Journal of Global Optimization **27:** 233–247, 2003.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

Constraint Selection and Deterministic Annealing

ALISTAIR I. MEES[1,2] and CAMERON TOVEY[2]
[1]*Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, NT, Hong Kong (E-mail: alistair@cado.uwa.edu.au);*
[2]*Centre for Applied Dynamics and Optimization, The University of Western Australia, Perth, Western Australia*

Abstract. The deterministic annealing optimization method is related to homotopy methods of optimization, but is oriented towards global optimization: specifically, it tries to tune a penalty parameter, thought of as "temperature", in such a way as to reach a global optimum. Optimization by deterministic annealing is based on thermodynamics, in the same sense that simulated annealing is based on statistical mechanics. It is claimed to be very fast and effective, and is popular in significant engineering applications. The language used to describe it is usually that of statistical physics and there has been relatively little attention paid by the optimization community; this paper in part attempts to overcome this barrier by describing deterministic annealing in more familiar terms.

The main contribution of this paper is to show explicitly that that constraints can be handled in the context of deterministic annealing by using constraint selection functions, a generalization of penalty and barrier functions. Constraint selection allows embedding of discrete problems into (non-convex) continuous problems.

We also show how an idealized version of deterministic annealing can be understood in terms of bifurcation theory, which clarifies limitations of its convergence properties.

1. Introduction

The method of deterministic annealing is a global optimization heuristic that was invented in the statistical physics community but is now used in other areas. It can be thought of as related to homotopy methods (Kojima et al., 1991), but is specifically oriented towards global optimization.

Like simulated annealing, deterministic annealing uses a temperature parameter that is decreased to zero as the algorithm progresses. It can be thought of as as a thermodynamic version of simulated annealing: that is, it works with continuous variables that obey deterministic laws. The continuous variables can be interpreted as bulk properties of a physical medium, in contrast to simulated annealing's probabilistic—and usually discrete—quantities which can be interpreted as molecular properties. The use of continuous variables and a deterministic optimization algorithm often has considerable advantages for speed of calculation.

As well as being used on the standard test problems such as travelling salesman (Yuille and Kosowsky, 1994), deterministic annealing is being used or proposed

for use in significant applications such as real-time image feature extraction (Puzicha et al., 1997), discrete multicommodity flow (Tovey, 1996), multidimensional scaling applied to data visualization and embedding dimension estimation (Klock and Buhmann, 1996, 1997), combinatorial problems (Dang, 2000; Tsuchiya et al., 2001) and other areas. These are many reports in the literature claiming that deterministic annealing outperforms other global optimization methods such as stochastic annealing or the EM algorithm: it is claimed to deliver optimal, or at least "good", solutions very quickly.

In spite of enthusiasm amongst practitioners, deterministic annealing has received little attention from the optimization community, perhaps because, like simulated annealing before it, deterministic annealing derives largely from the physics community (see for instance, Simic, 1990; Stolorz, 1991; Yuille and Kosowsky, 1994 and Acton and Bovik, 1996) so the language is unfamilliar. This is a pity, both because there are many unanswered questions about its convergence properties and because the method may be considerably improved in at least some applications by using standard methods of optimization theory.

In this paper we describe deterministic annealing in terms that do not require knowledge of statistical mechanics. We also show how it can be extended to handle constrained problems, using a generalization of penalty function and barrier function methods. This is done in Section 2.1, where we also develop a general method of embedding discrete problems in continuous problems. The discrete variables are relaxed to become continuous and, as the temperature parameter decreases, they are tightened back to discrete values in a way that allows continuous optimization methods to be used. We also briefly discuss the imposition of other kinds os constraints.

We should make it clear that the motivation for deterministic annealing is not purely to replicate well-known results involving penalty and barrier functions. The intent is that, by suitably tuning the temperature parameter, one can reach the global optimum in a non-convex problem, at least in idealized cases. In Section 3.2 we interpret deterministic annealing in terms of bifurcations and show that it may, in fact, not converge to the global optimum. In Section 4.1 we discuss how some of the limitations may be overcome in a practical problem, discrete multicommodity flow.

