DIRECT AND ITERATIVE METHODS FOR THE SOLUTION OF LINEAR OPERATOR EQUATIONS IN HILBERT SPACE(1)

BY

W. V. PETRYSHYN

Introduction. During the last two decades the problem of justifying existing methods and finding new ones for the solution of the equation $Lu = f$, where f is a given vector in Hilbert space and L is a given operator, has been investigated by many authors(2). This investigation led to the development of a group of direct methods of Ritz, Galerkin, least squares, and moments; and the iterative methods of steepest descent or gradient method, the method with minimal residuals, and their variants.

These methods have a long history and have been studied and extensively applied by earlier authors to integral and differential equations. Only recently, however, the Hilbert space operator theory has been used for their study. Thus, using the variational principle, Mikhlin [16] proved the convergence of Ritz method for a self-adjoint positive definite operator L and the convergence of the method of least squares for an invertible operator L. In case L is of the form $L = A + B$, where A is self-adjoint and positive definite and B satisfies some additional conditions, the method of Galerkin was investigated by Mikhlin [16], Polsky [19], and others [9; 13]. For similar differential operators the method of moments was studied by Kravchuk [11], Polsky [19], and Zdanov [22].

The convergence and the estimate of error of the gradient method for self-adjoint and positive definite operators L have been studied by Kantorovich [7] and Hayes [6]. The latter considered also a slightly more general problem. In case L is a finite, symmetric, and positive definite matrix the method with minimal residuals was developed by Krasnoselsky and Krein [10]. A powerful method related to direct methods was also developed by Murray [18].

The purpose of this paper is to extend the study and the applicability of these methods to a larger class of linear operator equations(3) than those considered by the above authors and to present these seemingly distinct methods in a more unified manner. In our investigation we do not use the variational principle

Received by the editors March 16, 1961.

(1) Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, under the Joint Committee on Graduate Instruction, Columbia University.

(2) For the discussion of methods considered in this paper and the literature on this subject see [3; 7; 8; 10; 11; 16].

(3) For the definition of the class of operators considered in this paper see §§1.1, 2.2, and 3.2.

applied by most of the above authors. Our argument is of a purely geometrical nature and is essentially based on the notion of projection. Such an approach furnishes not only the geometrical basis for these essentially variational methods but also unifies them and gives a basis for their comparison and for possible discovery of new methods as will be seen below.

The author is greatly indebted to Professor Francis J. Murray, who has been his research adviser, for his constructive criticism and helpful suggestions. The author also wishes to express his sincere appreciation to Professor Richard V. Kadison for his generous help in the preparation of this thesis.

We now begin with the summary of the material presented below.

The first chapter is concerned with the properties of our class of operators which we call the K-positive definite operators (K-p.d.)([3]). Thus in §1.1 we define the K-p.d. operators and discuss some of their properties while in §1.2 we extend the classical variational principle to our class of operators. §1.3 generalizes Friedrichs procedure of extending a symmetric positive definite operator to the extension of K-p.d. operators and other nonsymmetric operators. Some further properties of this extension, called here the solvable generalized Friedrichs extension, which slightly generalize the results of Lax and Milgram [14] are discussed. In §1.4 we derive some additional properties of K-p.d. operators.

The second chapter studies the direct methods. In fact, in §2.1 we study the generalized Ritz method for K-p.d. operators by means of a purely geometrical approach which for particular choices of K reduces to the ordinary Ritz method and the method of least squares. §§2.2 and 2.3 are concerned with the investigation of the generalized method of moments for the equation with $L = A + B$, where A is K-p.d. and B is some linear operator. It is shown that all other direct methods can be deduced as special cases of the generalized method of moments. At the same time the general theoretical justification for the ordinary method of moments, which was lacking, is supplied.

In §2.4 we describe a geometrical procedure which gives the necessary and sufficient conditions for solving an equation and which for a proper choice of coordinate elements reduces in its hypothesis and conclusions to the method proposed by Professsor Francis J. Murray [18]. For that reason we call it the generalized Murray method. The possible extension of the class of operators to which this method is applicable is indicated. §2.5 shows that after the proper renorming of the space in each case the generalized Murray method reduces to one of the methods of Galerkin, Ritz, or least squares.

The greater part of the third chapter deals with bounded operators and is devoted to the presentation of a modified iteration method, called here the method with relative minimal errors (RME-method) which unifies and extends the results of Kantorovich [7], Krasnoselsky and Krein [10], and Hayes [6] to a larger class of operators by means of a purely geometrical approach.

