality of the results.

REFERENCES

- [1] M. ANITESCU, *Degenerate nonlinear programming with a quadratic growth condition*, SIAM J. Optim., 10 (2000), pp. 1116–1135.
- [2] M. ANITESCU, *On the rate of convergence of sequential quadratic programming with nondifferentiable exact penalty function in the presence of constraint degeneracy*, Math. Program., 92 (2002), pp. 359–386.
- [3] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.
- [4] D. P. BERTSEKAS, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1995.
- [5] F. BONNANS, *Local analysis of Newton-type methods for variational inequalities and nonlinear programming*, Appl. Math. Optim., 29 (1994), pp. 161–186.
- [6] F. BONNANS AND A. IOFFE, *Second-order sufficiency and quadratic growth for nonisolated minima*, Math. Oper. Res., 20 (1995), pp. 801–817.
- [7] F. BONNANS AND A. SHAPIRO, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, New York, 2000.
- [8] J. V. BURKE, *A sequential quadratic programming method for potentially infeasible mathematical programs*, J. Math. Anal. Appl., 139 (1989), pp. 319–351.
- [9] J. V. BURKE, *A robust trust region method for constraint nonlinear programming problems*, SIAM J. Optim., 2 (1992), pp. 325–347.
- [10] J. V. BURKE AND S.-P. HAN, *A robust sequential quadratic programming method*, Math. Program., 43 (1989), pp. 277–303.
- [11] R. J. CARON, H. J. GREENBERG, AND A. G. HOLDER, *Analytic centers and repelling inequalities*, European J. Oper. Res., 143 (2002), pp. 268–290.
- [12] A. R. CONN, N. I. M. GOULD, AND PH. L. TOINT, *Trust-Region Methods*, MPS-SIAM Ser. Optim. 1, SIAM, Philadelphia, 2000.
- [13] F. FACCHINEI, *A robust recursive quadratic programming algorithm with global and super linear convergence properties*, J. Optim. Theory Appl., 92 (1997), pp. 543–579.
- [14] A. V. FIACCO, *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, Academic Press, New York, 1983.
- [15] R. FLETCHER, *Practical Methods of Optimization*, John Wiley & Sons, Chichester, 1987.
- [16] R. FLETCHER AND S. LEYFFER, *Solving mathematical programs with complementarity constraints as nonlinear programs*, Optim. Methods Softw., 19 (2004), pp. 15–40.
- [17] R. FLETCHER AND S. LEYFFER, *Nonlinear programming without a penalty function*, Math. Program., 91 (2002), pp. 239–269.
- [18] R. FLETCHER, S. LEYFFER, S. SHOLTES, AND D. RALPH, *Local Convergence of SQP Methods for Mathematical Programs with Equilibrium Constraints*, Tech. report NA 210, University of Dundee, Dundee, UK, 2002.
- [19] R. FLETCHER, S. LEYFFER, AND P. L. TOINT, *On the Global Convergence of a Filter-SQP Type Algorithm*, Numerical Analysis Report NA/197, Department of Mathematics, University of Dundee, Dundee, UK, 1999.
- [20] R. M. FREUND, R. ROUNDY, AND M. J. TODD, *Identifying the Set of Always-Active Constraints in a System of Linear Inequalities by a Single Linear Program*, Number 1674-85, Massachusetts Institute of Technology (MIT) Sloan School of Management, Cambridge, MA, <http://ideas/repec.org/p/mit/sloanp/2111.html> (29 April 2003).
- [21] M. FUKUSHIMA, Z.-Q. LUO, AND J.-S. PANG, *A globally convergent sequential quadratic programming algorithm for mathematical programs with linear complementarity constraints*, Comput. Optim. Appl., 10 (1998), pp. 5–34.
- [22] M. FUKUSHIMA AND P. TSENG, *An implementable active-set algorithm for computing a B-stationary point of a mathematical program with linear complementarity constraints*, SIAM J. Optim., 12 (2002), pp. 724–739.
- [23] J. GAUVIN, *A necessary and sufficient regularity condition to have bounded multipliers in nonconvex programming*, Math. Program., 12 (1977), pp. 136–138.
- [24] P. E. GILL, W. MURRAY, AND M. A. SAUNDERS, *User's guide for SNOPT 5.3: A fortran package for large-scale nonlinear programming*, Report NA 97-5, Department of Mathematics, University of California, San Diego, San Diego, CA, 1997.
- [25] X. HU AND D. RALPH, *Convergence of a penalty method for mathematical programming with complementarity constraints*, J. Optim. Theory Appl., 123 (2004), pp. 365–390.
- [26] A. IOFFE, *Necessary and sufficient conditions for a local minimum. III. Second order conditions and augmented duality*, SIAM J. Control Optim., 17 (1979), pp. 266–288.
- [27] A. IOFFE, *On sensitivity analysis of nonlinear programs in Banach spaces: The approach via composite unconstrained optimization*, SIAM J. Optim., 4 (1994), pp. 1–43.

- [28] H. JIANG AND D. RALPH, *Smooth SQP methods for mathematical programs with nonlinear complementarity constraints*, SIAM J. Optim., 10 (2000), pp. 779–808.
- [29] S. LUCIDI, *New results on a continuously differentiable exact penalty function*, SIAM J. Optim., 2 (1992), pp. 558–574.
- [30] Z.-Q. LUO, J.-S. PANG, AND D. RALPH, *Mathematical Programs with Complementarity Constraints*, Cambridge University Press, Cambridge, UK, 1996.
- [31] Z.-Q. LUO, J.-S. PANG, D. RALPH, AND S.-Q. WU, *Exact penalization and stationarity conditions of mathematical programs with complementarity constraints*, Math. Program., 34 (1986), pp. 142–162.
- [32] O. L. MANGASARIAN, *Nonlinear Programming*, McGraw-Hill, New York, 1969.
- [33] O. L. MANGASARIAN AND S. FROMOVITZ, *The Fritz John necessary optimality conditions in the presence of equality constraints*, J. Math. Anal. Appl., 17 (1967), pp. 34–47.
- [34] B. MURTAGH AND M. A. SAUNDERS, *Minos 5.0 User's Guide*, Tech. report SOL 83-20, Stanford University, Stanford, CA, 1983.
- [35] *The NEOS guide*, <http://www.mcs.anl.gov/otc/Guide>.
- [36] J. OUTRATA, M. KOCVARA, AND J. ZOWE, *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints: Theory, Applications and Numerical Results*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.
- [37] E. POLAK, *Optimization*, Springer-Verlag, New York, 1997.
- [38] S. M. ROBINSON, *Perturbed Kuhn-Tucker points and rates of convergence for a class of nonlinear-programming algorithms*, Math. Program., 7 (1974), pp. 1–16.
- [39] S. M. ROBINSON, *Generalized equations and their solutions*, II. *Applications to nonlinear programming*, Math. Prog. Stud., 19 (1982), pp. 200–221.
- [40] H. SCHEEL AND S. SCHOLTES, *Mathematical programs with equilibrium constraints: Stationarity, optimality, and sensitivity*, Math. Oper. Res., 25 (2000), pp. 1–22.
- [41] S. SCHOLTES, *Convergence properties of a regularization scheme for mathematical programs with complementarity constraints*, SIAM J. Optim., 11 (2001), pp. 918–936.
- [42] A. SHAPIRO, *Sensitivity analysis of nonlinear programs and differentiability properties of metric projections*, SIAM J. Control Optim., 26 (1988), pp. 628–645.
- [43] J. TELGEN, *Minimal representation of convex polyhedral sets*, J. Optim. Theory Appl., 38 (1982), pp. 1–24.
- [44] S. J. WRIGHT, *Superlinear convergence of a stabilized SQP method to a degenerate solution*, Comput. Optim. Appl., 11 (1998), pp. 253–275.
- [45] S. J. WRIGHT, *Modifying SQP for degenerate problems*, SIAM J. Optim., 13 (2002), pp. 470–497.

TAIL DECAY AND MOMENT ESTIMATES OF A CONDITION NUMBER FOR RANDOM LINEAR CONIC SYSTEMS*

DENNIS CHEUNG[†], FELIPE CUCKER[†], AND RAPHAEL HAUSER[‡]

Abstract. In this paper we study the distribution of $\mathcal{C}(A)$ and $\log \mathcal{C}(A)$, where $\mathcal{C}(A)$ is a condition number for the linear conic system $Ax \leq 0$, $x \neq 0$, with $A \in \mathbb{R}^{n \times m}$. For Gaussian matrices A we develop both upper and lower bounds on the decay rates of the distribution tails of $\mathcal{C}(A)$, showing that $\mathbf{P}[\mathcal{C}(A) \geq t] \sim c/t$ for large t , where c is a factor that depends only on the problem dimensions (m, n) . Using these bounds, we derive moment estimates for $\mathcal{C}(A)$ and $\log \mathcal{C}(A)$ and prove various limit theorems for the cases where m and/or n are large. Combined with condition number based complexity analyses, our results yield tail information on the distribution of running times for interior-point or relaxation methods designed to solve the feasibility problem $Ax \leq 0$, $x \neq 0$.

Key words. condition numbers, random matrices, linear programming, probabilistic analysis

AMS subject classifications. Primary, 90C31, 15A52; Secondary, 90C05, 90C60, 62H10

DOI. 10.1137/S105262340343470X

1. Introduction. Let $A \in \mathbb{R}^{n \times m}$ be a given matrix, and consider the systems

$$(1.1) \quad Ax \leq 0, \quad x \neq 0,$$

$$(1.2) \quad A^T y = 0, \quad y \geq 0, \quad y \neq 0.$$

It is well known that if A has full column rank, then one of these systems has a strict solution (one for which the satisfied inequality is strict in all coordinates) iff the other has no solution at all.

The following feasibility problem is a standard subproblem in linear programming:

(FP) Given a $n \times m$ real matrix A with full column rank, decide which of (1.1) or (1.2) is strictly feasible and return a strict solution for it.

Iterative algorithms that solve this problem generally have a cost that depends on some measure of conditioning of the matrix A besides the natural dependence on n and m . For instance, a finite-precision interior-point algorithm solving the problem above is analyzed in [17] to show a complexity of

$$(1.3) \quad \mathcal{O}\left((m+n)^{3.5}(\log(m+n) + \log C(A))^3\right).$$

Here $C(A)$ may be either the condition number introduced by Renegar [33, 34, 35], which we denote by $C_R(A)$, or a generalization of Goffin's "inner measure" [24, 25] introduced in [13], which we denote by $\mathcal{C}(A)$ in what follows.

*Received by the editors September 17, 2003; accepted for publication (in revised form) September 13, 2004; published electronically August 9, 2005.

<http://www.siam.org/journals/siopt/15-4/43470.html>

[†]Department of Mathematics, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong (50003110@student.cityu.edu.hk, macucker@math.cityu.edu.hk). The second author was substantially funded by a grant from the Research Grants Council of the Hong Kong SAR (project number CityU 1085/02P).

[‡]Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford, OX1 3QD, UK (hauser@comlab.ox.ac.uk). This author was supported by a grant from the Nuffield Foundation under the "Newly Appointed Lecturers" grant scheme (project number NAL/00720/G) and through grant GR/S34472 from the Engineering and Physical Sciences Research Council of the UK.

Complexity bounds that depend logarithmically on a condition number are typical for interior-point methods and variants of the ellipsoid method. Other algorithms have a complexity that is polynomial in $\mathcal{C}(A)$: the Agmon–Motzkin–Schönberg (AMS) relaxation method [2, 30], the cyclic projection method, the perceptron algorithm, and certain types of subgradient algorithms are of this type. The complexity of these methods is typically proportional to $\mathcal{C}(A)^2$. For example, for strictly feasible systems the perceptron algorithm is guaranteed to find a solution to $Ax < 0$ in

$$(1.4) \quad \mathcal{O}(\mathcal{C}(A)^2)$$

iterations; see Appendix B for further remarks. Although less interesting from a complexity perspective, such algorithms have appealing aspects in applications where m and n are both very large, for example in tumor radiation therapy planning; see [12, Chapter 11]. AMS relaxation is also the historic context in which the condition number $\mathcal{C}(A)$ was first studied, albeit only in the case of feasible systems; see [24, 25].

It is also worthwhile mentioning that there exist close links between $\mathcal{C}(A)$ and the notion of *margin*, which plays a key role in the learning theory literature. Furthermore, $\mathcal{C}(A)$ has applications in the backward error analysis and in estimating the stability of linear feasibility problems. For all these reasons, studying the moments of both $\log \mathcal{C}(A)$ and $\mathcal{C}(A)$ will be interesting in the probabilistic setting outlined below.

Unlike m and n , the condition number of A is not immediate to determine from the data A and seems to require a computation which is not easier than solving the feasibility problem instance described by A (see [32] for a discussion). In addition, there are no bounds on its magnitude as a function of m and n . It may actually be infinite. Thus, bounds such as (1.3) and (1.4) tell us little about the running time we can expect for a given input A .

One way to cope with this situation is to assume a probability measure on the input space of $n \times m$ matrices A and to derive some information about the distribution of the random variable $\mathcal{C}(A)$. The standard choice of an input distribution is to assume that A has independently and identically distributed (i.i.d.) standard normal components. We refer to such matrices as Gaussian in what follows. The distribution of the condition number for linear systems of equations under Gaussian input has been fully characterized [21], but since condition numbers for systems of linear inequalities are less explicit, their distribution has been studied almost exclusively through $\mathbf{E}[\log \mathcal{C}(A)]$. Previously, this question was studied in [43, 14] and [19]. Moreover, the so-called smoothed analysis of condition numbers [20] contains the study of $C_R(A)$ under Gaussian input as a special case when the distribution of A is centered at the origin.

In [14] it was established that if A is a Gaussian $n \times m$ matrix, then

$$(1.5) \quad \mathbf{E}[\log \mathcal{C}(A)] = \begin{cases} \mathcal{O}(\min\{m \log n, n\}) & \text{if } n > m, \\ \mathcal{O}(\log n) & \text{otherwise,} \end{cases}$$

$$\mathbf{E}[\log C_R(A)] \leq \mathbf{E}[\log \mathcal{C}(A)] + (5 \log n)/2 + (\log m)/2 + 2 \log 2.$$

In many practical occurrences of the feasibility problem (FP) one deals with matrices for which n is some orders of magnitude larger than m . The case $n \gg m$ is actually the case of interest among researchers in linear programming. In [19] the estimate (1.5) was refined to prove that, when n is moderately larger than m , one has

$$\mathbf{E}[\log \mathcal{C}(A)] \leq \max\{\log m, \log \log n\} + \mathcal{O}(1).$$

See also below for bounds obtained in [20].

Our paper contributes to this discussion in several ways:

(i) Our main results establish that the tail probabilities $\mathbf{P}[\mathcal{C}(A) \geq t]$ decay exactly at the rate t^{-1} : there exists constants $c_l(m, n), c_u(m, n)$ whose size depends only on the dimensions of A such that

$$(1.6) \quad c_l(m, n)t^{-1} \leq \mathbf{P}[\mathcal{C}(A) \geq t] \leq c_u(m, n)t^{-1};$$

see Theorem 6.4 for the upper bound and Theorem 7.3 for the lower bound. An upper bound of this type was already established in [19], but no lower bound was previously known. Equation (1.6) proves that for any fixed dimension the correct tail decay of $\mathcal{C}(A)$ is given by the regime t^{-1} . For comparison, the decay regime for $\mathbf{P}[C_R(A) \geq t]$ established in [20] is slower than $t^{-1} \times \log^2 t$.

(ii) Our inequalities (1.6) make it easy to give explicit estimates on any moment of $\mathcal{C}(A)$ or $\log \mathcal{C}(A)$; see Corollaries 9.1, 9.3, and 9.4.