2. Deterministic annealing

Statistical mechanics has given rise to several ideas that are useful in optimization and other areas outside physics. For example, the Gibbs distribution is used in Bayesian statistics and in image processing (see for example, Ruanaidh and Fitzgerald, 1996). The best known optimization method inspired by statistical mechanics is stochastic (or simulated) annealing (Kirkpatrick et al., 1982; Lundy and Mees, 1986; Geman and Geman, 1985), which enables approximate solution of

Table 1. Comparison of deterministic and stochastic annealing

	Stochastic	Deterministic
Concepts	Molecular level	Bulk level
Physical analogy	Spin glass	Elastic net
Temperature	Allows uphill steps	Smooths local fluctuations
Variables	Usually discrete	Mostly continuous
Optimization	Random walk with drift	Gradient dynamics
Speed	Fairly slow	Fast
Convergence	Yes if cooled slowly enough	Not necessarily, but is claimed to give "good" solutions.
Key result	Markov chain equilibrium	Mean field theory

difficult optimization problems using a probabilistic approach inspired by models of energy minimization at the molecular level. Most recently (Simic, 1990; Stolorz, 1991; Yuille and Kosowsky, 1994; Tovey and Mees, 1995; Tovey, 1996) deterministic annealing has appeared as a generalization and improvement of an earlier approach called "mean field annealing".

Table 1 compares stochastic and deterministic annealing. There are two significant differences from the point of view of optimization. The first difference is that the problem to be solved is a deterministic optimization over continuous variable, which makes it possible for deterministic annealing to use the well-known powerful results of continuous optimization theory. The second difference is that there is no guarantee of convergence, even if the temperature parameter is lowered adiabatically ("infinitely slowly"); we will give a counter-example later showing on showing that the global optimum may not be attained even in an idealized version. We remark that one minor advantage of deterministic annealing is that for any problem there are fixed minimum and maximum values of T which may in principle be calculated: this part of the algorithm, at least, is not heuristic.

There are interpretations of deterministic annealing in probabilistic terms (Puzicha et al., 1997) but this does not seem necessary. In this paper our approach is entirely deterministic, in accord with the way the algorithm is commonly used. We assume the problem is

$$\text{minimize } f(x) \quad \text{over } x \in \Omega \quad \text{with } c(x) = 0 \tag{1}$$

where $\Omega \subset \mathbb{R}^n$ is a discrete set, and $f : \Omega \to \mathbb{R}$ and $c : \Omega \to \mathbb{R}^m$ can be extended in some natural way to $\tilde{f} : \mathbb{R}^n \to \mathbb{R}$ and $\tilde{c} : \mathbb{R}^n \to \mathbb{R}^m$. There may be many ways of embedding a given set Ω in $\mathbb{R}^n$ but we assume this has already been done, and that Ω is a finite set of points. (We deal with the functional constraints $c(x) = 0$ separately, in Section 3.1 and Section 4.1.) The extended functions $\tilde{f}$ and $\tilde{c}$ are

assumed to differentiable. From now on we do not distinguish between $\tilde{f}$ and f and between $\tilde{c}$ and c.

In interesting cases, the structure of Ω makes the optimization difficult using standard methods. This may be because the problem is NP-hard (for example, in the travelling salesman problem, Ω is all permutations of $2, \ldots, n$, up to symmetry). Alternatively, Ω may be defined via constraints that are difficult to work with in some way.

The deterministic annealing approach, in the form described in this paper, expresses Ω via some constraints parametrized by $T \in \mathbb{R}$. The constraints are designed so that every global minimum of the continuous problem converges to a global minimum of the discrete problem (1) as $T \to 0$.

2.1. CONSTRAINT SELECTION FUNCTIONS

The fundamental requirement for solving discrete problems by deterministic annealing is to find a method of treating the variables as continuous while representing the discreteness requirements as functional constraints. A useful way to think about discreteness constraints is that any feasible solution to the discrete problem will correspond to a particular subset chosen from among some equality constraints applied to the continuous problem. For example, if we embed an n-dimensional binary optimization problem in the unit cube $[0, 1]^n$ then the constraints are of the form

$$\text{for each } i, \qquad x_i = 0 \quad \text{OR} \quad x_i = 1.$$

It is intuitively clear that "or-ing" constraints in this way cannot be done with conventional penalty functions while retaining convexity; for a proof, see Tovey (1996). This difficulty can be overcome by generalizing the idea of penalty functions.

We introduce the idea of constraint selection functions, which are parametrized by T and which are convex for large T but as $T \to 0$ they have minima which approach all the points of Ω.