Using the recent result of Greub and Rheinboldt [5] we derive in §3.1 an inequality critical for our work of which the well-known inequalities of Pólya-Szegö, Kantorovich, and Krasnoselsky-Krein are its special cases. There we also show the equivalence of the three special inequalities. In §3.2 we describe the RME-method and derive its convergence and estimate of error while in §3.3 we discuss its special cases. The first of these is the well-known gradient method and the second is the iterative method with minimal residuals studied in [10] in case of a finite, symmetric, and positive definite matrix. Thus we obtain the geometrical basis for these methods and at the same time extend the results of Krasnoselsky and Krein to operators in a Hilbert space. The third special case of the RME-method is a new iterative method, called here the method with minimal errors, which appears to be a very useful and effective procedure especially when applied to nonsymmetric operator equations and which, it seems, has escaped the notice of various investigators in this field.

§3.4 compares the relative merits of these special cases, while §3.5 generalizes the gradient method to the operator equations involving the unbounded K-p.d. operators. The paper is completed by considering in §3.6 a simple iterative method which coincides in principle but not in form with the generalized Ritz method. To avoid the solution of the corresponding linear algebraic equations we give a compact computational scheme based on this iterative method which seems to be very convenient in practice. This scheme was first used by Altman [1].

I. K-POSITIVE DEFINITE OPERATORS AND GENERALIZED VARIATIONAL PRINCIPLE

In this chapter we define our class of operators and discuss some of their important properties.

1.1. **K-positive definite operators.** Let H be a complex and separable Hilbert space(4). We shall first consider the problem of solving the equation

$$Au = f, \quad f \in H, \tag{1.1}$$

where A is a linear unbounded operator defined on a dense domain $D(A)$ in H with the property that there exists a continuously $D(A)$-invertible(5) closed operator K with $D(K) \supseteq D(A)$ and a constant $\alpha > 0$ such that

$$(Au, Ku) \geqq \alpha \| Ku \|^2, \quad u \in D(A). \tag{1.2}$$

The operators A having such property will be called *K-positive definite* (K-p.d.). Let us first observe that the class of K-p.d. operators contains, among others, the class of positive definite operators (when $K = I$) and the class of invertible operators (when $K = A$) as its subclasses. Moreover, it can be easily shown that

(4) The results obtained below are also valid for real Hilbert spaces provided A also satisfies (b) of Lemma 1.1.

(5) If Q is a dense linear set in H, then the operator K will be referred to as continuously Q-invertible if the range R_Q of K, considered as an operator on Q, is dense in H and K has a bounded inverse on R_Q.

for a proper choice of K the ordinary differential operators of odd order, the weakly elliptic partial differential operators of odd order, and others are members of our class. When operators are bounded, the class of K-p.d. operators forms a subclass of symmetrizable operators investigated by Reid [20].

LEMMA 1.1. *If A is K-p.d., then*

(a) *A has a bounded inverse.*

(b) *$(Au,Kv) = (Ku,Av)$ for all u and v in $D(A)$.*

(c) *For all u and v in $D(A)$ one has the generalized Schwarz inequality*

$$|(Au,Kv)|^2 \leqq (Au,Ku)\,(Av,Kv).$$

(d) *A is closeable.*

Proof. (a) Since $(Au,Ku) \leqq \|Au\|\,\|Ku\|$ and $\|Ku\| \geqq \beta\|u\|$ for some $\beta > 0$ (1.2) implies that $\|Au\| \geqq \alpha\beta\|u\|$. Thus, A^{-1} exists and is bounded on the range $R(A)$.

(b) Consider the identity valid for u and v in $D(A)$

$$\begin{aligned}(A(u+v), K(u+v)) - (A(u-v), K(u-v)) \\ + i(A(u+iv), K(u+iv)) - i(A(u-iv), K(u-iv)) = 4\,(Au, Kv)\end{aligned}$$

and an analogous second identity with A and K interchanged. Since the product (Au,Ku) is real-valued, the first members of the above identities are equal and consequently we obtain (b).

(c) For any u and v in $D(A)$, let $w = u + \lambda(Au,Kv)v$. Then for all real λ

$$0 < (Aw,Kw) = (Au,Ku) + 2\lambda|(Au,Kv)|^2 + \lambda^2|(Au,Kv)|^2(Av,Kv)$$

is a polynomial of λ of second degree with real coefficients. Hence it cannot have two real distinct roots. This gives (c).