(iii) When m and/or n are large, the estimates mentioned in (ii) give some indication on the moment asymptotics. Corollary 11.1 improves the asymptotic bounds from [19]: for $n \geq m$ and m large enough,

$$\mathbf{E}[\log \mathcal{C}(A)] \leq m \log 2,$$

and for $n \geq 5m$ and m large enough,

$$\mathbf{E}[\log \mathcal{C}(A)] < m^\gamma$$

for any $\gamma > 0$; that is, $\mathbf{E}[\log \mathcal{C}(A)]$ grows at a subalgebraic rate. In [20] a bound on $\mathbf{E}[\log C_R(A)]$ is proven that implies that $\mathbf{E}[\log \mathcal{C}(A)] \leq \mathcal{O}(\log n)$. This bound is stronger than ours, unless n is exponentially large in m .

(iv) Although the results mentioned in (iii) give some indication of the behavior of the distribution of $\mathcal{C}(A)$ when the dimensions of A are large, they usually fail to reflect the correct qualitative asymptotics. We recover the correct limiting behavior in a string of results based on more direct methods than the estimates (1.6). In particular, we show that for all fixed $m \geq 3$,

$$\begin{aligned} \mathcal{C}(A) &\xrightarrow[n \rightarrow \infty]{P} 1, \\ \lim_{n \rightarrow \infty} \mathbf{E}[(\log \mathcal{C}(A))^\gamma] &= 0 \quad \forall \gamma > 0, \\ \lim_{n \rightarrow \infty} \mathbf{E}[(\mathcal{C}(A))^\gamma] &= 1 \quad \forall \gamma \in [0, 1); \end{aligned}$$

see Corollaries 10.2 and 10.3.

(v) Previous analyses of $\log \mathcal{C}(A)$ for Gaussian A relied typically on linear algebra and on the fact that linear combinations of Gaussian variables are Gaussians. Our approach is completely different and relies on spherical geometry as the basic technique. This can be of interest even in contexts where the specific bounds we derive are not. It is interesting to compare our results with those obtained in [20] because these are based on the yet completely different geometric idea that the integral of a Gaussian density over the boundary of a convex set in $\mathbb{R}^d$ is bounded by $4d^{1/4}$; see [4]. The paper [20] and our own work were written independently of one another and roughly during the same time period.

2. The complexity theoretic motivation. The interest in the distribution of $\log \mathcal{C}(A)$ stems to a large extent from the conjecture that there exist so-called *strongly polynomial* algorithms for linear programming (LP). It is well known that under the classical Turing machine model, interior-point and ellipsoid methods can solve LP in polynomial time when the input data are rational. By contrast, an algorithm is called *strongly polynomial* if the number of operations on reals needed until termination is bounded by a polynomial in the dimension of the input data, which are assumed to consist of real numbers. This complexity model was formalized in the theory of real Turing machines [7]. No strongly polynomial algorithm for LP is known at present: the known versions of the simplex algorithm have exponential upper and lower complexity bounds in the real Turing machine model, while interior-point and ellipsoid algorithms have no a priori termination bound.

The situation is however not hopeless, for the latter classes of algorithms can be bounded by a polynomial in the problem dimension and the logarithm of a condition number; see, e.g., (1.3). Earlier relevant papers on this subject and on other applications of LP condition numbers include [33, 34, 35, 23, 22, 47, 17, 45, 46]. The idea of condition-based complexity analysis is also known in other areas of numerical analysis: in numerical linear algebra it occurs in the context of the conjugate gradient algorithm and other iterative methods (cf. [31, 44]); in polynomial equation solving it occurs in the analysis of homotopy methods [37] or of grid-based methods [18, 16]. A recent survey for conditioning in LP is [15].

The question of whether LP is strongly polynomial time solvable is regarded as an important open problem in complexity theory and appears in Smale's list of 18 mathematical problems for the twenty-first century [39]. In order to get around the difficult issue of strong polynomiality, it is often argued that the study of average running times of an algorithm over a reasonable input distribution can be more relevant than the study of worst-case scenarios. Such an analysis may reveal that an algorithm is strongly polynomial "on average." The earliest method to be studied in this framework was the simplex method [8, 9, 10, 38, 1, 28, 40] and [11]. In these studies the input data are usually assumed to be Gaussian. More recently, attention has shifted to the average case analysis of interior-point methods [3, 41, 42, 29, 26] and [27]. While all these papers followed an analysis pertaining to particular algorithms, it is also possible to derive similar results by analyzing the expected value of condition numbers under random input data. Combined with a condition-based complexity analysis this approach yields average case running time bounds for entire families of algorithms. This motivates the investigation of the distribution of $\mathcal{C}(A)$.

3. Basic notions and notation. If there exists $x \in \mathbb{R}^m$, $x \neq 0$, such that $Ax \leq 0$ we say that A is *feasible*. Otherwise, we say that A is *infeasible*. Also, if there exists a vector x such that $Ax < 0$ componentwise, then we say that A is *strictly feasible*. If A is feasible but not strictly feasible, then we say that A is *ill-posed*.

In what follows we consider $\arccos(t)$ as a function from $[-1, 1]$ into $[0, \pi]$. In this region, both $\cos$ and $\arccos$ are decreasing functions. Let a_i be the i th row of A . Denote by $\theta_i(A, x)$ the angle between a_i and x , that is, $\arccos(a_i \cdot x / \|a_i\| \|x\|)$. Let $\theta(A, x) = \min_{1 \leq i \leq n} \theta_i(A, x)$ and $\bar{x}$ be any vector in $\mathbb{R}^m \setminus \{0\}$, such that $\theta(A) = \theta(A, \bar{x}) = \sup_{x \in \mathbb{R}^m} \theta(A, x)$. The condition number $\mathcal{C}(A)$ is defined as

$$\mathcal{C}(A) = |\cos(\theta(A))|^{-1}.$$

In the case where the system $Ax \leq 0$ is feasible, $\mathcal{C}(A)$ is the same as Goffin's condition number ν [24, 25], which was developed to analyze step-length rules that

guarantee that the relaxation method applied to strictly feasible systems of linear inequalities converges in finite time. $\mathcal{C}(A)$ is also closely related to Agmon's condition number λ [2], with which it coincides in some cases, but again λ is only defined for feasible systems. Thus, $\mathcal{C}(A)$ is more general because it is defined both for feasible and infeasible systems.

It is not difficult to see that A is strictly feasible iff $\theta(A) > \pi/2$, ill-posed iff $\theta(A) = \pi/2$, and infeasible iff $\theta(A) < \pi/2$. Using a compactness argument it is also easy to show that a vector $\bar{x}$ such as that used in the definition of $\mathcal{C}(A)$ exists. Note that since $\mathcal{C}(A)$ is defined purely in terms of angles between vectors, $\mathcal{C}$ is invariant under positive rescaling of the rows of A . Hence, we may assume without loss of generality that all rows of A have been rescaled to unit length. Our analysis then reduces to geometry on the unit sphere.

Let S^{m-1} denote the unit sphere in $\mathbb{R}^m$. For $p \in S^{m-1}$ and $\rho \in [0, \pi]$ we denote by $\text{cap}(p, \rho)$ the spherical cap with its center in p and angular radius ρ , that is, $\text{cap}(p, \rho) = \{x \in S^{m-1} : x \cdot p \geq \cos \rho\}$. By $\partial \text{cap}(p, \rho)$ and $\text{int}(\text{cap}(p, \rho))$ we denote the boundary and interior of $\text{cap}(p, \rho)$, respectively, in the standard topology on S^{m-1} , that is, $\partial \text{cap}(p, \rho) = \{x \in S^{m-1} : x \cdot p = \cos \rho\}$, $\text{int}(\text{cap}(p, \rho)) = \{x \in S^{m-1} : x \cdot p > \cos \rho\}$.

PROPOSITION 3.1. $0 < \theta(A) \leq \rho \leq \pi \Leftrightarrow \bigcup_{i=1}^n \text{cap}(a_i, \rho) = S^{m-1}$.

Proof. $\theta(A) \leq \rho \Leftrightarrow \max_{x \in \mathbb{R}^m \setminus \{0\}} \theta(A, x) \leq \rho \Leftrightarrow \forall x \in S^{m-1}, \min_{1 \leq i \leq n} \theta_i(A, x) \leq \rho \Leftrightarrow \forall x \in S^{m-1}, \exists i \in \{1, 2, \dots, n\}, x \in \text{cap}(a_i, \rho) \Leftrightarrow \bigcup_{i=1}^n \text{cap}(a_i, \rho) = S^{m-1}$. $\square$

4. Characterizing extremal spherical caps. In this section A is a real $n \times m$ matrix with unit row vectors a_i . A *largest excluding spherical cap* (LEC) for A is a closed spherical cap $\text{cap}(p^*, \rho^*)$, where $\rho^* = \rho(p^*)$, $\rho(p) = \min_i \arccos(p \cdot a_i)$, and p^* is a maximizer for the optimization problem

$$(4.1) \quad \max_{p \in S^{m-1}} \rho(p).$$

A *smallest including spherical cap* (SIC) for A is the closure of the complement of a LEC for A . In other words, a LEC is maximal amongst all closed spherical caps on S^{m-1} that contain none of the row vectors of A in their interior, while a SIC is minimal amongst the closed spherical caps on S^{m-1} that contain all rows of A . The following result explains why the notions of LEC and SIC are important in the analysis of $\mathcal{C}(A)$.

PROPOSITION 4.1. *Let $\text{cap}(p, \rho)$ be a LEC for A . Then $x^* = p$ maximizes the function $\theta(A, x)$, and $\theta(A) = \theta(A, p) = \rho$, that is, $\mathcal{C}(A) = 1/|\cos \rho|$.*

Proof. Let $\bar{x} \in S^{m-1}$ be a maximizer of $\theta(A, x)$. Then $\theta(A, \bar{x}) = \theta(A)$, and $\text{int}(\text{cap}(\bar{x}, \theta(A)))$ contains none of the rows of A . This shows that $\theta(A) \leq \rho$. On the other hand, since $a_i \in \text{cap}(-p, \pi - \rho)$ for all i , we have $\rho \leq \min_{1 \leq i \leq n} \arccos(p \cdot a_i) = \theta(A, p) \leq \theta(A)$. $\square$

It is worth investigating the properties of LECs and SICs a bit further. A simple compactness argument shows that LECs and SICs exist for any A . If A is infeasible, then these spherical caps are not unique in general. If A is strictly feasible, however, it is not difficult to prove that there exists a unique LEC and SIC for A . Let $\text{cap}(p, \rho)$ be a LEC for A . We say that $S = \{i : a_i \in \partial \text{cap}(p, \rho)\}$ is the *blocking set* of $\text{cap}(p, \rho)$. The blocking set corresponds to the vectors that locally keep the LEC from having a larger radius. In fact, S is the active set in the following equivalent reformulation of (4.1):

$$\max \{ \rho \in [0, \pi] : \exists p \in S^{m-1} \text{ subject to } p \cdot a_i \leq \cos \rho \forall i \in \mathbb{N}_n \}.$$

By straightforward extension we also speak of the blocking set of a SIC.

If A is strictly feasible, then the blocking set of the (unique) LEC can have any cardinality $\geq \min(2, n)$, as can easily be illustrated by the example of three points on S^2 that lie on a single great circle and within a sufficiently small angle of one another. The situation is rather different if A is infeasible (in which case $n \geq m$ always holds). In this case all blocking sets have cardinality at least m . In fact, let $\text{cap}(p, \rho)$ be a LEC of an infeasible A , that is, $\theta(A) = \rho < \pi/2$, let S be the blocking set of $\text{cap}(p, \rho)$, and suppose that $|S| < m$. Then there exists a vector $u \in S^{m-1} \cap \text{Span}\{a_i : i \in S\}^\perp$ such that $u \cdot p \geq 0$. Let $p_\delta = p + \delta u$. For $\delta > 0$ we have $\|p_\delta\|^2 = 1 + \delta^2 + 2\delta u \cdot p > 1$. Let $\hat{p}_\delta = p_\delta / \|p_\delta\|$ and $\rho_\delta = \arccos(\cos \rho / \|p_\delta\|)$. Then $\rho_\delta > \rho$. We have $a_i \notin \text{cap}(\hat{p}_\delta, \rho_\delta)$ for all $i \notin S$ and $\delta > 0$ sufficiently small, since $a_i \notin \text{cap}(p, \rho)$ and because $\text{cap}(\hat{p}_\delta, \rho_\delta)$ varies continuously as a function of δ . Moreover, for $i \in S$ we have $p_\delta \cdot a_i = p \cdot a_i + \delta u \cdot a_i = p \cdot a_i = \cos \rho$. Therefore $\hat{p}_\delta \cdot a_i = \cos \rho / \|p_\delta\| = \cos \rho_\delta$, and this shows that $a_i \in \partial \text{cap}(p_\delta, \rho_\delta)$ for all $i \in S$. We conclude that $\text{int}(\text{cap}(\hat{p}_\delta, \rho_\delta))$ contains none of the a_i , and since $\rho_\delta > \rho = \theta(A)$, this is a contradiction. Therefore, our assumption was wrong, and $|S| \geq m$. Let us summarize what we have found so far.

PROPOSITION 4.2. *Let $A \in \mathbb{R}^{n \times m}$ have unit rows. Then,*

- (i) *if A is strictly feasible, then there exists a unique LEC and SIC for A , but the cardinality of the blocking set is arbitrary;*
- (ii) *if A is infeasible, then there exist LECs and SICs for A , which are generally not unique, and their blocking sets always have cardinality at least m .* $\square$

Let us now further explore the link between blocking sets and extremal spherical caps in the case where $n \geq m$. We consider the set of index sets of cardinality m , $\mathcal{P}_m = \{S \subseteq \mathbb{N}_n : |S| = m\}$, where $\mathbb{N}_n := \{1, \dots, n\}$. If $S \in \mathcal{P}_m$ and $A \in \mathbb{R}^{n \times m}$, we denote by A_S the $m \times m$ matrix obtained by removing all rows from A with an index not in S . Let $e = [1 \dots 1]^T \in \mathbb{R}^m$. If A_S is nonsingular, then the vectors $u_S = A_S^{-1}e$ and $\hat{u}_S = u_S / \|u_S\|$ are well defined.

LEMMA 4.3. *Let $A \in \mathbb{R}^{n \times m}$, $n \geq m$, and $S \in \mathcal{P}_m$ be such that A_S is nonsingular. Let $p \in S^{m-1}$ and $\rho \in [0, \pi/2)$ be such that $a_i \in \partial \text{cap}(p, \rho)$ for $(i \in S)$. Then $p = \hat{u}_S$ and $\cos \rho = \|u_S\|^{-1}$.*

Proof. If $a_i \in \partial \text{cap}(p, \rho)$, then $a_i \cdot p = \cos \rho$. Therefore, $A_S p = (\cos \rho)e$ and $p = (\cos \rho)A_S^{-1}e = (\cos \rho)u_S$. Since $\|p\| = 1$, we have $|\cos \rho| = \|u_S\|^{-1}$ and, using $\rho < \pi/2$, $\cos \rho = \|u_S\|^{-1}$. We conclude that $p = \hat{u}_S$. $\square$

Blocking sets are the key tools that allow us to gain information about the distribution of $\mathcal{C}(A)$ when the unit rows of A are random vectors with known distribution. The fact that the blocking set S may have cardinality $|S| < m$ when A is strictly feasible is an obstacle to this analysis requiring further attention. The following two lemmas allow us to overcome this problem in the analysis of upper and lower bounds, respectively, on the tails of $\mathcal{C}(A)$.