DEFINITION 2.1. **(Constraint Selection Function)**. A twice differentiable function $g : \mathbb{R}^n \times \mathbb{R}_+ \to \mathbb{R}_+$ is a constraint selection function for a set $\Omega \in \mathbb{R}^n$ if all of the following hold:

1. there exists $T_{\max} \geqslant 0$ such that $g(x, T)$ is convex in x for all $T \geqslant T_{\max}$;
2. for each $T > 0$ let the set of minima of g be $X(T)$; then $g(x, T) = \theta(T) \geqslant 0$ for all $x \in X(T)$ (i.e., the minima are all equal); and $\theta(T) \to 0$ as $T \to 0$.
3. for small T, the elements of Ω correspond to the minima of g: that is, for all $\varepsilon > 0$ there exists some $T_{\min}$ such that when $T < T_{\min}$, for every $z \in \Omega$ there is some $x \in X(T)$ such that $|z - x| < \varepsilon$, and for every $x \in X(T)$ there is some $z \in \Omega$ such that $|z - x| < \varepsilon$.

4. for small T, $g(x, T)$ is unbounded away from Ω: specifically,

$$g(x, T) \to \infty \quad as\ T \to 0 \quad \text{if } x \notin \Omega$$

A constraint selection function is therefore a function that is convex for large T, and for small T has equal minima at points that approach all the elements of Ω as $T \to 0$. Condition 3 is stronger than necessary but ensures that g has no spurious local minima to confuse an optimization algorithm, and likewise condition 2 could be weakened but by requiring equal values for all minima we ensure g does not bias the solution.

THEOREM 2.2. *The optimization problem*

$$minimize\ f(x) + g(x, T) \quad over\ x \in \mathbb{R}^n \quad with\ c(x) = 0 \tag{2}$$

has the same solution as (1) in the limit $T \to 0$.

The theorem is a straightforward consequence of the way we have defined constraint selection functions. It states that as $T \to 0$, the function g enforces the constraints $x \in \Omega$.

The definition of constraint selection functions does not place any restrictions on ω. So, for example, conventional penalty and barrier functions can be considered as special cases of constraint selection functions. However, in this paper we use constraint selection functions only to handle discreteness constraints and in addition, for notational simplicity, we restrict attention to the common case $\Omega = \{0, 1.\ldots, K\}^n$. For a constraint selection function that works in the case of general discrete sets, see Tovey (1996). In the present case, consider

$$g(x, T) = \sum_{i=1}^{n} h(x_i, T) \tag{3}$$

where

$$h(x_i, T) = -T \log \sum_{k=0}^{K} \exp(-(x_i - k)^2 / T^2). \tag{4}$$

This choice of g is separable in the components of x, though this is not usually directly useful in solving (2) since f is not normally separable.

Although we do not need the statistical physics background, it may help in reading the deterministic annealing literature to know that the functions h are inspired by entropies in thermodynamics and it can be shown (Tovey, 1996) that all thermodynamic averages corresponding to problem (1) are the same as those of problem (2) with the above choice of $g(x, T)$. The key statistical mechanics fact is that in the limit $T \to 0$), the average value of x turns out to be equal to the minimizer of (1) or (2), which is why solving (2) also solves (1). For a proof of

Theorem 2.2 in statistical physics terms, see Tovey (1996) and Tovey and Mees (1995).

2.2. PROPERTIES OF g

To get an informal idea of why g works as a constraint selection function look at its components h:

$$h(x_i, T) = -T \log \left(e^{-x_i^2/T^2} + \sum_{k=1}^{K} e^{-(x_i-k)^2/T^2} \right).$$

Factor out the first term to give

$$h(x_i, T) = \frac{x^2}{T} - T \log \left(1 + \sum_{k=1}^{K} e^{k(2x-k)/T^2} \right).$$

For small T, this looks like a penalty function x^2/T when x is close to 0. In particular, $h(0, T) \to 0$ as $T \to \infty$ and for fixed Δ with $0 < |\Delta| < 1/2$, $h(\Delta, T) \to \infty$ as $T \to \infty$. Since h is symmetric in $k = 0, \ldots, K$, we could instead have factored out any term to see that h looks like $(x - k)^2/T$ for x close to k and $h(k, T) \to 0$ as $T \to \infty$.

To prove that g is a constraint selection function requires checking the conditions in the definition. Most of this is straightforward. We only show the proof for convexity here, which has the nice feature that it gives an explicit value for $T_{\max}$.