(d) Let $\{u_n\}$ be a sequence in $D(A)$ and f an element in H such that $u_n \to 0$ and $Au_n \to f$ as $n \to \infty$. (1.2) implies that $\|Au\| \geqq \alpha\|Ku\|$ for all u in $D(A)$. Hence the convergence of $\{Au_n\}$ implies the convergence of $\{Ku_n\}$ and since $u_n \to 0$ and K is closed we have $Ku_n \to 0$ as $n \to \infty$. Let v be any element in $D(A)$; then, in view of (b),

$$(f,Kv) = \lim_{n\to\infty}(Au_n,Kv) = \lim_{n\to\infty}(Ku_n,Av) = 0.$$

Since K is continuously $D(A)$-invertible, $f = 0$, i.e., A is closeable.

THEOREM 1.1. *If A is K-p.d. and $D(A) = D(K)$, then there exists a constant $\theta > 0$ such that for all u in $D(K)$*

$$\|Au\| \leqq \theta\|Ku\|. \tag{1.3}$$

Furthermore, the operator A is closed, $R(A) = H$, and the equation $Au = f$ has a unique solution for each f in H[6].

---

[6] In case K is continuous and A is closed a result similar to the second part of Theorem 1.1 was recently proved by Browder [2].

Proof. Introduce in $D(K)$ a new inner product and norm by

$$[u,v]_1 = (Ku,Kv), \quad |u|_1 = \|Ku\|.$$

Since K is closed, $R(K)$ is dense, and K^{-1} is bounded on $R(K)$, then $R(K) = H$ and hence $D(K)$ becomes a complete Hilbert space in this new metric which we shall denote by H_1. Moreover, A considered as a linear operator of H_1 to H is closed. To show this note that since A is defined in all of H_1 it is sufficient to show that A admits in H_1 a closed linear extension i.e., if $\{u_n\}$ is a sequence in H_1 such that $|u_n - 0|_1 \to 0$ and $Au_n \to f$ as $n \to \infty$, then $f = 0$. Since K is continuously invertible $|u_n - 0|_1 \to 0$ implies that $\|u_n - 0\| \to 0$ as $n \to \infty$ and since by Lemma 1.1 (d) A is closeable in H we have $f = 0$. Thus, A is closed in H_1 and being everywhere defined in H_1 it follows from the closed-graph theorem [21] that A is bounded in H_1. Hence there exists a constant $\theta > 0$ such that $\|Au\| \leqq \theta |u|_1 = \theta \|Ku\|$ for each u in H_1. This proves (1.3).

To prove that A is closed let us consider a sequence $\{u_n\}$ in $D(A)$ such that $u_n \to u$ and $Au_n \to f$ as $n \to \infty$. (1.2) implies that $\{Ku_n\}$ converges. But, K is closed and consequently $u \in D(K)$ and $Ku_n \to Ku$ as $n \to \infty$. By (1.3) and the fact that $D(K) = D(A)$ we have

$$\|A(u_n - u)\| \leqq \theta \|K(u_n - u)\| \to 0, \quad \text{as } n \to \infty.$$

Also, $Au_n \to Au$ as $n \to \infty$. This shows that $Au = f$.

To prove the rest of the theorem observe that since A is closed, then, in view of Lemma 1.1 (a), it is sufficient to show that the zero space $Z(A^*) = \{0\}$. Let g be in $Z(A^*)$. Since $D(A^*) \subset R(K)\ (= H)$ and $D(K) = D(A)$, there exists u in $D(K)$ such that $Ku = g$. For such u we have

$$0 = |(u, A^*g)| = |(Au, Ku)| \geqq \alpha \|Ku\|^2 \geqq \alpha\beta^2 \|u\|^2.$$

Thus, $u = 0, g = Ku = 0$, implying that $Z(A^*) = \{0\}$, as was to be shown.

1.2. **Generalized variational principle.** Corresponding to eq. (1.1) construct the functional

$$F(u) = (Au, Ku) - (Ku, f) - (f, Ku) \tag{1.4}$$

The generalization of the variational principle to eq. (1.1) is based on

THEOREM 1.2. *If A is K-p.d., then a necessary and sufficient condition that w be a solution of eq.* (1.1) *is that w realize the minimum of $F(u)$*([7]).