LEMMA 4.4. *Let $A \in \mathbb{R}^{n \times m}$ be a rank m matrix with $n \geq m$ unit rows, and let $\text{cap}(p, \rho)$ with $\rho < \pi/2$ contain all rows of A . Then there exist $p' \in S^{m-1}$ and $\rho \leq \rho' < \frac{\pi}{2}$ such that $\text{cap}(p', \rho')$ also contains all rows of A , and at least m of them lie on the boundary $\partial \text{cap}(p', \rho')$.*

Proof. Suppose $S = \{i : a_i \in \partial \text{cap}(p, \rho)\}$ contains fewer than m rows of A . Then there exists a vector $u \in S^{m-1} \cap \text{Span}\{a_i : i \in S\}^\perp$ (where we set $\text{Span} \emptyset^\perp = \mathbb{R}^m$) such that $T := \{i \in S^c : a_i \cdot u \neq 0\} \neq \emptyset$. Without loss of generality we may assume that $p \cdot u \geq 0$. Let i^* be a minimizer of $\min_{i \in T} |(\cos \rho - a_i \cdot p) / a_i \cdot u|$, $\delta = (\cos \rho - a_{i^*} \cdot p) / a_{i^*} \cdot u$, and $p_\delta = p + \delta u$. Then $p \cdot u \geq 0$ implies that $\|p_\delta\| \geq 1$, and hence $\hat{p}_\delta = \|p_\delta\|^{-1} p_\delta$

and $\rho_\delta = \arccos(\|p_\delta\|^{-1} \cos \rho) \in [\rho, \frac{\pi}{2})$ are well defined. The proof of Proposition 4.2 shows that $a_i \in \partial \text{cap}(\widehat{p}_\delta, \rho_\delta)$ for all $i \in S$. Moreover, the definition of δ shows that $a_{i^*} \cdot p_\delta = \cos \rho$, and hence $a_{i^*} \cdot \widehat{p}_\delta = \cos \rho_\delta$. It follows that $a_{i^*} \in \partial \text{cap}(\widehat{p}_\delta, \rho_\delta)$, and thus $|S'| > |S|$, where $S' = \{i : a_i \in \partial \text{cap}(\widehat{p}_\delta, \rho_\delta)\}$. If there exists $a_i \notin \text{cap}(\widehat{p}_\delta, \rho_\delta)$, then $i \notin S$, and

$$(4.2) \quad a_i \cdot p \geq \cos \rho, \quad a_i \cdot p_\delta = a_i \cdot p + \delta a_i \cdot u < \cos \rho$$

implies that $a_i \cdot u \neq 0$, so that $i \in T$. Moreover, (4.2) shows that $\delta(a_i \cdot u)/|a_i \cdot u| < (\cos \rho - a_i \cdot p)/|a_i \cdot u| \leq 0$, which implies that $|\delta| > |(\cos \rho - a_i \cdot p)/a_i \cdot u|$ and contradicts the choice of i^* . Therefore, $\text{cap}(\widehat{p}_\delta, \rho_\delta)$ contains all rows of A . Using this construction recursively, we eventually arrive at the desired p' and ρ' . $\square$

LEMMA 4.5. *Let $\{a_i : i \in \mathbb{N}_m\}$ be linearly independent elements of $\partial \text{cap}(p, \rho) \subset S^{m-1}$ with $\rho \in [0, \pi/2)$. Let $\pi_p : \mathbb{R}^m \rightarrow \text{Span}\{p\}^\perp$ be the orthogonal projection along p . Then $\text{cap}(p, \rho)$ is the SIC of $A = [a_1 \dots a_m]^T$ iff $0 \in \text{conv}(\{\pi_p a_i : i \in \mathbb{N}_m\})$.*

Proof. Let us first remark that it is trivial to see that $\rho < \pi/2$ implies that

$$(4.3) \quad 0 \in \text{cone}\{\pi_p a_i : i \in \mathbb{N}_m\} \Leftrightarrow p \in \text{cone}\{a_i : i \in \mathbb{N}_m\}.$$

Note also that the linear independence of the a_i implies that A is strictly feasible and the SIC for A is unique. We first show the “only if” part. Suppose $0 \notin \text{conv}(\{\pi_p a_i : i \in \mathbb{N}_m\})$. Because of (4.3), Farkas’ lemma implies that there exists $q \in S^{m-1}$ such that $q \cdot a_i > 0$ ($i \in \mathbb{N}_m$) and $q \cdot p < 0$. For $\delta > 0$ let

$$(4.4) \quad p_\delta = p + \delta q, \quad \widehat{p}_\delta = \frac{p_\delta}{\|p_\delta\|}, \quad \text{and} \quad \rho_\delta = \arccos \min_i (a_i \cdot \widehat{p}_\delta).$$

Then $\|p_\delta\|^{-1} = 1 - \delta p \cdot q + \mathcal{O}(\delta^2)$ and $a_i \cdot \widehat{p}_\delta = (\cos \rho + \delta a_i \cdot q)(1 - \delta p \cdot q + \mathcal{O}(\delta^2)) > \cos \rho$ for $i \in \mathbb{N}_m$ and for all $\delta > 0$ small enough. Therefore, $\text{cap}(\widehat{p}_\delta, \rho_\delta)$ contains all a_i and $\rho_\delta < \rho$ for $0 < \delta \ll 1$, showing that $\text{cap}(p, \rho)$ is not the SIC for A . It remains to show the “if” part of the lemma. Because of (4.3) we may assume that $p = \sum_{i=1}^m \lambda_i a_i$ for some $\lambda_i \geq 0$ ($i \in \mathbb{N}_m$). Let us assume that $\text{cap}(p, \rho)$ is not the SIC for A . Then it is not difficult to show that there exists a direction $d \in \text{Span}\{p\}^\perp$ such that $\text{cap}(\widehat{p}_\delta, \rho_\delta)$ contains all a_i and $\rho_\delta < \rho$ for $0 < \delta \ll 1$, where $\widehat{p}_\delta$ and ρ_δ are defined as in (4.4) with q replaced by d . That is to say, $a_i \cdot \widehat{p}_\delta = (a_i \cdot p + \delta a_i \cdot d)/\sqrt{1 + \delta^2} > \cos \rho = a_i \cdot p$ for ($i \in \mathbb{N}_m$) and $0 < \delta \ll 1$, which shows that $a_i \cdot d > 0$ for all i . But then $0 = p \cdot d = \sum_{i=1}^m \lambda_i a_i \cdot d > 0$, which shows that our assumption was wrong and $\text{cap}(p, \rho)$ is indeed the SIC for A . $\square$

5. The input distribution. It is well known that if A is a Gaussian matrix with rows $\{A_i : i \in \mathbb{N}_n\}$, then $\{A_i/\|A_i\| : i \in \mathbb{N}_n\}$ and $\{\|A_i\|^2 : i \in \mathbb{N}_n\}$ are independent random vectors and variables with distributions $\mathcal{U}(S^{m-1})$ and χ_m^2 , respectively, where $\mathcal{U}(S^{m-1})$ is the uniform distribution on the sphere. In the remainder of this paper $(\Omega, \mathcal{F}, \mathbf{P})$ denotes a probability space on which there are defined n i.i.d. random vectors $A_i : \Omega \rightarrow S^{m-1}$ ($i \in \mathbb{N}_n$) with distribution $\mathcal{U}(S^{m-1})$. We then say that $A = [A_1 \dots A_n]^T$ is a *uniform random matrix*. Since $\mathcal{C}(A)$ is invariant under row scaling of A , the distribution of $\mathcal{C}(A)$ is the same as for a random matrix with i.i.d. Gaussian components.

For the remainder of this paper we assume that $m \leq n$ (see section 12 for a remark about the case $m > n$). We adhere to the usual practice in probability theory and say that a property holds *almost surely* (or for almost all $\omega \in \Omega$) if it holds with probability 1. An event that occurs with probability zero is also called a null set.

Remark 5.1. If A is a uniform random matrix, then almost surely A is not ill-posed, all subsets of m row vectors of A are linearly independent, and all subsets of $m + 1$ row vectors are affinely independent. Therefore, almost surely the matrices $\{A_S : S \in \mathcal{P}_m\}$ are all nonsingular, and the random vectors u_S and $\hat{u}_S$ are well defined.

We denote the uniform probability measure on S^{m-1} by ν_{m-1} . An elementary calculation shows that the $(m-1)$ -dimensional volume of $\text{cap}(p, \rho)$ equals the volume of the unit sphere in $\mathbb{R}^{m-1}$ times the integral $I_{m-2}(\rho) = \int_0^\rho \sin^{m-2} x \, dx$. Therefore,

$$(5.1) \quad \mathbf{P}[A_i \in \text{cap}(p, \rho)] = \nu_{m-1}(\text{cap}(p, \rho)) = \frac{I_{m-2}(\rho)}{I_{m-2}(\pi)}.$$

It is also elementary to check by induction that

$$(5.2) \quad I_m(\pi) \geq \frac{2}{\sqrt{m}}.$$

6. Upper tail bounds. The goal of this section is to derive an upper bound on the tail probability $\mathbf{P}[\mathcal{C}(A) \geq t]$ when A is a uniform random $n \times m$ matrix with $n \geq m$. For this event to happen, an LEC and SIC, respectively, for A must have an angular radius close to $\pi/2$. If A is strictly feasible, then by Lemmas 4.3 and 4.4 there exists $S \in \mathcal{P}_m$ such that $\|u_S\| \geq t$ and all rows of A lie on one side of the hyperplane defined by $u(S)$. This motivates the decomposition of the event

$$\mathcal{A}_t^{\text{sf}} = \{\omega \in \Omega : (A(\omega) \text{ strictly feasible}) \wedge (\mathcal{C}(A(\omega)) \geq t)\}$$

up to a null set into simpler events expressed in terms of

$$\begin{aligned} \mathcal{N}_t(S) &= \{\omega \in \Omega : \|u_S(\omega)\| \geq t\}, \\ \mathcal{B}_{\pi/2}^{\text{in}}(S) &= \{\omega \in \Omega : A_i(\omega) \in \text{cap}(\hat{u}_S(\omega), \pi/2) \forall i \notin S\}, \\ \mathcal{S}_{t, \pi/2}(S) &= \mathcal{N}_t(S) \cap \mathcal{B}_{\pi/2}^{\text{in}}. \end{aligned}$$

Likewise, if A is infeasible, then Proposition 4.2 and Lemma 4.3 imply the existence of an index set $S \in \mathcal{P}_m$ such that $\|u_S\| \geq t$, and $\{A_i : i \in S\}$ is the blocking set of an LEC for A . Again up to a null set the event

$$\mathcal{A}_t^{\text{if}} = \{\omega \in \Omega : (A(\omega) \text{ infeasible}) \wedge (\mathcal{C}(A(\omega)) \geq t)\}$$

can be decomposed into simpler events which can be expressed in terms of $\mathcal{N}_t(S)$ and the events

$$\begin{aligned} \mathcal{B}_t^{\text{out}}(S) &= \{\omega \in \Omega : A_i(\omega) \notin \text{cap}(\hat{u}_S(\omega), \arccos(1/t)) \forall i \notin S\}, \\ \mathcal{S}_t(S) &= \mathcal{N}_t(S) \cap \mathcal{B}_t^{\text{out}}. \end{aligned}$$

We will use Lemma 6.1 to bound $\mathbf{P}[\mathcal{C}(A) \geq t]$ in terms of $\mathbf{P}[\mathcal{S}_{t, \pi/2}(S)]$ and $\mathbf{P}[\mathcal{S}_t(S)]$, which in turn will be bounded in terms of $\mathbf{P}[\mathcal{N}_t(S)]$ in (6.6) and (6.7). Finally, a bound on $\mathbf{P}[\mathcal{N}_t(S)]$ is developed in Lemmas 6.2 and 6.3.

LEMMA 6.1. *Let A be a random uniform $n \times m$ matrix with $n \geq m$ and $t \in [1, \infty)$. Then $\mathcal{A}_t^{\text{if}} \setminus (\bigcup_{S \in \mathcal{P}_m} \mathcal{S}_t(S))$ and $\mathcal{A}_t^{\text{sf}} \setminus (\bigcup_{S \in \mathcal{P}_m} \mathcal{S}_{t, \pi/2}(S))$ are null sets.*

Proof. Let $A = A(\omega)$ be infeasible and $\mathcal{C}(A) = 1/\cos \theta \geq t$, where $\theta = \theta(A)$. Then $\arccos(1/t) \leq \theta(A) < \pi/2$. Let $\text{cap}(p, \theta)$ be a LEC for A , and let S be the

corresponding blocking set. By Proposition 4.2 and Remark 5.1, we have $|S| = m$ for almost all $\omega \in \Omega$. Lemma 4.3 then implies that $p = \widehat{u}_S$ and $\arccos(1/\|u_S\|) = \theta$, which shows that $\|u_S\| \geq t$ and $\text{cap}(\widehat{u}_S, \arccos(1/t)) \subseteq \text{cap}(\widehat{u}_S, \arccos(1/\|u_S\|))$ does not contain any of the A_i ($i \notin S$). This shows that $A_t^{\text{if}} \setminus (\bigcup_{S \in \mathcal{P}_m} \mathcal{S}_t(S))$ is a null set. If A is strictly feasible and $\mathcal{C}(A) = 1/|\cos \theta| \geq t$, then $\pi/2 < \theta \leq \arccos(-1/t)$, and hence $\pi/2 > \pi - \theta \geq \arccos(1/t)$. Let $\text{cap}(-p, \pi - \theta)$ be the (unique) SIC for A . By Remark 5.1 for almost all $\omega \in \Omega$ Lemma 4.4 is applicable, and there exist $\rho' \in [\pi - \theta, \pi/2)$ and $p' \in S^{m-1}$ such that $\text{cap}(p', \rho')$ contains all rows of A and $S := \{i : A_i(\omega) \in \partial \text{cap}(p', \rho')\}$ is of cardinality m . Lemma 4.3 implies that $p' = \widehat{u}_S$ and $\rho' = \arccos(1/\|u_S\|)$ for these ω . Moreover, $\text{cap}(\widehat{u}_S, \pi/2) \supset \text{cap}(\widehat{u}_S, \rho')$ contains all rows of A and, in particular, $A_i(\omega)$ ($i \notin S$). This shows that $A_t^{\text{sf}} \setminus (\bigcup_{S \in \mathcal{P}_m} \mathcal{S}_{t, \pi/2}(S))$ is a null set. $\square$