PROPOSITION 2.3. *If $T \geqslant K$ then $g(x, T)$ is convex in $x \in \mathbb{R}^n$*

Proof. Define

$$S = \sum_{k=0}^{K} e^{-(x_i-k)^2/T^2}$$

so that $h = -T \log S$. Then

$$\frac{\partial h}{\partial x_i} = -\frac{T}{S} \frac{\partial S}{\partial x_i}, \qquad \frac{\partial^2 h}{\partial x_i^2} = \frac{T}{s^2} \left(\frac{\partial S}{\partial x_i} \right) - \frac{T}{S} \frac{\partial^2 S}{\partial x_i^2}.$$

But

$$\frac{\partial S}{\partial x_i} = -\frac{2}{T^2} L, \qquad \frac{\partial^2 S}{\partial x_i^2} = \frac{4}{T^4} Q - \frac{2}{T^2} S$$

where $L = \sum_k (x_i - k) \exp(-(x_i - k)^2/T^2)$ and $Q = \sum_k (x_i - k)^2 \exp(-(x_i - k)^2/T^2)$.

Hence

$$\frac{\partial^2 h}{\partial x_i^2} = \frac{2}{T^3 S^2}(2L^2 - 2QS + T^2 S^2).$$

Writing out explicitly the term in parentheses we get

$$\begin{aligned} 2L^2 - 2QS + T^2 S^2 &= \sum_{k,j} e^{-(x_i-k)^2/T^2} e^{-(x_i-j)^2/T^2} \left(2(x_i-k)(x_i-j)\right. \\ &\quad \left. -2(x_i-k)^2 + T^2\right) \\ &= \sum_{k,j} e^{-(x_i-k)^2/T^2} e^{-(x_i-j)^2/T^2} \left(2(x_i-k)(k-j) + T^2\right) \end{aligned}$$

In the double summation, when $k = j$ the part in parentheses reduces to T^2 and is positive. The terms with $k \neq j$ the occur symmetrically, so we can add each pair to get

$$\begin{aligned} 2((x_i-k)-(x_i-j))(k-j) + 2T^2 &= 2(T^2 - (k-j)^2) \\ &\geqslant 2(T^2 - K^2) \end{aligned}$$

where we have replaced $(k-j)^2$ by its maximum value K^2. Hence every term in the summation is non-negative if $T \geqslant K$, so h has positive second derivative under that condition.

Since g is additive in h, g is also convex when $T \geqslant K$, completing the proof. □

Remark: It is easily verified by calculation of the second derivative that in the binary case ($K = 1$), $T_{\max} = 1/\sqrt{2}$, showing that the bound $T \geqslant K$ is conservative.

The following result is obtained from the Proposition 2.3 together with a straightforward check on the remaining conditions of Definition 2.1.

LEMMA 2.4. *$g(x, T)$ defined by (3) and (4) is a constraint selection function.*

3. Solution methods

The idea of deterministic annealing is to try to solve the problem

$$\text{minimize } L(x, T) = f(x) + g(x, T) \quad \text{over } x \in \mathbb{R} \quad \text{with } c(x) = 0 \tag{5}$$

for temperature T starting at a large value and decreasing to 0. We have seen that g is convex for large enough T and looks like a quadratic penalty function near each element of Ω. Proposition 2.3 and Lemma 2.4 show that solving (5) is equivalent to solving (1).

Any reasonable minimization algorithm will find a local minimum of (5) and hence a local minimum of (1) as $T \to 0$. The difficulty is, of course, that we have

to find a *global* minimum of (5) as $T \to 0$. It is hoped that, if we start with a global minimum $\hat{x}(T)$ at a high enough value of T (say, $T = K$ if f is convex) and follow it as T decreases adiabatically, then $\hat{x}(T)$ will converge to the global optimum of (5) and hence of (1).

As we have seen, choosing a suitable initial temperature is easy if the optimization problem without g (i.e., the problem of minimizing $f(x)$ over $\mathbb{R}^n$ subject to $c(x) = 0$) is convex. If f is not convex, it may be possible to generalize it to $\tilde{f}(x, T)$ where $\tilde{f}(x, T)$ is convex in x for large T; we will not consider this case in the present paper.