Proof. Let w satisfy eq. (1.1). Then

$$F(u) = (Au, Ku) - (Ku, Aw) - (Aw, Ku),$$

([7]) In case A is of the form $A = CK$, where C is self-adjoint, Theorem 1.2 was stated without proof by Martyniuk [15].

in view of Lemma 1.1 (b), can be written in the form

$$F(u) = (A(u - w), K(u - w)) - (Aw, Kw)$$

showing that $F(u)$ attains its minimum at $u = w$. This disposes of the necessity.

To prove sufficiency, suppose w realizes the minimum of $F(u)$. Let v be any element in $D(A)$ and t any real number. Then $F(w + tv)$, as a function of t, attains its minimum when $t = 0$; therefore, $dF(u + tv)/dt\big|_{t=0} = 0$. Using Lemma 1.1 (b) we get $dF(w + tv)/dt\big|_{t=0} = 2\,\mathrm{Re}\,(Aw - f, Kv) = 0$. Replacing v by iv we obtain $\mathrm{Im}\,(Aw - f, Kv) = 0$. Consequently, $(Aw - f, Kv) = 0$ for every v in $D(A)$. Since K is continuously $D(A)$-invertible this implies that $Aw - f = 0$, i.e., w satisfies eq. (1.1).

SPECIAL CASE. If $K = I$, we have an ordinary variational principle for symmetric positive definite operators which has been extensively studied([8]).

Theorem 1.2 allows us to replace the problem of solving eq. (1.1) by the problem of minimizing $F(u)$. However, it may happen that A is defined on a too restricted domain. In that case the minimum problem (1.4) may have no solution which, however, will exist if $D(A)$, on which $F(u)$ is defined, and with it A, could be somewhat extended. It will be shown that for a K-p.d. operator A it is always possible to extend $D(A)$ so that the problem of minimizing $F(u)$ has a solution.

For that purpose we form the pre-Hilbert space structure on $D(A)$ with the inner product $[u, v] = (Au, Kv)$. The corresponding norm is defined by $|u| = [u, u]^{1/2}$. We complete $D(A)$ to the Hilbert space H_K with respect to the norm which we have just defined. By (1.2) and $\|Ku\| \geqq \beta \|u\|$

$$|u| \geqq \gamma_1 \|Ku\|, \quad \gamma_1 = \alpha^{1/2} > 0, \ u \in D(A), \tag{1.5}$$

$$|u| \geqq \gamma_2 \|u\|, \quad \gamma_2 = \gamma_1 \beta > 0, \ u \in D(A). \tag{1.6}$$

LEMMA 1.2. (a) *The set $D(A)$ is dense in H_K.*

(b) *H_K is a subspace of H in the sense of identifying elements from H_K with elements in H.*

(c) *K can be extended to a bounded linear operator of all of H_K to H.*

(d) *The inequalities* (1.5) *and* (1.6) *are valid for all u in H_K*([9]).

Proof. (a) follows from the definition of the space H_K.

(b) Let h_1 be in H_K. By (a) there exists a sequence $\{u_n\}$ in $D(A)$ such that $|h_1 - u_n| \to 0$ as $n \to \infty$. Hence $|u_n - u_m| \to 0$ and by (1.6), valid for all u in $D(A)$, we have $\|u_n - u_m\| \to 0$ as $n \to \infty$. Thus, u_n converges to some h in H. We can thus assign to each h_1 in H_K (ideal or not) a definite h in H, the correspondence being linear and such that $h_1 = h$ for h_1 in $D(A)$. By (1.5) Ku_n converges. Since $u_n \to h$ and K is closed, $h \in D(K)$ and $Kh = \lim_{n\to\infty} Ku_n$. Because the

([8]) See, for instance, Mikhlin [16].

([9]) For $K = I$, Lemma 1.2 was first proved by Friedricks [4].

equation $[v,h_1]=(Av,Kh)$, where h denotes the element of H we have just made correspond to h_1 of H_K, is valid by definition if v and h_1 are in $D(A)$, by continuity it remains valid in case v is in $D(A)$ and h_1 is any element in H_K. It follows from this equation that the identification is one-to-one for if $h=0$, then $[v,h_1]=0$ for all v in $D(A)$ showing that $h_1=0$. Thus we can identify elements of H_K with the corresnonding elements of H.

(c) Since for all $u\in D(A)\subseteq H_K$ the operator K satisfies the inequality $\|Ku\| \leqq (1/\gamma_1)|u|$ we see that K is a bounded operator from $D(A)$, considered as a manifold in H_K, into H. Thus K can be extended by continuity to a bounded operator of all H_K to H. We shall also denote this extension by K.