Let $S^* = \mathbb{N}_m \in \mathcal{P}_m$, and let us write $\mathcal{S}_t^* = \mathcal{S}_t(S^*)$, $\mathcal{S}_{t, \pi/2}^* = \mathcal{S}_{t, \pi/2}(S^*)$. For $t \geq 1$,

$$\begin{aligned} \mathbf{P}[\mathcal{C}(A) \geq t] &= \mathbf{P}[A_t^{\text{if}}] + \mathbf{P}[A_t^{\text{sf}}] + \mathbf{P}[(A \text{ ill-posed}) \wedge (\mathcal{C}(A) \geq t)] \\ (6.1) \quad &= \mathbf{P}\left[A_t^{\text{if}} \cap \bigcup_{S \in \mathcal{P}_m} \mathcal{S}_t(S)\right] + \mathbf{P}\left[A_t^{\text{sf}} \cap \bigcup_{S \in \mathcal{P}_m} \mathcal{S}_{t, \pi/2}(S)\right] + 0 \\ (6.2) \quad &\leq \sum_{S \in \mathcal{P}_m} \mathbf{P}[\mathcal{S}_t(S)] + \sum_{S \in \mathcal{P}_m} \mathbf{P}[\mathcal{S}_{t, \pi/2}(S)] \leq \binom{n}{m} \left(\mathbf{P}[\mathcal{S}_t^*] + \mathbf{P}[\mathcal{S}_{t, \pi/2}^*]\right), \end{aligned}$$

where (6.1) follows from Remark 5.1 and Lemma 6.1. Note that

$$\begin{aligned} (6.3) \quad \mathbf{P}[\mathcal{S}_t^*] &= \mathbf{P}[\mathcal{B}_t^{\text{out}}(S^*) \|\mathcal{N}_t(S^*)] \times \mathbf{P}[\mathcal{N}_t(S^*)] \\ (6.4) \quad &= \int_{S^{m-1}} \mathbf{P}[\mathcal{B}_t^{\text{out}}(S^*) \|\mathcal{N}_t(S^*), \widehat{u}_{S^*} = x] \nu_{m-1}(dx) \times \mathbf{P}[\mathcal{N}_t(S^*)] \\ (6.5) \quad &= \int_{S^{m-1}} \left(1 - \nu_{m-1}(\text{cap}(x, \arccos(1/t)))\right)^{n-m} \nu_{m-1}(dx) \times \mathbf{P}[\mathcal{N}_t(S^*)] \\ (6.6) \quad &= \left(1 - \frac{I_{m-2}(\arccos(1/t))}{I_{m-2}(\pi)}\right)^{n-m} \times \mathbf{P}[\mathcal{N}_t(S^*)], \end{aligned}$$

where (6.4) holds because $\mathcal{D}(\widehat{u}_{S^*} \|\mathcal{N}_t(S^*)) \sim \mathcal{U}(S^{m-1})$, (6.5) holds because the random vectors A_i ($i \notin S^*$) are independent of A_j ($j \in S^*$), and (6.6) follows from (5.1). A similar argument shows that

$$(6.7) \quad \mathbf{P}[\mathcal{S}_{t, \pi/2}^*] = 2^{-(n-m)} \mathbf{P}[\mathcal{N}_t(S^*)].$$

The task remains to determine a bound on $\mathbf{P}[\mathcal{N}_t(S^*)]$. For $j \in S^*$ let B_j be the unique unit vector in the orthogonal complement of $\text{Span}\{A_i : i \in S^* \setminus \{j\}\}$ such that $\{A_1, \dots, A_{j-1}, B_j, A_{j+1}, \dots, A_m\}$ forms a positively oriented basis of $\mathbb{R}^m$. Then the random vectors A_j and B_j are independent because A_j is independent of $\{A_i : i \in \mathbb{N}_m \setminus \{j\}\}$.

LEMMA 6.2. *Let $\mathcal{C}_t := \{\omega \in \Omega : \bigcup_{j \in S^*} \{|B_j(\omega) \cdot A_j(\omega)| \leq m/t\}\}$. Then $\mathcal{N}_t(S^*) \setminus \mathcal{C}_t$ is a null set for all $t \geq 1$.*

Proof. Almost surely A_{S^*} is nonsingular, and then $\|u_{S^*}\| \leq \|A_{S^*}^{-1}\| \times \|e\| \leq \|A_{S^*}^{-1}\|_F \sqrt{m}$, so that the inequality $\|u_{S^*}\| \geq t$ implies that $\|A_{S^*}^{-1}\|_F \geq t/\sqrt{m}$. Whenever this holds there exists an index $j \in S^*$ such that $\|A_j^{-1}\| \geq t/m$, where A_j^{-1} denotes the j th column of $A_{S^*}^{-1}$. Now the equation $A_{S^*} A_{S^*}^{-1} = I$ implies that $A_i \cdot A_j^{-1} = \delta_{ij}$

(the Kronecker symbol), which shows that $A_j^{-1}/\|A_j^{-1}\| = \pm B_j$ and $|B_j \cdot A_j| = 1/\|A_j^{-1}\| \leq m/t$. This proves the result. $\square$

LEMMA 6.3. *For all $m \geq 3$, $u \in S^{m-1}$, and $i \in \mathbb{N}_n$ it is true that*

$$\mathbf{P}[|A_i \cdot u| \leq m/t] \leq m^{\frac{3}{2}} t^{-1}.$$

Proof. Since the statement is trivially true when $m \geq t$, we may assume without loss of generality that $m < t$ so that $\arccos(m/t)$ is well defined. Therefore, $|A_i \cdot u| \leq m/t \Leftrightarrow A_i \in \mathbf{cap}(u, \arccos(-m/t)) \cap \mathbf{cap}(-u, \arccos(-m/t))$. Thus,

$$\begin{aligned} \mathbf{P}[|A_i \cdot u| \leq m/t] &= (I_{m-2}(\arccos(-m/t)) - I_{m-2}(\arccos(m/t))) / (I_{m-2}(\pi)) \\ &\leq \frac{2 \int_{\arccos(m/t)}^{\pi/2} \sin x \, dx}{\int_0^\pi \sin^{m-2} x \, dx} \leq \frac{m\sqrt{m-2}}{t}, \end{aligned}$$

where the first inequality follows from the definition of I_{m-2} and the fact that $m \geq 3$, and the last inequality follows from (5.2). $\square$

From Lemma 6.3 we compute the bound we seek as follows: for $t \geq 1$ and $m \geq 3$,

$$\begin{aligned} \mathbf{P}[\mathcal{N}_t(S^*)] &\stackrel{\text{Lem 6.2}}{\leq} \mathbf{P}[\mathcal{C}_t] + 0 \leq \sum_{j \in S^*} \mathbf{P}[|B_j \cdot A_j| \leq mt^{-1}] \\ (6.8) \quad &= \sum_{j \in S^*} \int_{S^{m-1}} \mathbf{P}[|B_j \cdot A_j| \leq mt^{-1} | B_j = x] \nu_{m-1}(dx) \end{aligned}$$

$$(6.9) \quad \leq m^{\frac{5}{2}} t^{-1},$$

where (6.8) uses that B_j is uniformly distributed on the sphere because the A_i ($i \neq j$) are, and (6.9) follows from Lemma 6.3 and the fact that A_j is independent of B_j . We can now give an upper bound on the tail decay of $\mathcal{C}(A)$.

THEOREM 6.4. *For all $t \geq 1$, $m \geq 3$, and $n \geq m$ we have*

$$\mathbf{P}[\mathcal{C}(A) \geq t] \leq \binom{n}{m} \times 2m^{5/2} \times (1 - I_{m-2}(\arccos 1/t)/I_{m-2}(\pi))^{n-m} \times t^{-1}.$$

Proof. The claim follows immediately from (6.2), (6.6), (6.7), and (6.9), together with the fact that $1 - I_{m-2}(\arccos(1/t))/I_{m-2}(\pi) \geq 1/2$. $\square$

7. Lower tail bounds. For fixed problem dimensions (m, n) the upper bound on $\mathbf{P}[\mathcal{C}(A) \geq t]$ derived in Theorem 6.4 decays like $O(t^{-1})$. To show that this is the correct functional dependence in t we will now derive a lower bound of the same form. The idea is to use Lemma 4.5 to identify locations for $\{A_i : i \in \mathbb{N}_m\}$ that guarantee that the corresponding SIC $\mathbf{cap}(P_A, R_A)$ has an angular radius R_A smaller than $\arccos(1/t)$. If all the remaining A_i lie in this SIC, then $\mathcal{C}(A) > t$. In other words, our analysis will rely on the inequality

$$\begin{aligned} \mathbf{P}[\mathcal{C}(A) \geq t] &\geq \mathbf{P}\left[\{A_{m+1}, \dots, A_n \in \mathbf{cap}(P_A, R_A)\} \cap \{R_A \geq \arccos 1/t\}\right] \\ &\geq \left(\frac{I_{m-2}(\arccos 1/t)}{I_{m-2}(\pi)}\right)^{n-m} \times \mathbf{P}[R_A \geq \arccos 1/t]. \end{aligned}$$

We start by defining the notion of angle space. Let $K \subseteq \mathbb{R}^m$ be a convex cone. The convexity of K implies that $K \cap \partial K \cap S^{m-1}$ is a ν -null set and that $K \cap S^{m-1} =$

$(\text{int}(K) \cap S^{m-1}) \cup (K \cap \partial K \cap S^{m-1})$ is a disjoint union. Since $\text{int}(K) \cap S^{m-1}$ is open in S^{m-1} , this shows that $K \subseteq \mathbb{R}^m$ is ν -measurable and the *angle space* $\mathfrak{as}(K) = \nu_{m-1}(K \cap S^{m-1})$ of K is a well-defined concept. The remarks of section 5 show that equivalent definitions of angle space are provided by the relations $\mathfrak{as}(K) = \mathbf{P}[X \in K \cap S^{m-1}] = \mathbf{P}[Y \in K]$, where $X \sim \mathcal{U}(S^{m-1})$ and $Y \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ is a standard multivariate Gaussian vector.

Let $A_i \sim \mathcal{U}(S^{m-1})$ ($i \in \mathbb{N}_m$) be i.i.d. random vectors, and let $\mathcal{LI}_k \subset \Omega$ be the event that $\{A_i(\omega) : i \in \mathbb{N}_k\}$ is a linearly independent set of vectors. Then $\Omega \setminus \mathcal{LI}_k$ is a null set, and for all $\omega \in \mathcal{LI}_k$ there exists a unique orthogonal basis $\{E_i : i \in \mathbb{N}_k\}$ of $\text{Span}\{A_i : i \in \mathbb{N}_k\}$ such that $E_i \cdot A_i > 0$ and $\text{Span}\{E_j : j \in \mathbb{N}_i\} = \text{Span}\{A_j : j \in \mathbb{N}_i\}$ for all $(i \in \mathbb{N}_k)$. In fact, the vectors E_i are obtained by Gram-Schmidt orthogonalization of the vectors A_i . Let us consider the event

$$\mathcal{C}_k^{m-1} = \{\omega \in \mathcal{LI}_k : \text{cone}\{A_i : i \in \mathbb{N}_k\} \supseteq \text{cone}\{E_i : i \in \mathbb{N}_k\}\}.$$

The superscript $(m-1)$ reminds us that the random vectors A_i take values in S^{m-1} . The event $\mathcal{C}_{m-1}^{m-1}$ is useful in simplifying the event that the SIC of $\{A_i : i \in \mathbb{N}_m\}$ has an angular radius smaller than $\arccos(1/t)$: indeed, when $\mathcal{C}_{m-1}^{m-1}$ holds, then

$$\text{cone}\{E_1, \dots, E_{m-1}, A_m\} \subseteq \text{cone}\{A_i : i \in \mathbb{N}_m\},$$

so that if the SIC for $\{E_1, \dots, E_{m-1}, A_m\}$ has an angular radius R_E greater than $\arccos(1/t)$, then so does the SIC for $\{A_i : i \in \mathbb{N}_m\}$. The probabilities $\mathbf{P}[\mathcal{C}_{m-1}^{m-1}]$ and $\mathbf{P}[R_E \geq \arccos(1/t)]$ can be bounded separately, the latter by use of Lemma 4.5. This last part involves an integration over $\pi_{p_\Theta} \text{cone}\{E_i : i \in \mathbb{N}_{m-1}\}$, where π_{p_Θ} is a projection defined below. In order to give a lower bound on this integral, we need to understand how the angle space of a cone changes under orthogonal projections. The following lemma provides such a tool.

LEMMA 7.1. *Let $p_1, p_2 \in S^{m-1}$. Denote by $\mathfrak{as}_i$ ($i = 1, 2$) the angle space defined on $\text{Span}\{p_i\}^\perp$, let π_{p_2} be the orthogonal projection of $\mathbb{R}^m$ along p_2 into $\text{Span}\{p_2\}^\perp$, and let $\pi_{1,2} = \pi_{p_2}|_{\text{Span}\{p_1\}^\perp}$ be its restriction to $\text{Span}\{p_1\}^\perp$. Finally, let K be a convex cone in $\text{Span}\{p_1\}^\perp$. Then*

$$\mathfrak{as}_2(\pi_{1,2}K) \geq \mathfrak{as}_1(K) \times |p_1 \cdot p_2|.$$

Proof. If $p_1 \cdot p_2 = 0$, then the bound is trivial. Therefore, without loss of generality, $p_1 \cdot p_2 \neq 0$, and then $\pi_{1,2}$ is a vector space isomorphism between $\text{Span}\{p_1\}^\perp$ and $\text{Span}\{p_2\}^\perp$. Let $\{e_j : j \in \mathbb{N}_{m-2}\}$ be an orthonormal basis of $\text{Span}\{p_1\}^\perp \cap \text{Span}\{p_2\}^\perp$. For $(i = 1, 2)$, let $e_j^{(i)} = e_j$ ($j \in \mathbb{N}_{m-2}$), and let $e_{m-1}^{(i)}$ be chosen so that $\{e_j^{(i)} : j \in \mathbb{N}_{m-1}\}$ is an orthonormal basis of $\text{Span}\{p_i\}^\perp$. We denote the coordinates defined by these bases on $\text{Span}\{p_1\}^\perp$ and $\text{Span}\{p_2\}^\perp$ by $y \in \mathbb{R}^{m-1}$ and $z \in \mathbb{R}^{m-1}$, respectively. Expressed in these coordinate systems, $\pi_{1,2}$ becomes the matrix

$$\pi_{1,2} = \begin{pmatrix} \mathbf{I} & 0 \\ 0 & e_{m-1}^{(2)} \cdot e_{m-1}^{(1)} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & 0 \\ 0 & \pm |p_1 \cdot p_2| \end{pmatrix}.$$

Now let the random vector Z have the multivariate standard normal density $f_Z(z) = (2\pi)^{-(m-1)/2} \exp(-\sum_{i=1}^{m-1} z_i^2/2)$. Then $\mathfrak{as}_2(\pi_{1,2}K) = \mathbf{P}[Z \in \pi_{1,2}(K)] = \mathbf{P}[Y \in K]$,

where the random vector $Y = \pi_{1,2}^{-1}Z$ has the density

$$\begin{aligned} f_Y(y) &= f_Z(z(y)) \left| \det \left(\frac{\partial z_i}{\partial y_j} \right) \right| \\ &= (2\pi)^{-\frac{m-1}{2}} \exp \left(-\frac{1}{2} \left(\sum_{i=1}^{m-2} y_i^2 + y_{m-1}^2 |p_1 \cdot p_2|^2 \right) \right) \times |p_1 \cdot p_2| \\ &\geq (2\pi)^{-\frac{m-1}{2}} \exp \left(-\frac{1}{2} \sum_{j=1}^{m-1} y_j^2 \right) \times |p_1 \cdot p_2|. \end{aligned}$$

The last inequality holds because $|p_1 \cdot p_2| < 1$. Therefore,

$$\mathbf{P}[Y \in K] \geq |p_1 \cdot p_2| \int_K (2\pi)^{-\frac{m-1}{2}} \exp \left(-\sum_{i=1}^{m-1} y_i^2 / 2 \right) dy = |p_1 \cdot p_2| \times \mathbf{a}\mathbf{s}_1(K). \quad \square$$

The Stiefel manifold $V_{m,k}$ of $m \times k$ real matrices with orthonormal columns is a compact homogeneous manifold on which the group of $m \times m$ orthogonal matrices acts transitively as follows:

$$\begin{aligned} O_m \times V_{m,k} &\rightarrow V_{m,k}, \\ (\Omega, X) &\mapsto \Omega X. \end{aligned}$$

The uniform measure $\nu_{m,k}$ on $V_{m,k}$ is the unique O_m -invariant Borel measure that satisfies $\nu_{m,k}(V_{m,k}) = 1$. We also write $\mathcal{U}(V_{m,k})$ for the uniform distribution on $V_{m,k}$. Note that, in particular, $S^{m-1} \equiv V_{m,1}$ and $\nu_{m-1} \equiv \nu_{m,1}$. Since the distribution $\mathcal{D}(A_i) = \mathcal{U}(V_{m,1})$ is O_m -invariant for all i , the following identities follow from spatial symmetry, where $[X_1 \dots X_k]$ denotes the matrix with columns $X_1, \dots, X_k$:

$$(7.1) \quad \mathcal{D}([E_1 \dots E_{m-1}] | \mathcal{C}_{m-1}^{m-1}) = \mathcal{U}(V_{m,m-1}),$$

$$(7.2) \quad \mathcal{D} \left(\frac{\pi_{E_m}(-A_m)}{\|\pi_{E_m}(-A_m)\|} \middle| [E_1 \dots E_{m-1}], \mathcal{C}_{m-1}^{m-1} \right) = \mathcal{U}(S^{m-2}),$$

$$(7.3) \quad \mathbf{P}[\mathcal{C}_k^{m-1}] = \mathbf{P}[\mathcal{C}_k^{k-1}], \quad (k \leq m).$$

Together with (7.3), the following lemma provides the required bound on $\mathbf{P}[\mathcal{C}_{m-1}^{m-1}]$.