3.1. PENALTIES AND CONSTRAINTS

One way to incorporate some or all of the functional constraints $c(x) = 0$ is to use penalty functions, for example, $\sum_{i \in I} c_i(x)^2 / T$ where I indexes those constraints chosen to be applied by penalty functions. This seems to be the commonest way to incorporate constraints in the published deterministic annealing literature, but there is no reason to insist on applying constraints in this way. To allow the use of penalty methods without excluding other methods, we redefine the optimization problem as

$$\text{minimize } L(x, T) = f(x, T) + g(x, T) \quad \text{over } x \in \mathbb{R}^n \quad \text{with } \tilde{c}(x) = 0 \quad (6)$$

where now $f(x, T)$ may incorporate some constraints as penalties if desired, Any remaining constraints represented by $\tilde{c}(x) = 0$ are to be dealt with in some other way, such as Lagrangian methods. We replace $\tilde{c}$ by c from now on.

We also remark that although we do not include them explicitly, inequality constraints may either be left as they are or treated by barrier functions, again absorbing the barrier functions into $f(x, T)$.

3.2. BIFURCATIONS

An idealized representation of deterministic annealing is as a solution continuation method. We solve the problem for a large enough value of T, and use a continuation method (Kojima et al., 1991) to explicitly follow the initial solution as T decreases. Different local minima will appear as a result of bifurcations on the path being followed, or will appear in a different path. The question of whether a global optimum is reached is equivalent to asking whether the solution branch that the continuation method follows, starting at $\hat{x}(T_0)$, leads to the global optimum. Some simple examples show that the global solution may or may not be reached.

Figure 1 shows the path for the problem

$$\text{minimize } (x - 1/4)^2 \quad \text{over } x \in \{0, 1\}$$

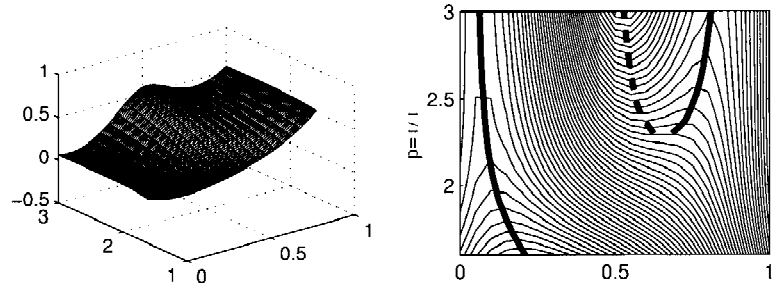

Figure 1. Minimizing $(x - 1/4)^2$ over $x \in \{0, 1\}$ by deterministic annealing. The surface and contour plot both show the total energy function (objective function plus constraint selection function) as a function of x and $\beta = 1/T$. The surface plot demonstrates convexity in x at high temperatures and lack of convexity at low temperatures. The contour plot also shows the locations of stationary points. (Minima: solid lines, maxima: dashed line.) As T decreases the number of stationary points in x increases from 1 to 3, but the global solution is on the path traced out by the continuation of the single minimum which occurs for all values of $T > 1/\sqrt{2}$.

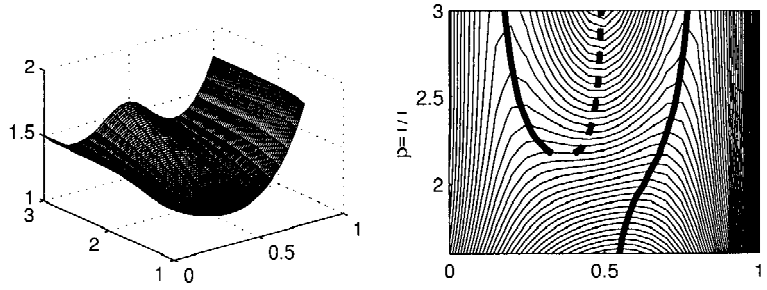

Figure 2. Minimizing $x^4 + 1.5/x + 1)$ over $x \in \{0.1\}$ by deterministic annealing. As before, the number of stationary points in x increases from 1 to 3, but the global solution is not on the path traced out by the continuation of the single minimum but on an isola, created by a saddle-node bifurcation which generates a minimum and a maximum at a point far from the minimum for the then current value of T.

and we see that for large T there is a unique minimum while for small T there are two minima (solid lines) and one local maximum (dashed line). The minimum at high values of T is connected by a non-branching path to the solution as $T \to 0$, and the deterministic annealing method will deliver the correct solution.