(d) (1.5) and (1.6) carry over by continuity to all of u in H_K.

THEOREM 1.3. *If A is K-p.d. the problem of minimizing $F(u)$ has in H_K a unique solution for every f in H. If $\{\phi_i\}, i=1,2,\ldots,$ is a complete orthonormal sequence in H_K, then the element w realizing the minimum of $F(u)$ can be developed in the series.*

$$w=\sum_{i=1}^{\infty}(f,K\phi_i)\phi_i \tag{1.7}$$

converging in H_K (and in H)[10].

Proof. If f is a fixed element in H and u is any element in H_K, then by Lemma 1.2 (c) (f,Ku) is a bounded conjugate linear functional of u in H_K. By Fréchet-Riesz theorem there exists a unique element w in H_K such that for all u in H_K

$$(f,Ku)=[w,u]. \tag{1.8}$$

Consequently, $F(u)$ in (1.4) can be written in the form

$$F(u)=|u|^2-[u,w]-[w,u] \tag{1.9}$$

valid for all u in $D(A)$. But its right-hand side has meaning for all u in H_K. Therefore, we may use (1.9) to extend $F(u)$ to all of H_K and seek the minimum of $F(u)$ in H_K. The last problem is solved very simply. Putting (1.9) in the form $F(u)=|u-w|^2-|w|^2$ we see that $F(u)$ attains its minimum at $u=w$ with $\min F(u)=F(w)=-|w|^2$. This proves the first part of the theorem.

To prove the second part observe that since $w\in H_K$ and $\{\phi_i\}$ is a complete orthonormal sequence in H_K the element w can be expanded into the series

$$w=\sum_{i=1}^{\infty}[w_1\phi_i]\phi_i \tag{1.10}$$

converging in H_K. Using (1.8) and replacing in it u by ϕ_i we can put (1.10) as

$$w=\sum_{i=1}^{\infty}(f,K\phi_i)\phi_i. \tag{1.11}$$

---

[10] In case A is of the form $A=CK$, where C is self-adjoint, Theorem 1.3 was stated without proof by Martyniuk [15].

The series (1.11) converges in the metric of H_K and, in view of (1.6), also in the metric of H. This proves our theorem.

REMARK 1. The formula (1.11) can be used in actual computation of the solution of the minimum problem (1.4). However, in practice it is difficult to obtain an orthonormal sequence $\{\phi_i\}$ which is complete in H_K. We shall consider this question in Lemma 1.4 below.

SPECIAL CASE. If we choose $K = I$, then A is positive definite and our theorem reduces to an analogous theorem for such operators [16].

1.3. **Solvable generalized Friedrichs extensions.** Let us note that w may not be in $D(A)$ so that eq. (1.1) may not have a solution if A is considered only on $D(A)$. In this section we show that a K-p.d. operator A can be extended to a closed K-p.d. operator whose domain consists of all elements realizing the minimum of $F(u)$ in H_K and whose range is H and which has a bounded inverse.

THEOREM 1.4. *If A is K-p.d., then A can be extended to a closed K-p.d. operator A_1 such that $A_1 \supseteq A$ and A_1 has a bounded inverse A_1^{-1} on $R(A_1) = H$.*

Proof. By Lemma 1.2 (c) the product (f, Kv) is a bounded conjugate linear functional of v in H_K for each f in H. Hence there exists a linear bounded operator G defined on all of H into H_K such that

$$(f, Kv) = [Gf, v] \tag{1.12}$$

for all v in H_K and f in H and $|G| \leqq 1/\gamma_1$ since by (1.12) and (1.5)

$$|G| = \sup_{\|f\|=1,\,|v|=1} |[Gf, v]| = \sup_{\|f\|=1,\,|v|=1} |(f, Kv)| \leqq \sup_{|v|=1} \|Kv\| \leqq \frac{1}{\gamma_1} \sup_{|v|=1} |v| = \frac{1}{\gamma_1}.$$

Let us add that G is also bounded as an operator from H to H and, as is easily seen from (1.5) and (1.6), $\|G\| \leqq 1/\alpha\beta$.

Also, if $Gf = 0$, then, by (1.12), $(f, Kv) = 0$ for all v in H_K. Since K is continuously $D(A)$-invertible this implies that $f = 0$. Thus, G has an inverse $A_1 = G^{-1}$ which is closed.