LEMMA 7.2. *For $(m \geq 1)$ let $\{A_i : i \in \mathbb{N}_m\}$ be i.i.d. uniform S^{m-1} -valued random vectors. Then*

$$\mathbf{P}[\mathcal{C}_m^{m-1}] \geq 2^{-\frac{2+m(m-1)}{2}}.$$

Proof. We use induction over m . For $m = 1$ we have $\mathbf{P}[\text{cone}\{A_1\} \supseteq \text{cone}\{E_1\}] = 1 \geq 2^{-1}$, which shows that the claim holds true in the base case. Suppose that the result holds true for $m - 1$, and let us show that it holds for m . For $\omega \in \mathcal{L}I_m$, $A_m \notin \text{Span}\{E_i : i \in \mathbb{N}_{m-1}\}$. When this relation holds, let π_{E_m} be the orthogonal projection of $\mathbb{R}^m$ along $E_m(\omega)$ onto $\text{Span}\{E_i : i \in \mathbb{N}_{m-1}\}$. Let

$$\mathcal{D}_m = \{\omega \in \mathcal{L}I_m : \pi_{E_m}(-A_m) \in \text{cone}\{E_i : i \in \mathbb{N}_{m-1}\} \setminus \{0\}\}.$$

We claim that

$$(7.4) \quad \mathcal{D}_m \subseteq \{E_m \in \text{cone}\{E_1, \dots, E_{m-1}, A_m\}\}.$$

In fact, if $\omega \in \mathcal{D}_m$, then there exist numbers $\mu_i \geq 0$ ($i \in \mathbb{N}_{m-1}$) and $\mu_m > 0$ such that $A_m = \mu_m E_m + \pi_{E_m} A_m = \mu_m E_m - \sum_{i=1}^{m-1} \mu_i E_i$. Hence, $E_m = \mu_m^{-1}(A_m + \sum_{i=1}^{m-1} \mu_i E_i)$, which proves (7.4). Clearly, (7.4) implies that $\mathcal{D}_m \cap \mathcal{C}_{m-1}^{m-1} \subseteq \mathcal{C}_m^{m-1}$. The induction hypothesis, the identities (7.1)–(7.3), and the fact that $\text{as}(\text{cone}\{E_i : i \in \mathbb{N}_m\}) = 2^{-m}$ ($m \in \mathbb{N}$) therefore imply

$$\begin{aligned} \mathbf{P}[\mathcal{C}_m^{m-1}] &\geq \mathbf{P}[\mathcal{D}_m \cap \mathcal{C}_{m-1}^{m-1}] \\ &= \int_{V_{m,m-1}} \mathbf{P}[\mathcal{D}_m \| [E_1 \dots E_{m-1}] = v, \mathcal{C}_{m-1}^{m-1}] \times \mathbf{P}[dv \| \mathcal{C}_{m-1}^{m-1}] \times \mathbf{P}[\mathcal{C}_{m-1}^{m-1}] \\ &\geq \int_{V_{m,m-1}} 2^{-(m-1)} \times \nu_{m,m-1}(dv) \times 2^{-\frac{2+(m-1)(m-2)}{2}} = 2^{-\frac{2+m(m-1)}{2}}. \quad \square \end{aligned}$$

THEOREM 7.3. *Let A be a uniform random $n \times m$ matrix, where $n \geq m \geq 2$. Then there exists a constant $c(m) > 0$ that depends only on m such that for all $t \geq \sqrt{m-1}/\sin(\pi/4)$ we have*

$$\mathbf{P}[\mathcal{C}(A) \geq t] \geq c(m) \times (I_{m-2}(\arccos 1/t)/I_{m-2}(\pi))^{n-m} \times t^{-1}.$$

Proof. For $\omega \in \mathcal{C}_m^{m-1}$, $\{A_i : i \in \mathbb{N}_m\}$ and $\{E_1, \dots, E_{m-1}, A_m\}$ are sets of linearly independent vectors. By Proposition 4.2(i), the SICs $\text{cap}(P_A(\omega), R_A(\omega))$ and $\text{cap}(P_E(\omega), R_E(\omega))$ that correspond to these sets are uniquely defined, and then $\omega \in \mathcal{C}_m^{m-1}$ furthermore implies that $\text{cap}(P_E, R_E) \subseteq \text{cap}(P_A, R_A)$ and $0 \leq R_E(\omega) \leq R_A(\omega) \leq \pi/2$. Together with the bound implied by (7.3) and Lemma 7.2, this shows that for $t \geq 1$,

$$(7.5) \quad \mathbf{P}[R_A \geq \arccos 1/t] \geq \mathbf{P}[R_E \geq \arccos 1/t \| \mathcal{C}_{m-1}^{m-1}] \times 2^{-\frac{2+(m-1)(m-2)}{2}}.$$

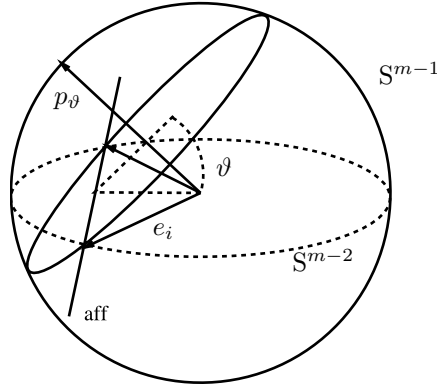
Let $\{e_i : i \in \mathbb{N}_m\}$ be the canonical basis of $\mathbb{R}^m$. Then $\{e_1, \dots, e_{m-1}, A_m\}$ is linearly independent almost surely, and it follows from Proposition 4.2(i) that there exists a unique SIC $\text{cap}(P_e, R_e)$ that corresponds to these vectors. Since A_m is independent of the event $\mathcal{C}_{m-1}^{m-1}$, (7.1) and the O_m -invariance of $\mathcal{D}(A_m)$ imply that

$$\begin{aligned} &\mathbf{P}[R_E \geq \arccos 1/t \| \mathcal{C}_{m-1}^{m-1}] \\ &= \int_{V_{m,m-1}} \mathbf{P}[R_E \geq \arccos 1/t \| [E_1 \dots E_{m-1}] = v, \mathcal{C}_{m-1}^{m-1}] \times \mathbf{P}[dv \| \mathcal{C}_{m-1}^{m-1}] \\ (7.6) \quad &= \int_{V_{m,m-1}} \mathbf{P}[R_e \geq \arccos 1/t] \times \nu_{m,m-1}(dv) = \mathbf{P}[R_e \geq \arccos 1/t]. \end{aligned}$$

To make use of (7.5) and (7.6), we need a lower bound on $\mathbf{P}[R_e \geq \arccos 1/t]$. Deriving such a bound is our next goal. Let $\bar{e} = \sum_{i=1}^{m-1} e_i / \sqrt{m-1}$, and consider the vectors $p_\vartheta = (\cos \vartheta)e_m + (\sin \vartheta)\bar{e}$ for $\vartheta \in [-\pi/2, \pi/2]$. Then a random angle $\Theta : \Omega \rightarrow [-\pi/2, \pi/2]$ is defined almost everywhere by the condition $A_m(\omega) \in \text{Span}\{p_{\Theta(\omega)}\}^\perp + e_1$. In fact,

$$\Theta = \begin{cases} \arctan \frac{\sqrt{m-1} A_m \cdot e_m}{1 - \sum_{i=1}^{m-1} A_m \cdot e_i} & \text{if } \sum_{i=1}^{m-1} A_m \cdot e_i \neq 1, \\ -\frac{\pi}{2} & \text{if } \sum_{i=1}^{m-1} A_m \cdot e_i = 1 \text{ and } A_m \cdot e_m \neq 0, \\ \text{undefined} & \text{otherwise.} \end{cases}$$

Note that the definition of Θ is symmetric with respect to permutations of the vectors e_i , because $\text{Span}\{p_\vartheta\}^\perp + e_1 = \text{Span}\{p_\vartheta\}^\perp + e_i$ ($i = 2, \dots, m-1$) for

FIG. 7.1. *Tilt-and-shrink of S^{m-2} .*

all $\vartheta \in [-\pi/2, \pi/2)$. Θ has a continuous density function $f_{\Theta} > 0$ and such that $f_{\Theta}(-\vartheta) = f_{\Theta}(\vartheta)$ for all $\vartheta \in (-\pi/2, \pi/2)$. Therefore there exists a constant $c_{\Theta}(m) > 0$ such that

$$(7.7) \quad f_{\Theta}(\vartheta) \geq c_{\Theta}(m) \quad \forall \vartheta \in [-\pi/4, \pi/4].$$

Θ is of interest to us because $\{e_1, \dots, e_{m-1}, A_m\} \subset \partial \text{cap}(p_{\Theta}, \arccos((\sin \Theta)/\sqrt{m-1}))$, which makes it possible to apply Lemma 4.5 to derive the required lower bound on $\mathbf{P}[R_e \geq \arccos 1/t]$.

This also motivates the parameterization of S^{m-1} via the mapping

$$S^{m-2} \times [-\pi/2, \pi/2) \rightarrow S^{m-1}$$

$$\left(\sum_{i=1}^{m-1} \lambda_i e_i, \vartheta \right) \mapsto \sqrt{1 - \frac{\sin^2 \vartheta}{m-1}} \sum_{j=1}^m \mu_j e_j + \frac{\sin \vartheta}{\sqrt{m-1}} p_{\vartheta},$$

where $\sum_{i=1}^{m-1} \lambda_i^2 = 1$ and

$$\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_m \end{pmatrix} = \begin{pmatrix} \frac{\cos \vartheta + (m-2)}{m-1} & \frac{\cos(\vartheta)-1}{m-1} & \cdots & \frac{\cos(\vartheta)-1}{m-1} & \frac{\sin \vartheta}{\sqrt{m-1}} \\ \frac{\cos(\vartheta)-1}{m-1} & \frac{\cos \vartheta + (m-2)}{m-1} & \cdots & \frac{\cos(\vartheta)-1}{m-1} & \frac{\sin \vartheta}{\sqrt{m-1}} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\cos(\vartheta)-1}{m-1} & \frac{\cos(\vartheta)-1}{m-1} & \cdots & \frac{\cos \vartheta + (m-2)}{m-1} & \frac{\sin \vartheta}{\sqrt{m-1}} \\ -\frac{\sin \vartheta}{\sqrt{m-1}} & -\frac{\sin \vartheta}{\sqrt{m-1}} & \cdots & -\frac{\sin \vartheta}{\sqrt{m-1}} & \cos \vartheta \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_{m-1} \\ 0 \end{pmatrix}.$$

The matrix in the display is orthogonal, and its last column represents p_{ϑ} in the basis $\{e_i : i \in \mathbb{N}_m\}$. Thus, the chosen parameterization corresponds to tilting the unit hypersphere $S^{m-2} \subset \text{Span}\{e_i : i \in \mathbb{N}_{m-1}\}$ by the angle ϑ about the affine hull $\text{aff}(\{e_i : i \in \mathbb{N}_{m-1}\})$ and shrinking it by $(1 - \sin^2 \vartheta / (m-1))^{1/2}$ to fit the radius of the hypersphere $S^{m-1} \cap (\text{Span}\{p_{\vartheta}\}^{\perp} + e_1)$ cut out of S^{m-1} by the tilted plane; see Figure 7.1. Varying ϑ from $-\pi/2$ to $\pi/2$, the hypersphere $S^{m-1} \cap (\text{Span}\{p_{\vartheta}\}^{\perp} + e_1)$ skims over all of S^{m-1} with only the points in the ν_{m-1} -null set $\text{aff}(\{e_i : i \in \mathbb{N}_{m-1}\}) \cap S^{m-1}$ being multiply covered. On S^{m-2} our parameterization defines a conditional distribution $\mathcal{D}(A_m | \Theta = \vartheta)$ whose continuous Radon–Nikodym derivative $f_{A_m | \Theta}$ with respect to

ν_{m-2} is zero only at points in $\text{aff}(\{e_i : i \in \mathbb{N}_{m-1}\}) \cap S^{m-2}$. By compactness there exists therefore a constant $c_A(m) > 0$ such that

$$(7.8) \quad f_{A_m \parallel \Theta}(x \parallel \vartheta) \geq c_A(m) \quad \forall (x, \vartheta) \in \{x \in S^{m-2} : x \cdot \bar{e} \leq 0\} \times [-\pi/4, \pi/4].$$

We are finally ready to derive a bound on $\mathbf{P}[R_e \geq \arccos(1/t)]$: Lemma 4.5 shows that for all $t \geq \sqrt{m-1}/\sin(\pi/4)$,

$$\begin{aligned} & \mathbf{P}[R_e \geq \arccos(1/t)] \\ &= \mathbf{P} \left[\left\{ \omega : |\Theta| \leq \arcsin \left(\frac{\sqrt{m-1}}{t} \right) \right\} \cap \left\{ \omega : -\pi_{p_\Theta} A_m \in \text{cone}\{\pi_{p_\Theta} e_i : i \in \mathbb{N}_{m-1}\} \right\} \right] \\ &= \int_{-\arcsin \frac{\sqrt{m-1}}{t}}^{\arcsin \frac{\sqrt{m-1}}{t}} \int_{-\pi_{p_\Theta} \text{cone}\{e_i : i \in \mathbb{N}_{m-1}\}} f_{A_m \parallel \Theta}(x \parallel \vartheta) \times \nu_{m-2}(dx) \times f_\Theta(\vartheta) \times d\vartheta \\ &\stackrel{(7.8), \text{Lem 7.1}}{\geq} \int_{-\arcsin \frac{\sqrt{m-1}}{t}}^{\arcsin \frac{\sqrt{m-1}}{t}} c_A(m) 2^{-(m-1)} |\cos \vartheta| \times f_\Theta(\vartheta) d\vartheta \\ &\stackrel{(7.7)}{\geq} 2^{-(m-1)} c_A(m) c_\Theta(m) \times 2 \int_0^{\arcsin \frac{\sqrt{m-1}}{t}} \cos \vartheta d\vartheta \\ &= 2^{-(m-2)} c_A(m) c_\Theta(m) \sqrt{m-1} \times t^{-1}. \end{aligned}$$

Combining this bound with (7.5) and (7.6), we find

$$\mathbf{P}[R_A \geq \arccos 1/t] \geq 2^{-(1 + \frac{(m-1)(m-2)}{2})} 2^{-(m-2)} c_A(m) c_\Theta(m) \sqrt{m-1} \times t^{-1}.$$

Finally, if $A_{m+1}, \dots, A_n \in \text{cap}(P_A, R_A)$, then $\text{cap}(P_A, R_A)$ is the SIC for $\{A_i : i \in \mathbb{N}_n\}$, and it then follows that $\mathcal{C}(A) = |\cos R_A|^{-1}$. Therefore, we have

$$\begin{aligned} \mathbf{P}[\mathcal{C}(A) \geq t] &\geq \mathbf{P}[\{\omega : A_{m+1}, \dots, A_n \in \text{cap}(P_A, R_A)\} \cap \{\omega : R_A \geq \arccos 1/t\}] \\ &\geq \left(\frac{I_{m-2}(\arccos 1/t)}{I_{m-2}(\pi)} \right)^{n-m} \times 2^{-(m-1 + \frac{(m-1)(m-2)}{2})} c_A(m) c_\Theta(m) \sqrt{m-1} \times t^{-1}. \quad \square \end{aligned}$$

8. Exact tail decay rates. As discussed in the introduction, the complexity of several important families of algorithms for solving the linear feasibility problem depends on $\log \mathcal{C}(A)$ rather than $\mathcal{C}(A)$. Theorems 6.4 and 7.3 show that when A is Gaussian and $n \geq m \geq 2$, then

$$\begin{aligned} (8.1) \quad c(m) \left(\frac{I_{m-2}(\arccos e^{-t})}{I_{m-2}(\pi)} \right)^{n-m} \times e^{-t} &\leq \mathbf{P}[\mathcal{C}(A) \geq e^t] \\ &= \mathbf{P}[\log \mathcal{C}(A) \geq t] \leq \binom{n}{m} 2m^{\frac{5}{2}} \left(1 - \frac{I_{m-2}(\arccos e^{-t})}{I_{m-2}(\pi)} \right)^{n-m} \times e^{-t}, \end{aligned}$$

where the first inequality holds for all $t \geq \sqrt{m-1}/\sin(\pi/4)$ and the second for $t \geq 1$. Since $I_{m-2}(\arccos e^{-t})/I_{m-2}(\pi)$ increases monotonically to $1/2$ for $t \rightarrow \infty$, these formulas show that the exponential decay rate of $\mathbf{P}[\log \mathcal{C}(A) \geq t]$ is exactly -1 .