Figure 2 shows the path for the problem

$$\text{minimize } x^4 + \frac{3/2}{1+x} \quad \text{over } x \in \{0, 1\}$$

and we see that, as before, for large T there is a unique minimum while for small T there are two local minima (solid lines) and one local maximum (dashed line). This time, however, the minimum at high values of T is connected by a non-branching

path to a point which is *not* the solution as $T \to 0$, and the deterministic annealing problem will deliver the wrong solution.

The behavior can be understood via bifurcation theory. For the function $f =$ constant, the path undergoes a pitchfork bifurcation (Glendinning, 1994) at $T = 1/\sqrt{2}$. The two different choices of f in the above examples unfold the degenerate pitchfork into a non-bifurcating path and a saddle-node bifurcation (at the points marked S in the figures) in two different ways. In one case, the path followed happens to lead to the optimum and in the other case it does not.

In general g will have a highly degenerate bifurcation (a pitchfork simultaneously in all n dimensions) which will be unfolded into non-degenerate, or at least less degenerate, forms by the presence of f. Degenerate bifurcations are difficult to handle numerically and the branch taken by an algorithm will depend on the details of the algorithm. On the other hand, since any unfolding depends on f, we cannot expect to be able to solve all possible problems this way: we have already seen a trivial example where the unfolding is such as to give the wrong answer.

In the examples, the idealized deterministic annealing method in effect just "rounds off" the continuous solution in the sense that it chooses the point in Ω nearest the unconstrained minimum of f. This may be an over-simplified description of the method in higher dimensions, but it is plausible that this idealized version of deterministic annealing will only perform well on problems for which there are many good solutions, or for which rounding off the continuous solution gives good results.

3.3. FINAL TEMPERATURE

If Ω has a finite number of elements there can only by finitely many bifurcations as T decreases, and hence there is a value $T_{\min}$ beyond which no further bifurcations can take place. The system can then be "quenched", that is, cooled quickly to $T = 0$. This is not very helpful in practice because determining $T_{\min}$ is unlikely to be possible. Most implementations seem to be designed to terminate when the solution is much closer to one element of Ω than to any other.

3.4. OTHER SOLUTION METHODS

In practice, continuation methods do not seem to be used directly. More often, practitioners use gradient descent, either as simple steepest descent, or by converting the minimization to a differential equation. A typical scheme might be

$$\dot{x} = -\nabla_x L(x, T)$$
$$\dot{T} = -\epsilon T.$$

where $L(x, T) = f(x, T) + g(x, T)$. The T dynamics are analogous to geometric cooling for stochastic annealing, which is known to be be a bad choice in the stochastic algorithm.

Such a differential equation method can be thought of as approaching but never reaching the solution continuation curve. The loci of maxima divide the space into basins of attraction and the final solution reached may depend on the starting point, if there have been bifurcations along the path from the high-temperature minimum, or if convergence to the minima is not particularly fast (e.g., if ϵ is not small enough). Dependence of solutions on starting conditions, and convergence to poor solutions (or even infeasible solutions, if penalty functions are used) occurs in realistic applications such as knapsack problems.

An improvement suggested by Stolorz (1991) and by Tovey (1996) is to replace $\dot{T}$ equation by

$$\frac{\partial \beta}{\partial t} = +\nabla_{\beta} L(x, \beta)$$

where $\beta = 1/T$. Now a saddle-point method can be used to simultaneously solve the primal and the dual problems. Stolorz calls this *adaptive annealing* and asserts that it performs well. A different β can be used for each constraint, giving "local temperatures".

There is no reason to suppose that gradient descent or any of its improvements will give better solutions in general than curve tracking, with the possible exception of local temperatures if it is made possible for local temperatures to increase provided the overall temperature is decreasing.

A possibly useful approach, if a sophisticated solution continuation method (Doedel and Kernevez, 1998) is used, is to keep track of several branches and pick the best. One retains the best N branches, for some moderate value of N, dropping the worst old ones whenever new branches appear. This improves the chance of finding a good solution, but depends on there being bifurcations on the curve being followed. If all the good solutions are on isolae then this method would not find any of them.

Another possibility is to reintroduce randomness into the algorithm, in the hope of jumping out of local minima. This would complicate the analysis immensely but seems to be used in some applications.

Finally, we should point out that most numerical methods would have difficulties with high order bifurcations: for example, as we pointed out earlier, the constraint selection function has simultaneous pitchfork bifurcations in all variables, so unless the degeneracy is unfolded by f, the Hessian has all its eigenvalues passing through zero at the same value of T. And it does not help that a common way of adding an equality constraint $c(x) = 0$ is to add $1/T\ c(x)^2$ to L, which could lead to further degeneracies.

It is clear that much work is needed to understand which of the proposed improvements is valuable in practice.

4. Applications

In spite of the reservations we have expressed about the theoretical ability of deterministic annealing to find good solutions, it does often perform well in practice. Here we look at one problem; for others, consult the literature already cited.

4.1. MULTICOMMODITY FLOW

The binary multicommodity flow problem is

$$\text{minimize} \sum_{a,j} f_a\left(x_a^1, \ldots, x_a^J\right) \quad x_a^j \in \{0, 1\} \quad \forall a, j$$

with the constraints

$$Ex^j = b^j \ \forall j \quad \text{and} \quad \sum_j |x_a^j| \leqslant c_a \ \ \forall a$$

where x_a^j is the flow of commodity j on arc a. The capacity of arc a is c_a and has to be shared between commodities. If the arcs are unidirectional then $|x_a^j|$ can be replaced by x_a^j in the capacity constraint. The matrix E is the node-arc incidence matrix of the network (Rockafellar, 1984) and the vector b^j is the supply-demand vector for the jth commodity, so $Ex^j = b^j$ represents supply-demand satisfaction for the jth commodity (and flow conservation at nodes which are not supply-demand nodes). This is an NP-hard problem (Garey and Johnson, 1979).

To solve it using deterministic annealing we allow the variables x_a^j to be continuous and add the standard constraint selection function g. The problem is now a continuous nonlinear multicommodity flow problem, which can be solved in any of several ways (Boland et al., 1992). The constraints $Ex^j = b^j$ could be handled by penalty functions but the resulting problem turns out to behave badly (Tovey, 1996); it is better to use active set methods or other standard methods of nonlinear programming.

One way to apply deterministic annealing to this problem is to start with x such that $Ex^j = b^j$ for all j and then only change x^j within the null space of E. In this way we will always have a feasible network flow. Doing so corresponds to changing x^j round a cycle or cycles in the network, that is, to constraining δx^j for every j to lie in circulation space (Rockafellar, 1984). To ensure $E\delta x^j = 0$ we project the gradient onto the appropriate subspace. We can enforce the capacity constraint either by penalty functions or by active set methods; in the following we used penalty functions.

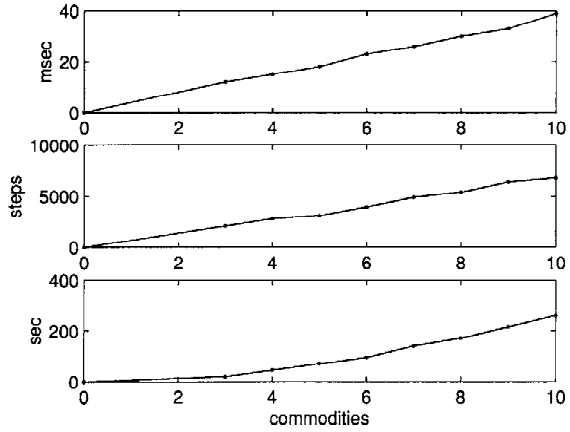

Figure 3. Average performance of deterministic annealing on 50 binary multicommodity flow problems with randomly generated networks of 35 arcs and 10 nodes. In each case the horizontal axis is number of commodities. The upper graph shows time per step (in milliseconds), the middle graph is average number of steps to termination, and the lower graph shows total computation time. It should be noted that an antique computer was used (a 70 MHz 486 processor) and the times should be scaled accordingly for current computers.

4.2. PERFORMANCE

We solved a large number of binary multicommodity flow problems, all having capacity $c_a = 1$ on each arc. The graphs were randomly generated subject to connectivity between all sources and their sinks, and the sources and sinks were themselves randomly selected. In Figure 3 we show some computational results for solution by deterministic annealing, using a projected gradient method as outlined above. The computation time per step and the number of steps to solution are both roughly linear in the number of commodities over the range 1–10 commodities, with the result that the total computation time is approximately quadratic. In every case the terminating solutions were better than or comparable to those obtained by a much slower stochastic annealing progress.