COROLLARY 8.1. *If A is a random uniform $n \times m$ matrix, then*

$$\lim_{t \rightarrow \infty} \frac{\log \mathbf{P}[\log \mathcal{C}(A) \geq t]}{t} = -1. \quad \square$$

Thus, although the multiplicative constant in (8.1) is a conservative estimate, and hence too large, the formula captures the correct qualitative behavior of the tails of $\log \mathcal{C}(A)$, and any tight upper bound on $\mathbf{P}[\log \mathcal{C}(A) \geq t]$ must be of the form

$$(8.2) \quad \mathbf{P}[\log \mathcal{C}(A) \geq t] \leq c(m, n) \times e^{-t}$$

for some constant $c(m, n)$ that depends on m and n .

Remark 8.2. We strongly suspect that in any tight upper bound the constant $c(m, n)$ grows only polynomially in m and n . The exponentially large binomial term $\binom{n}{m}$ in (8.1) is due to the union bound applied in the proof of Theorem 6.4, which seems unnecessarily conservative.

9. Moment estimates. The so-called probabilistic analysis of LP is primarily concerned with the average running time of algorithms on random input data. Since complexity bounds for interior-point methods are polynomial in $\log \mathcal{C}(A)$, upper bounds on the expectation and higher moments of the running time are easily derived from upper bounds on the corresponding moments of $\log \mathcal{C}(A)$. In the remainder of this paper we will show how the bounds of Theorems 6.4 and 7.3 can be used to estimate the moments of $\mathcal{C}(A)$ and $\log \mathcal{C}(A)$. This is easily achieved using the fact that

$$\mathbf{E}[X] = \int_0^\infty \mathbf{P}[X > x] dx$$

holds for any random variable X that takes only nonnegative values.

Let us consider the function $\varphi : \mathbb{R}_+ \rightarrow (\frac{1}{2}, 1]$ defined by

$$\varphi(t) = 1 - I_{m-2}(\arccos e^{-t}) / I_{m-2}(\pi),$$

and note that φ is continuous and strictly decreasing with $\varphi(0) = 1$ and $\lim_{t \rightarrow \infty} \varphi(t) = 1/2$. We define $f(m, n) := \varphi^{-1}(2^{-1/\sqrt{n}})$. Since $f(m, n) > 0$, the following result follows.

COROLLARY 9.1. *Let A be a uniform random $n \times m$ matrix with $n \geq m \geq 3$. Then for all $\gamma \in \mathbb{R}_+$, we have*

$$\mathbf{E}[(\log \mathcal{C}(A))^\gamma] \leq f(m, n)^\gamma + \binom{n}{m} 2m^{5/2} 2^{-(n-m)/\sqrt{n}} \Gamma(\gamma + 1).$$

Proof. Using (8.1), we find

$$\begin{aligned} \mathbf{E}[(\log \mathcal{C}(A))^\gamma] &= \int_0^\infty \mathbf{P}[(\log \mathcal{C}(A))^\gamma > t] dt \\ &\leq f(m, n)^\gamma + \int_{f(m, n)^\gamma}^\infty \binom{n}{m} 2m^{\frac{5}{2}} (1 - I_{m-2}(\arccos e^{-f(m, n)})) / I_{m-2}(\pi))^{n-m} e^{-t^{1/\gamma}} dt \\ &\leq f(m, n)^\gamma + \binom{n}{m} 2m^{\frac{5}{2}} 2^{-\frac{n-m}{\sqrt{n}}} \int_{f(m, n)^\gamma}^\infty e^{-t^{1/\gamma}} dt \\ &\leq f(m, n)^\gamma + \binom{n}{m} 2m^{\frac{5}{2}} 2^{-\frac{n-m}{\sqrt{n}}} \Gamma(\gamma + 1). \quad \square \end{aligned}$$

In section 10 we will see that the bounds of Corollary 9.1 are particularly useful for understanding the behavior of $\mathcal{C}(A)$ when $n \gg m$. Note, however, that this

bound grows exponentially in m and thus does not directly lead to an “average strong polynomiality” result for LP. In order to obtain such a result, we need to change the cut-off point $f(m, n)^\gamma$. We begin with a technical lemma.

LEMMA 9.2. *Let $\{X_{m,n} : (m, n) \in \mathbb{N} \times \mathbb{N}\}$ be a set of random variables and $\gamma \geq 1$ a real number. Furthermore, let $p(m, n)$ and $t(m, n)$ be functions of m and n such that, for all $t \geq t(m, n)$, one has $\mathbf{P}[X_{m,n} > t] \leq \exp(p(m, n) - t^{1/\gamma})$. Then*

$$\mathbf{E}[X_{m,n}] \leq \max(2^{\gamma-1}p(m, n)^\gamma, t(m, n)) + \Gamma(\gamma + 1)2^{\gamma-1}.$$

Proof. Let $T = \max(2^{\gamma-1}p(m, n)^\gamma, t(m, n))$ and $q = p(m, n)^\gamma$, so that $2^{1/\gamma-1}T^{1/\gamma} \geq q^{1/\gamma}$. Since $\gamma \geq 1$, the function $x \mapsto x^{1/\gamma}$ is concave, so that we have

$$(T + t)^{\frac{1}{\gamma}} \geq 2^{-1}(2T)^{\frac{1}{\gamma}} + 2^{-1}(2t)^{\frac{1}{\gamma}} = 2^{\frac{1}{\gamma}-1}(T^{\frac{1}{\gamma}} + t^{\frac{1}{\gamma}}).$$

Therefore, $p(m, n) - (T + t)^{1/\gamma} \leq q^{1/\gamma} - 2^{1/\gamma-1}T^{1/\gamma} - 2^{1/\gamma-1}t^{1/\gamma} \leq -2^{1/\gamma-1}t^{1/\gamma}$, and hence

$$\begin{aligned} \mathbf{E}[X] &= \int_0^\infty \mathbf{P}[X > t] dt \leq \int_0^T 1 dt + \int_T^\infty \exp(p(m, n) - t^{\frac{1}{\gamma}}) dt \\ &\leq T + \int_0^\infty \exp(-t^{\frac{1}{\gamma}} 2^{\frac{1}{\gamma}-1}) dt = T + \Gamma(\gamma + 1)2^{\gamma-1}. \quad \square \end{aligned}$$

COROLLARY 9.3. *Let A be a uniform random $n \times m$ matrix, where $n \geq m \geq 3$, and let $\gamma \geq 1$ be a constant. Then*

$$\mathbf{E}[(\log \mathcal{C}(A))^\gamma] \leq 2^{\gamma-1}(m \log n + 5/2 \log m + \log 2)^\gamma + \Gamma(\gamma + 1)2^{\gamma-1}.$$

In particular, we have

$$\begin{aligned} \mathbf{E}[\log \mathcal{C}(A)] &\leq m \log n + 5/2 \log m + \log 2 + 1, \\ \mathbf{VAR}(\log \mathcal{C}(A)) &\leq 2(m \log n + 5/2 \log m + \log 2)^2 + 4. \end{aligned}$$

Proof. Equation (8.1) shows that for all $t \geq t(m, n) = 1$,

$$\begin{aligned} \mathbf{P}[(\log \mathcal{C}(A))^\gamma > t] &= \mathbf{P}[\log \mathcal{C}(A) > t^{1/\gamma}] \\ &\leq \binom{n}{m} 2m^{\frac{5}{2}} (1 - I_{m-2}(\arccos e^{-t^{1/\gamma}}) / I_{m-2}(\pi))^{n-m} e^{-t^{1/\gamma}} \\ &\leq n^m 2m^{\frac{5}{2}} e^{-t^{1/\gamma}} = \exp\left(m \log n + \frac{5}{2} \log m + \log 2 - t^{1/\gamma}\right). \end{aligned}$$

All claims now follow from Lemma 9.2. $\square$

Note that if the hunch of Remark 8.2 were true, then the bounds of Corollary 9.3 would be logarithmic in m and n . Although Corollary 9.3 recovers the main result in [14], it still does not exhaust the full potential of Theorem 6.4: in section 11 we will show that $\mathbf{E}[\log \mathcal{C}(A)]$ is asymptotically bounded by $m \log 2$ for arbitrary $n \geq m$ and by $\mathcal{O}(m^\gamma)$ for all $\gamma > 0$ when $n \geq 5m$. Let us finally bound the moments of $\mathcal{C}(A)$ itself.

COROLLARY 9.4. *For a uniform random $n \times m$ matrix A with $n \geq m \geq 3$,*

$$\mathbf{E}[\mathcal{C}(A)^\gamma] \begin{cases} = +\infty & \text{if } \gamma \geq 1, \\ \leq 1 + \binom{n}{m} 2m^{\frac{5}{2}} \frac{\gamma}{1-\gamma} & \text{if } \gamma \in (0, 1). \end{cases}$$

Proof. Theorem 7.3 shows that for all $t \geq (\sqrt{m-1}/\sin(\pi/4))^\gamma$,

$$\mathbf{P}[\mathcal{C}(A)^\gamma > t] \geq c(m)(I_{m-2}(\arccos t^{-1/\gamma})/I_{m-2}(\pi))^{n-m} \times t^{-1/\gamma}.$$

Therefore, for all $\gamma \geq 1$,

$$\begin{aligned} \mathbf{E}[\mathcal{C}(A)^\gamma] &= \int_0^\infty \mathbf{P}[\mathcal{C}(A)^\gamma > t] dt \\ &\geq c(m) \times \left(\frac{I_{m-2}\left(\arccos\left(\frac{\sin(\pi/4)}{\sqrt{m-1}}\right)\right)}{I_{m-2}(\pi)} \right)^{n-m} \times \int_{\left(\frac{\sqrt{m-1}}{\sin(\pi/4)}\right)^\gamma}^\infty t^{-\frac{1}{\gamma}} t^{-\frac{1}{\gamma}} dt = \infty. \end{aligned}$$

On the other hand, using Theorem 6.4, we find that for all $\gamma \in (0, 1)$,

$$\begin{aligned} \mathbf{E}[\mathcal{C}(A)^\gamma] &= 1 + \int_1^\infty \mathbf{P}[\mathcal{C}(A)^\gamma > t] dt \\ &\leq 1 + \binom{n}{m} 2m^{\frac{5}{2}} \int_1^\infty t^{-\frac{1}{\gamma}} dt = 1 + \binom{n}{m} \frac{2m^{\frac{5}{2}}\gamma}{1-\gamma}. \quad \square \end{aligned}$$

The exponential growth of the moment bound of Corollary 9.4 for $\gamma < 1$ does not reflect the correct limiting behavior in n : in Corollary 10.3 we will show that $\lim_{n \rightarrow \infty} \mathbf{E}[\mathcal{C}(A)^\gamma] = 1$ for all $\gamma < 1$ and $m \in \mathbb{N}$ fixed.

10. Limit theorems for $n \gg m$. For the purposes of this section let $m \geq 3$ be a fixed dimension, $(A_i)_\mathbb{N}$ a sequence of i.i.d. random vectors with uniform distribution $\mathcal{D}(A_i) \sim \mathcal{U}(S^{m-1})$ on the sphere, and let $(A^{[n]})_\mathbb{N}$ be the sequence of random matrices $A^{[n]} = [A_1 \dots A_n]^\mathsf{T}$.

THEOREM 10.1. *With $A^{[n]}$ defined as above,*

$$\mathbf{P}\left[\lim_{n \rightarrow \infty} \mathcal{C}(A^{[n]}) = 1\right] = 1.$$

Proof. For all $\omega \in \Omega$ and $n \in \mathbb{N}$ let $\mathbf{cap}(P_n(\omega), R_n(\omega))$ be a LEC of $A^{[n]}$. By virtue of Proposition 4.1 it suffices to prove that

$$(10.1) \quad \mathbf{P}\left[\lim_{n \rightarrow \infty} R_n = 0\right] = 1.$$

Let $\rho \in (0, \pi/2)$ be a fixed radius. Since S^{m-1} is compact, there exists a finite set of vectors $\{p_1, \dots, p_k\} \subset S^{m-1}$ such that $\bigcup_{i=1}^k \mathbf{cap}(p_i, \rho/2) = S^{m-1}$. Denote by $\mathcal{CE}_{i,n} = \{\omega \in \Omega : A_j \notin \mathbf{cap}(p_i, \rho/2) \ (j \in \mathbb{N}_n)\}$ the event that the i th spherical cap does not contain any of the n first vectors of $(A_i)_\mathbb{N}$. Then $\mathbf{P}[\mathcal{CE}_{i,n}] = (I_{m-2}(\pi - \rho/2)/I_{m-2}(\pi))^n$, and hence

$$(10.2) \quad \mathbf{P}\left[\bigcup_{i=1}^k \mathcal{CE}_{i,n}\right] \leq \sum_{i=1}^k \mathbf{P}[\mathcal{CE}_{i,n}] = k \times \left(\frac{I_{m-2}(\pi - \rho/2)}{I_{m-2}(\pi)}\right)^n.$$

We now claim that

$$(10.3) \quad \{\omega \in \Omega : R_n \geq \rho\} \subseteq \bigcup_{i=1}^k \mathcal{CE}_{i,n}.$$

Indeed, if $\omega \notin (\bigcup_{i=1}^k \mathcal{CE}_{i,n})$, then there exist indices $i \in \{1, \dots, k\}$ and $j \in \mathbb{N}_n$ such that $P_n(\omega) \in \text{cap}(p_i, \rho/2)$ and $A_j(\omega) \in \text{cap}(p_i, \rho/2)$. Using the triangular inequality on the sphere we find $R_n \leq \arccos(P_n(\omega) \cdot A_j(\omega)) < 2 \times \rho/2$. This shows that $\{\omega \in \Omega : R_n < \rho\} \supseteq (\bigcup_{i=1}^k \mathcal{CE}_{i,n})^c$, which is equivalent to our claim. Now (10.2) and (10.3) show that $\mathbf{P}[R_n \geq \rho] \leq k \times (I_{m-2}(\pi - \rho/2)/I_{m-2}(\pi))^n \xrightarrow{n \rightarrow \infty} 0$. Finally, since $n_1 \leq n_2$ implies that $R_{n_1} \geq R_{n_2}$, we have $0 \leq \mathbf{P}[\lim_{n \rightarrow \infty} R_n > \rho] \leq \lim_{n \rightarrow \infty} \mathbf{P}[R_n > \rho] = 0$. Since this is true for all $\rho \in (0, \pi/2)$, (10.1) follows. $\square$

COROLLARY 10.2. *Let $(A^{[n]})_{\mathbb{N}}$ be as above, and let us denote convergence in probability by $\xrightarrow{P}$ and weak convergence by $\Rightarrow$. Then the following are standard consequences of Theorem 10.1; see for example Theorem 25.2 in [5]:*

- (i) $\mathcal{C}(A^{[n]}) \xrightarrow{P} 1$, (iv) $\log \mathcal{C}(A^{[n]}) \xrightarrow{P} 0$,
- (ii) $\mathcal{C}(A^{[n]}) \xrightarrow{\infty} 1$, (v) $\log \mathcal{C}(A^{[n]}) \xrightarrow{\infty} 0$,
- (iii) $\mathbf{P}\left[\lim_{n \rightarrow \infty} \log \mathcal{C}(A^{[n]}) = 0\right] = 1$. $\square$

Using (8.1) and Corollary 10.2 one can analyze the asymptotic behavior of $\mathbf{E}[\mathcal{C}(A)]$ and $\mathbf{E}[\log \mathcal{C}(A)]$; see Corollary 10.3. For example, using Skorohod's theorem, one can show that $\lim_{n \rightarrow \infty} \mathbf{E}[\log \mathcal{C}(A^{[n]})] = 0$. However, using the estimates of Corollary 9.1 one can prove this result directly.

COROLLARY 10.3. *For $m \in \mathbb{N}$ fixed and $(A^{[n]})_{\mathbb{N}}$ defined as above, we have*

- (i) $\lim_{n \rightarrow \infty} \mathbf{E}[(\log \mathcal{C}(A^{[n]}))^{\gamma}] = 0 \quad \forall \gamma > 0$,
- (ii) $\lim_{n \rightarrow \infty} \mathbf{E}[(\mathcal{C}(A^{[n]}))^{\gamma}] = 1 \quad \forall \gamma \in [0, 1)$.

Proof. (i) Since $2^{-(n-m)/\sqrt{n}}$ is exponentially decreasing in n and $\binom{n}{m} \leq n^m$ increases only polynomially in n , the second term in the estimate of Corollary 9.1 tends to zero as n tends to infinity. In addition, $f(m, n)$ tends to zero as n tends to infinity by the definition of $f(m, n)$.

(ii) The result is trivial for $\gamma = 0$. Let us therefore assume that $\gamma \in (0, 1)$. Let $t_0 \in \mathbb{R}_+$ be large enough so that $1 - I_{m-2}(\arccos t^{-1/\gamma})/I_{m-2}(\pi) \leq 3/4$ for all $t \geq t_0$. Since $\mathcal{C}(A^{[n]}) \xrightarrow{P} 1$ by Corollary 10.2, for all $\epsilon > 0$ there exists a number $n_{\epsilon} \in \mathbb{N}$ such that

$$(10.4) \quad \mathbf{P}[\mathcal{C}(A^{[n]}) > (1 + \epsilon)^{\frac{1}{\gamma}}] < \epsilon \quad \forall n \geq n_{\epsilon}.$$

Therefore, for $n \geq n_{\epsilon}$,

$$\begin{aligned} \mathbf{E}[(\mathcal{C}(A^{[n]}))^{\gamma}] &= \int_0^{\infty} \mathbf{P}[(\mathcal{C}(A^{[n]}))^{\gamma} > t] dt \\ &\leq \int_0^{(1+\epsilon)^{1/\gamma}} 1 dt + \int_{(1+\epsilon)^{1/\gamma}}^{t_0} \mathbf{P}[\mathcal{C}(A^{[n]}) > (1 + \epsilon)^{1/\gamma}] dt + \int_{t_0}^{\infty} \mathbf{P}[\mathcal{C}(A^{[n]}) > t^{1/\gamma}] dt \\ &\stackrel{\text{Thm 6.4, (10.4)}}{\leq} (1 + \epsilon)^{1/\gamma} + \epsilon(t_0 - (1 + \epsilon)^{1/\gamma}) + \binom{n}{m} 2m^{5/2} (3/4)^{n-m} \int_{t_0}^{\infty} t^{-1/\gamma} dt \\ &\leq (1 + \epsilon)^{1/\gamma} + \epsilon(t_0 - (1 + \epsilon)^{1/\gamma}) + 2n^m m^{5/2} (3/4)^{n-m} (1/\gamma - 1)^{-1} t_0^{1-1/\gamma}. \end{aligned}$$

Taking limits as $n \rightarrow \infty$ and observing that $\epsilon > 0$ was arbitrary, the claim follows. $\square$

We find $\lim_{n \rightarrow \infty} \mathbf{E}[\log \mathcal{C}(A^{[n]})] = 0$, and $\lim_{n \rightarrow \infty} \mathbf{VAR}(\log \mathcal{C}(A^{[n]})) = 0$ when m is fixed. These are trivial consequences of the above result. Note also that $\lim_{n \rightarrow \infty} \mathbf{E}[\mathcal{C}(A)^\gamma] = +\infty$ for $\gamma \geq 1$, as follows from Corollary 9.4.

11. Limit theorems for $m \gg 1$. One of the effects of the so-called concentration of measure phenomenon is that the uniform measure of a high-dimensional sphere is largely concentrated around any given great circle. In order to analyze the probability tails of $\mathcal{C}(A)$ for large m , it therefore becomes necessary to know a sufficient condition for a sequence $(\rho_m)_{\mathbb{N}} \subset (0, \pi/2]$ with $\lim_{m \rightarrow \infty} \rho_m = \pi/2$ to have the property that $\lim_{m \rightarrow \infty} \mathbf{P}[A^{[m]} \in \text{cap}(p, \rho_m)] = 1/2$. Such a condition is developed in Lemma 12.1 of Appendix A, where we prove that for any fixed real exponent $\gamma > 0$,

$$(11.1) \quad \lim_{m \rightarrow \infty} \frac{I_{m-2}(\arccos e^{-m^\gamma})}{I_{m-2}(\pi)} = \frac{1}{2}.$$

In Corollary 11.1 we will show that $\mathbf{E}[\log \mathcal{C}(A)]$ is asymptotically linearly bounded in m . The linear bound is due to the appearance of a binomial term in (8.1). If the hunch of Remark 8.2 is true, then for $n \geq m \gg 1$, $\mathbf{E}[\log \mathcal{C}(A)]$ should decrease faster than any algebraic rate in m ; that is,

$$(11.2) \quad \mathbf{E}[\log \mathcal{C}(A)] = o(m^\gamma)$$

for all $\gamma > 0$. When $n \geq 5m$, one can actually prove (11.2).

COROLLARY 11.1. *If $A^{[n,m]}$ are $n \times m$ matrices with i.i.d. uniform random rows, then*

$$\begin{aligned} \text{(i)} \quad & \limsup_{m \rightarrow \infty} \sup_{n \geq m} \frac{1}{m} \mathbf{E}[\log \mathcal{C}(A^{[n,m]})] \leq \log 2, \\ \text{(ii)} \quad & \limsup_{m \rightarrow \infty} \sup_{n \geq 5m} \frac{1}{m^\gamma} \mathbf{E}[\log \mathcal{C}(A^{[n,m]})] = 0 \quad \forall \gamma > 0. \end{aligned}$$

Proof. (i) Let $0 < \epsilon \ll 1$, and consider $X \sim \mathcal{B}(m+k, (1-\epsilon)/2)$, a binomial random variable. Then

$$\begin{aligned} \binom{m+k}{m} \left(\frac{1+\epsilon}{2} \right)^k &= \left(\frac{2}{1-\epsilon} \right)^m \mathbf{P}[X=m] = \left(\frac{2}{1-\epsilon} \right)^m \\ &\times \mathbf{P} \left[\frac{X - \frac{(m+k)(1-\epsilon)}{2}}{\frac{\sqrt{(m+k)(1-\epsilon^2)}}{2}} \in \left[\frac{m-k-1+\epsilon(m+k)}{\sqrt{(m+k)(1-\epsilon^2)}}, \frac{m-k+1+\epsilon(m+k)}{\sqrt{(m+k)(1-\epsilon^2)}} \right] \right]. \end{aligned}$$

By the central limit theorem,

$$\mathcal{D} \left(\frac{X - \frac{(m+k)(1-\epsilon)}{2}}{\frac{\sqrt{(m+k)(1-\epsilon^2)}}{2}} \right) \xrightarrow{m+k \rightarrow \infty} \mathcal{N}(0, 1).$$

Hence, there exists $m_\epsilon \in \mathbb{N}$ such that for $m + k \geq m_\epsilon$,

$$\begin{aligned} \mathbf{P} \left[\frac{X - \frac{(m+k)(1-\epsilon)}{2}}{\frac{\sqrt{(m+k)(1-\epsilon^2)}}{2}} \in \left[\frac{m-k-1+\epsilon(m+k)}{\sqrt{(m+k)(1-\epsilon^2)}}, \frac{m-k+1+\epsilon(m+k)}{\sqrt{(m+k)(1-\epsilon^2)}} \right] \right] \\ \leq (1+\epsilon) \times \left(\Phi \left(\frac{m-k+1+\epsilon(m+k)}{\sqrt{(m+k)(1-\epsilon^2)}} \right) - \Phi \left(\frac{m-k-1+\epsilon(m+k)}{\sqrt{(m+k)(1-\epsilon^2)}} \right) \right) \\ < (1+\epsilon) \times \frac{2}{\sqrt{(m+k)(1-\epsilon^2)}} \\ \times \frac{1}{\sqrt{2\pi}} \times \exp \left(-\frac{1}{2} \min \left(\frac{(m-k+1+\epsilon(m+k))^2}{(m+k)(1-\epsilon^2)}, \frac{(m-k-1+\epsilon(m+k))^2}{(m+k)(1-\epsilon^2)} \right) \right). \end{aligned}$$

Therefore, $m \geq m_\epsilon$ implies $m + k \geq m_\epsilon$, and then by the above inequalities,

(11.3)

$$\begin{aligned} \binom{m+k}{m} \left(\frac{1+\epsilon}{2} \right)^k &< \exp \left(m \log \frac{2}{1-\epsilon} + \frac{1}{2} \log \frac{2}{\pi} + \frac{1}{2} \log \left(\frac{1+\epsilon}{1-\epsilon} \right) \right. \\ &\quad \left. - \frac{1}{2} \log(m+k) - \frac{1}{2} \min \left(\frac{(m-k+1+\epsilon(m+k))^2}{(m+k)(1-\epsilon^2)}, \frac{(m-k-1+\epsilon(m+k))^2}{(m+k)(1-\epsilon^2)} \right) \right). \end{aligned}$$

Furthermore, (11.1) implies that there exists $m_\epsilon^{[1/2]} \in \mathbb{N}$ such that for $m \geq m_\epsilon^{[1/2]}$,

$$(11.4) \quad \frac{I_{m-2}(\arccos e^{-\sqrt{m}})}{I_{m-2}(\pi)} \leq \frac{1+\epsilon}{2}.$$

Equations (8.1), (11.3), and (11.4) show that for all $m \geq \max(m_\epsilon, m_\epsilon^{[1/2]})$, $n = m+k \geq m$, and $t \geq \sqrt{m}$,

$$\begin{aligned} \mathbf{P} \left[\log \mathcal{C}(A^{[n,m]}) \geq t \right] &\leq \binom{m+k}{m} 2m^{\frac{5}{2}} \left(\frac{1+\epsilon}{2} \right)^k e^{-t} \\ &\leq \exp \left(m \log \frac{2}{1-\epsilon} + \log 2 + \frac{1}{2} \log \left(\frac{1+\epsilon}{1-\epsilon} \right) + \frac{5}{2} \log m - t \right), \end{aligned}$$

and Lemma 9.2 then implies that for the same range of parameter values,

$$\begin{aligned} \mathbf{E} \left[\log \mathcal{C}(A^{[n,m]}) \right] &\leq 1 \\ &\quad + \max \left(\sqrt{m}, (m+1) \log 2 - m \log(1-\epsilon) + \frac{1}{2} \log \left(\frac{1+\epsilon}{1-\epsilon} \right) + \frac{5}{2} \log m \right). \end{aligned}$$

Since ϵ was arbitrary, the claim now follows easily.

(ii) We need to consider (11.3) again. For $k \geq 4m$ and ϵ small enough,

$$\begin{aligned} \frac{1}{2} \log \left(\frac{1+\epsilon}{1-\epsilon} \right) + m \log \frac{2}{1-\epsilon} \\ < \frac{1}{2} \min \left(\frac{(m-k+1+\epsilon(m+k))^2}{(m+k)(1-\epsilon^2)}, \frac{(m-k-1+\epsilon(m+k))^2}{(m+k)(1-\epsilon^2)} \right), \end{aligned}$$

and then, for all $m \geq m_\epsilon$,

$$(11.5) \quad \binom{m+k}{m} \left(\frac{1+\epsilon}{2} \right)^k \leq \exp \left(\frac{1}{2} \log \frac{2}{\pi} \right).$$

Furthermore, (11.1) shows that there exists $m_\epsilon^{[\gamma]} \in \mathbb{N}$ such that $m \geq m_\epsilon^{[\gamma]}$ implies $I_{m-2}(\arccos e^{-m^{\gamma/2}})/I_{m-2}(\pi) \leq \frac{1+\epsilon}{2}$. The claim now follows by the same argument as in (i), with $m_\epsilon^{[1/2]}$ replaced by $m_\epsilon^{[\gamma]}$ and (11.3) by (11.5). $\square$

12. A final remark about the case $n < m$. With little extra effort, uniform random $n \times m$ matrices with $n < m$ can be analyzed using the results derived for the case $n \geq m$, and one obtains the estimates

$$\begin{aligned} \mathbf{P}[\mathcal{C}(A) \geq t] &\leq n^{5/2} t^{-1}, \\ \mathbf{E}[\log \mathcal{C}(A)] &\leq 5/2 \log n + 1. \end{aligned}$$

We will not do this here, since it is already known that $\mathbf{E}[\log \mathcal{C}(A)] \leq 5/2 \log n + 2$; see [14].

Appendix A: A concentration of measure inequality.