5. Conclusions

The main contribution of this paper has been to show how deterministic annealing may be applied to constrained problems using constraint selection functions, which are generalized penalty or barrier function method. It may be useful for finding locally optimal solutions to discrete problems by embedding them in a parametrized class of continuous problems that converges to the discrete problem.

Although the intention of deterministic annealing is to find globally optimal solutions to non-convex problems, it is not clear that global optimality, or even any particular quality of solution, can be guaranteed by any of the currently pro-

posed variants of the method, though practical experience is encouraging. The combination of standard constrained and unconstrained optimization methods with constraint selection functions may be more powerful than deterministic annealing as currently applied.

References

Acton, S. and Bovik, A. (1992), Generalized deterministic annealling. *IEEE Transactions on Neural Networks*, 7(3), 686–699.

Boland, N., Goh, C. and Mees, A.I. (1992), An algorithm for non-linear network programming: Implementation, results and comparisons. *Journal of the Operational Research Society*, 43(10), 979–992.

Dang, C. (2000), Approximating a solution of the s-t max-cut problem with a deterministic annealing algorithm, *Neural Networks*, 13(7), 801–810.

Doedel, E. and Kernevez, J. (1988), Software for continuation problems in ordinary differential equations with applications. Technical report.

Garey, M. and Johnson, D. (1979), *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W.H. Freeman, San Francisco.

Geman, S. and Geman, D. (1985), Stochastic relaxation, gibbs distribution, and the Bayesian restoration of images. *IEEE Trans. PAMI*, 6(6), 721–741.

Glendinning, P. (1994), *Stability, Instability and Chaos*. Cambridge University Press, Cambridge.

Kirkpatrick, S., Gelatt, C.J. and Vecchi, M. (1982), Optimization of simulated annealing. Technical report, IBM, Yorktown Heights.

Klock, H. and Buhmann, J. (1996), Data visualization by multidimensional scaling: A deterministic annealing approach. Technical Report IAI-TR-96-8, Universitat Bonn, Institut fur Informatik III, Romerstrae 194, Bonn.

Klock, H. and Buhmann, J. (1997), Multidimensional scaling by deterministic annealing. In Klock, H. and Buhmann, J. (eds), *Proceedings EMMCVPR'97*, pp. 245–260. Springer Verlag.

Kojima, M., Megiddo, N. and Noma, T. (1991), Homotopy continuation methods for nonlinear complementarity problems. Mathematics of Operations Research, 16, 754–774.

Lundy, M and Mees, A.I. (1986), Convergence of an annealing algorithm. *Mathematical Programming*, 34, 111–124.

Puzicha, J., Hofmann, T. and Buhmann, J.M. (1997), Deterministic annealing: Fast physical heuristics for real-time optimization of large systems. In *Proceedings of the 15th IMACS World Conference on Scientific Computation, Modelling and Applied Mathematics*.

Rockafellar, R. (1984), *Network Flows and Monotropic Optimization*. Wiley, New York.

Ruanaidh, J.J.K.O. and Fitzgerald, W.J. (1996), *Numerical Bayesian Methods Applied to Signal Processing*, Springer.

Simic, P.D. (1990), Statistical mechanics as the underlying theory of 'elastic' and 'neural' optimizations. *Comp. Neural. Syst.*, 1, 89–103.

Stolorz, P. (1991), Merging constrained optimization with deterministic annealing to "solve" combinatorially hard problems. Technical report, Theoretical Division, los Alamos National Laboratory, September 27.

Tovey, C. (1996), A new deterministic annealing method with application to operational research and decision problems. Master's thesis, Dept. of Mathematics, University of Western Australia.

Tovey, C. and Mees, A.I. (1995), A general approach to deterministic annealing. In Agarwal, R.P. (ed.), *Recent Trends in Optimization Theory and Applications*, volume 5 of *World Scientific Series in Applicable Analysis*, pp. 409–421. World Scientific, Singapore.

Tsuchiya, K., Nishiyama, T. and Tsujita, K. (2001), A deterministic annealing algorithm for a combinatorial optimization problem using replicator equations. *Physica D.*, 149(3), 161–173.

Yuille, A. and Kosowsky, J.J. (1994), Statistical physics algorithms that converge. *Neural Computation*, 6, 341–356.