LEMMA 12.1. *Let $(X^{[m]})_{\mathbb{N}}$ be a sequence of random vectors such that $X^{[m]} \sim \mathcal{U}(S^{m-1})$, and let $(p^{[m]})_{\mathbb{N}}$ be a sequence of vectors such that $p^{[m]} \in S^{m-1}$. Then for any fixed $\gamma > 0$,*

$$\lim_{m \rightarrow \infty} \mathbf{P}[X \in \text{cap}(p, \arccos \exp(-m^\gamma))] = \frac{1}{2}.$$

Proof. For simplicity we write $X = X^{[m]}$ and $p = p^{[m]}$. Note that for $\rho < \pi/2$,

$$(12.1) \quad \mathbf{P}[X \in \text{cap}(p, \rho)] = \frac{I_{m-2}(\rho)}{I_{m-2}(\pi)} = \frac{1}{2} \frac{\int_0^\rho \sin^{m-2} \tau d\tau}{\int_0^\rho \sin^{m-2} \tau d\tau + \int_\rho^{\pi/2} \sin^{m-2} \tau d\tau}.$$

Let $\theta \in (0, 1/2)$ and $m_\theta \in \mathbb{N}$ be such that $\theta < (1 - 1/(m_\theta - 3))/2$. Then for $m \geq m_\theta$ (12.1) implies that

$$(12.2) \quad \mathbf{P}[X \in \text{cap}(p, \rho)] < \theta \Leftrightarrow \int_0^\rho \sin^{m-2} \tau d\tau < 2\theta \int_0^{\pi/2} \sin^{m-2} \tau d\tau.$$

It is easy to show by induction and partial integration that

$$(12.3) \quad \int_\rho^{\pi/2} \sin^m \tau d\tau = \begin{cases} \frac{\cos \rho}{m} \left[\sin^{m-1} \rho + \sum_{k=0}^{\frac{m-3}{2}} \frac{(m-1)(m-3)\dots(m-2k-1)}{(m-2)(m-4)\dots(m-2k-2)} \sin^{m-2k-3} \rho \right] & \text{if } m \text{ is odd,} \\ \frac{\cos \rho}{m} \left[\sin^{m-1} \rho + \sum_{k=0}^{\frac{m}{2}-2} \frac{(m-1)(m-3)\dots(m-2k-1)}{(m-2)(m-4)\dots(m-2k-2)} \sin^{m-2k-3} \rho \right] \\ \quad + \frac{(m-1)(m-3)\dots 3 \times 1}{m(m-2)\dots 2} \left(\frac{\pi}{2} - \rho \right) & \text{if } m \text{ is even.} \end{cases}$$

In particular,

$$(12.4) \quad \int_0^{\pi/2} \sin^m \tau d\tau = \begin{cases} \frac{(m-1)(m-3)\dots 4 \times 2}{m(m-2)\dots 3 \times 1} & \text{if } m \text{ is odd,} \\ \frac{(m-1)(m-3)\dots 3 \times 1}{m(m-2)\dots 2} \times \frac{\pi}{2} & \text{if } m \text{ is even.} \end{cases}$$

It follows from (12.3) and (12.4) that

$$\int_{\rho}^{\pi/2} \sin^m \tau d\tau < \begin{cases} \frac{(m-1)(m-3)\dots 2}{m(m-2)\dots 1} \times \frac{1-\sin^{m+1}\rho}{\cos \rho} = \frac{1-\sin^{m+1}\rho}{\cos \rho} \int_0^{\pi/2} \sin^m \tau d\tau & \text{if } m \text{ is odd,} \\ \frac{(m-1)(m-3)\dots 3}{m(m-2)\dots 2} \left[(\tan \rho)(1-\sin^m \rho) + \frac{\pi}{2} - \rho \right] \\ = \left[\frac{2 \sin \rho (1-\sin^m \rho)}{\pi \cos \rho} + 1 - \frac{2\rho}{\pi} \right] \times \int_0^{\pi/2} \sin^m \tau d\tau & \text{if } m \text{ is even.} \end{cases}$$

$$< \frac{(1-\sin^{m+1}\rho) + 2(1-\sin^2\rho)}{\cos \rho} \times \int_0^{\pi/2} \sin^m \tau d\tau \quad (\text{in both cases}),$$

where the last inequality holds at least for $\pi/4 \leq \rho < \pi/2$. Therefore,

$$\begin{aligned} \int_0^{\rho} \sin^m \tau d\tau &= \int_0^{\pi/2} \sin^m \tau d\tau - \int_{\rho}^{\pi/2} \sin^m \tau d\tau \\ (12.5) \quad &> \left[1 - \frac{(1-\sin^{m+1}\rho) + 2(1-\sin^2\rho)}{\cos \rho} \right] \times \int_0^{\pi/2} \sin^m \tau d\tau. \end{aligned}$$

Note that the combination of (12.2) and (12.5) implies that for $\pi/4 \leq \rho < \pi/2$ and all $k \in \mathbb{N}$ the inequality

$$(12.6) \quad 2\theta < 1 - \frac{(1-\sin^{m+k}\rho) + 2(1-\sin^2\rho)}{\cos \rho}$$

implies $\mathbf{P}[X \in \text{cap}(p, \rho)] \geq \theta$. We set $k = 2$ if m is even and $k = 1$ if m is odd, and we write $\rho_m = \arccos \exp(-m^\gamma)$. For all m large enough it is true that $(m+k)/2 < \exp(2m^\gamma)$, and then

$$1 - \frac{(1-\sin^{m+k}\rho_m) + 2(1-\sin^2\rho_m)}{\cos \rho_m} \geq 1 - e^{m^\gamma(1-2\log \frac{m+1}{2})} - 2e^{-m^\gamma} \xrightarrow{m \rightarrow \infty} 1.$$

Therefore, for $m \geq m_\theta$ large enough, $\rho_m \in (\pi/4, \pi/2)$, and the condition on the left-hand side of (12.6) is satisfied. This shows that

$$(12.7) \quad \lim_{m \rightarrow \infty} \mathbf{P} \left[X \in \text{cap} \left(p, \arccos e^{-m^\gamma} \right) \right] \geq \theta.$$

We now have $\lim_{m \rightarrow \infty} \mathbf{P}[X \in \text{cap}(p, \arccos \exp(-m^\gamma))] \geq 1/2$, since (12.7) is true for any $\theta \in (0, 1/2)$. Moreover, $\lim_{m \rightarrow \infty} \mathbf{P}[X \in \text{cap}(p, \arccos \exp(-m^\gamma))] \leq 1/2$ holds trivially, and the result follows. $\square$

Appendix B: Complexity of the relaxation method. This appendix is to illustrate that the typical complexity of relaxation methods for the solution of the linear system $Ax \leq 0$, $x \neq 0$, is proportional to $\mathcal{C}(A)^2$. This seems well known, but for lack of a reference in the required form and in the interest of completeness we include a short proof. The variant we consider is the so-called perceptron algorithm [36]. When applied to solving a strictly feasible system $Ax < 0$, where $A \in \mathbb{R}^{n \times m}$ with unit row vectors, this algorithm starts from an initial point $x_0 \in \mathbb{R}^m$ and constructs an iterative sequence of points $(x_i)_{\mathbb{N}} \subset \mathbb{R}^m$ as follows: if $Ax_i < 0$, then x_i is a

solution, and the algorithm stops. Otherwise, a row vector $a^{[i]}$ from A is chosen so that $a^{[i]}x_i \geq 0$, and the next point is computed by $x_{i+1} = x_i - a_i$.

The usual convergence analysis then proceeds as follows; see, e.g., [6]: let $w \in \mathbb{R}^m$ be a solution of $Ax < 0$, let $\alpha = \max_{1 \leq j \leq n} a_j w$, where a_j is the j th row vector of A , and let $w^* = w/|\alpha|$. It is then easy to show that if the algorithm has not stopped before or during iteration i , that is, if $x_{i+1} \neq x_i$, then

$$\|x_{i+1} - w^*\|^2 \leq \|x_i - w^*\|^2 - 1.$$

This shows that at most $\|x_0 - w^*\|^2$ iterations can take place. For simplicity, we can choose x_0 to be the origin, so that the algorithm has complexity $\mathcal{O}(\|w^*\|^2)$.

Now note that $\alpha = \cos \theta(A, w)$, where we use the notation of section 3. Without loss of generality we may assume that w is a unit vector, so that $\|w^*\| = |\cos \theta(A, w)|^{-1}$. In order to minimize the complexity estimate, we need to choose $w = \bar{x}$, so that we find that the algorithm terminates after at most $\mathcal{C}(A)^2$ iterations.

Acknowledgments. We would like to thank the two referees and the associate editor for their valuable input that helped improve the initial draft of the paper.

REFERENCES

- [1] I. ADLER AND N. MEGIDDO, *A simplex algorithm whose average number of steps is bounded between two quadratic functions of the smaller dimension*, J. ACM, 32 (1985), pp. 871–895.
- [2] S. AGMON, *The relaxation method for linear inequalities*, Canadian J. Math., 6 (1954), pp. 382–392.
- [3] K. ANSTREICHER, F. POTRA, AND Y. YE, *Probabilistic analysis of an infeasible-interior-point algorithm for linear programming*, Math. Oper. Res., 24 (1999), pp. 176–192.
- [4] K. BALL, *The reverse isoperimetric problem for Gaussian measure*, Discrete Comput. Geom., 10 (1993), pp. 411–420.
- [5] P. BILLINGSLEY, *Probability and Measure*, 3rd ed., John Wiley and Sons, New York, 1995.
- [6] H. BLOCK AND S. LEVIN, *On the boundedness of an iterative procedure for solving a system of linear inequalities*, Proc. Amer. Math. Soc., 26 (1970), pp. 229–235.
- [7] L. BLUM, F. CUCKER, M. SHUB, AND S. SMALE, *Complexity and real computation: A manifesto*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 6 (1996), pp. 3–26.
- [8] K. BORGWARDT, *Untersuchungen zur Asymptotik der mittleren Schrittzahl von Simplexverfahren in der linearen Optimierung*, Ph.D. thesis, Universität Kaiserslautern, Germany, 1977.
- [9] K. BORGWARDT, *The average number of pivot steps required by the simplex-method is polynomial*, Z. Oper. Res. Ser. A-B, 26 (1982), pp. 157–177.
- [10] K. BORGWARDT, *Some distribution-independent results about the asymptotic order of the average number of pivot steps of the simplex method*, Math. Oper. Res., 7 (1982), pp. 441–462.
- [11] K. BORGWARDT, *The Simplex Method—A Probabilistic Analysis*, Springer-Verlag, Berlin, 1987.
- [12] Y. CENSOR AND S. ZENIOS, *Parallel Optimization: Theory, Algorithms and Applications*, Oxford University Press, New York, 1997.
- [13] D. CHEUNG AND F. CUCKER, *A new condition number for linear programming*, Math. Program., 91 (2001), pp. 163–174.
- [14] D. CHEUNG AND F. CUCKER, *Probabilistic analysis of condition numbers for linear programming*, J. Optim. Theory Appl., 114 (2002), pp. 55–67.
- [15] D. CHEUNG, F. CUCKER, AND Y. YE, *Linear programming and condition numbers under the real number computation model*, in Handbook of Numerical Analysis, Vol. XI, P. Ciarlet and F. Cucker, eds., North-Holland, Amsterdam, 2003, pp. 141–207.
- [16] F. CUCKER, *Approximate zeros and condition numbers*, J. Complexity, 15 (1999), pp. 214–226.
- [17] F. CUCKER AND J. PEÑA, *A primal-dual algorithm for solving polyhedral conic systems with a finite-precision machine*, SIAM J. Optim., 12 (2001), pp. 522–554.
- [18] F. CUCKER AND S. SMALE, *Complexity estimates depending on condition and round-off error*, J. ACM, 46 (1999), pp. 113–184.
- [19] F. CUCKER AND M. WSCHEBOR, *On the expected condition number of linear programming problems*, Numer. Math., 94 (2002), pp. 419–478.

- [20] J. DUNAGAN, D. SPIELMAN, AND S.-H. TENG, *Smoothed Analysis of Renegar's Condition Number for Linear Programming*, <http://theory.lcs.mit.edu/~spielman>, (2003).
- [21] A. EDELMAN, *Eigenvalues and Condition Numbers of Random Matrices*, Ph.D. thesis, MIT, Cambridge, MA, 1989.
- [22] R. M. FREUND AND J. R. VERA, *Condition-based complexity of convex optimization in conic linear form via the ellipsoid algorithm*, SIAM J. Optim., 10 (1999), pp. 155–176.
- [23] R. FREUND AND J. VERA, *Some characterizations and properties of the “distance to ill-posedness” and the condition measure of a conic linear system*, Math. Program., 86 (1999), pp. 225–260.
- [24] J.-L. GOFFIN, *On the Finite Convergence of the Relaxation Method for Solving Systems of Inequalities*, Ph.D. thesis, University of California, Berkeley, CA, 1971.
- [25] J.-L. GOFFIN, *The relaxation method for solving systems of linear inequalities*, Math. Oper. Res., 5 (1980), pp. 388–414.
- [26] P. HUHN AND K. BORGWARDT, *An upper bound for the average number of iterations required in phase II of an interior-point method*, in Operations Research Proceedings 1997, Springer-Verlag, Berlin, 1998, pp. 19–24.
- [27] P. HUHN AND K. BORGWARDT, *Interior-point methods: Worst-case and average-case analysis of a phase-I algorithm and a termination procedure*, J. Complexity, 18 (2002), pp. 833–910.
- [28] N. MEGIDDO, *Improved asymptotic analysis of the average number of steps performed by the self-dual simplex algorithm*, Math. Program., 35 (1986), pp. 140–172.
- [29] M. MINOUX, *Probabilistic bounds on one step objective/potential function improvement in Karmarkar's algorithm*, RAIRO Rech. Opér., 28 (1994), pp. 329–355.
- [30] T. MOTZKIN AND I. SCHÖNBERG, *The relaxation method for linear inequalities*, Canadian J. Math., 6 (1954), pp. 393–404.
- [31] J. REID, *On the method of conjugate gradients for the solution of large sparse linear equations*, in Large Sparse Sets of Linear Equations, J. Reid, ed., Academic Press, London, 1971, pp. 231–254.
- [32] J. RENEGAR, *Is it possible to know a problem instance is ill-posed?*, J. Complexity, 10 (1994), pp. 1–56.
- [33] J. RENEGAR, *Some perturbation theory for linear programming*, Math. Program., 65 (1994), pp. 73–91.
- [34] J. RENEGAR, *Incorporating condition measures into the complexity theory of linear programming*, SIAM J. Optim., 5 (1995), pp. 506–524.
- [35] J. RENEGAR, *Linear programming, complexity theory and elementary functional analysis*, Math. Program., 70 (1995), pp. 279–351.
- [36] R. ROSENBLATT, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*, Spartan Books, Washington, DC, 1962.
- [37] M. SHUB AND S. SMALE, *Complexity of Bézout's theorem I: Geometric aspects*, J. Amer. Math. Soc., 6 (1993), pp. 459–501.
- [38] S. SMALE, *On the average number of steps of the simplex method of linear programming*, Math. Program., 27 (1983), pp. 241–262.
- [39] S. SMALE, *Mathematical problems for the next century*, Math. Intelligencer, 20 (1998), pp. 7–15.
- [40] M. TODD, *Polynomial expected behavior of a pivoting algorithm for linear complementarity and linear programming problems*, Math. Program., 35 (1986), pp. 173–192.
- [41] M. TODD, *Probabilistic models for linear programming*, Math. Oper. Res., 16 (1991), pp. 671–693.
- [42] M. TODD, *Erratum: Probabilistic models for linear programming*, Math. Oper. Res., 23 (1998), pp. 767–768.
- [43] M. TODD, L. TUNÇEL, AND Y. YE, *Characterizations, bounds and probabilistic analysis of two complexity measures for linear programming problems*, Math. Program., 90 (2001), pp. 59–69.
- [44] L. TREFETHEN AND D. BAU III, *Numerical Linear Algebra*, SIAM, Philadelphia, 1997.
- [45] S. VAVASIS AND Y. YE, *Condition numbers for polyhedra with real number data*, Oper. Res. Lett., 17 (1995), pp. 209–214.
- [46] S. VAVASIS AND Y. YE, *A primal-dual interior point method whose running time depends only on the constraint matrix*, Math. Program., 74 (1996), pp. 79–120.
- [47] J. VERA, *On the complexity of linear programming under finite precision arithmetic*, Math. Program., 80 (1998), pp. 91–123.