Furthermore, $D(A_1)$ is dense in H_K in the sense of both metrics for otherwise there would exist $h \neq 0$ in H_K such that, by (1.12), $(A_1 u, Kh) = [GA_1 u, h] = [u, h] = 0$ for all u in $D(A_1)$. Since $R(A_1) = H, Kh = 0$ implying that $h = 0$. Thus, $D(A_1)$ is dense in the H_K-metric and, in view of (1.6), also in the H-metric.

Finally we prove that A_1 is K-p.d. and $A_1 \supseteq A$. Indeed, if $w = Gf$ is in $D(A_1)$, then by (1.12) and (1.5)

$$(A_1 w, Kw) = [w, w] \geqq \alpha \|Kw\|^2$$

showing that A_1 is K-p.d. To prove that $A_1 \supseteq A$ note that by the definition of the H_K-metric and (1.12)

$$[u, v] = (Au, Kv) = [GAu, v]$$

for all pairs u, v in $D(A)$. Since $D(A)$ is dense in H_K, the last equality implies that $GAu = u$ for all u in $D(A)$. This shows that $u \in D(A_1)$ and $A_1 u = Au$, i.e., $A_1 \supseteq A$ and thus completes the proof.

The operator A_1 will be called a *solvable generalized Friedrichs extension of A*. For $K = I$, Theorem 1.4 furnishes a self-adjoint extension of a symmetric positive definite operator constructed by Friedrichs([11]), whose procedure we have generalized to the nonsymmetric K-p.d. operators.

REMARK 2. A may still have other K-p.d. extensions. But among these extensions there is only one, the operator A_1 we have just constructed, whose domain is contained in H_K. Also, if A^1 is an arbitrary K-p.d. extension of A such that $D(A^1) \subset H_K$, then $A_1 \supseteq A^1$. To prove it, let v^1 be an element in $D(A^1)$ and u an element in $D(A)$; then, by (1.12), $[u, GA^1v^1] = (Ku, A^1v^1) = (A^1u, Kv^1) = (Au, Kv^1)$ and, since $(Au, Kv) = [u, v]$ is valid for all $u \in D(A)$ and v in H_K, we have $[u, GA^1v^1] = [u, v^1]$. This implies that $v^1 = GA^1v^1$. Thus, $v^1 \in D(A_1)$ and $A_1v^1 = A^1v^1$, i.e., $A_1 \supseteq A^1$.

We complete this section by proving three theorems on Friedrichs extensions which seem to be both of theoretical and practical interest. In Theorem 1.5 below we construct a solvable generalized Friedrichs extension for a more general class of nonsymmetric operators. The other two theorems are concerned with the extensions of the perturbed K-p.d. operators. In case H is real and $K = I$ the latter two theorems were recently proved by Lax and Milgram [14].

THEOREM 1.5. *Let L be a linear operator defined over $D(A)$ and η_1 and η_2 two positive constants such that for all u and v in $D(A)$*

$$\eta_1 |(Au, Kv)| \leqq |(Lu, Kv)| \leqq \eta_2 |(Au, Kv)|. \tag{1.13}$$

Then L has a generalized solvable Friedrichs extension L_1 such that L_1 is a closed operator from $D(L_1)$ onto H, $L \subseteq L_1$, and L_1 has a bounded inverse on H.

Proof. By Lemma 1.2 (c) (Lu, Kv) is a bounded conjugate linear functional of v in H_K for each u in $D(A)$. Hence there exists a linear operator S in H_K with domain $D(S) = D(A)$, the latter considered as a manifold in H_K, such that for all v in H_K and u in $D(A)$

$$(Lu, Kv) = [Su, v]. \tag{1.14}$$

By the right-hand inequality (1.13) and Lemma 1.1 (c), $|[Su, v]| \leqq \eta_2 |u| \, |v|$ for all $u \in D(S)$ and $v \in H_K$. This shows that S is bounded. By the left-hand inequality (1.13), $|[Su, u]| \geqq \eta_1 [u, u]$. Hence S has also a bounded inverse S^{-1}. Therefore, S has a closure $\bar{S}$ in H_K such that $R(\bar{S})$ is closed, $R(\bar{S}) = \overline{R(S)}$, $\bar{S}$ is bounded in H_K and has a bounded inverse $\bar{S}^{-1}$ on $R(\bar{S})$. Moreover, if v is a given element in H_K and $[Su, v] = 0$ for all u in $D(A)$, then, in view of (1.14) and (1.13),

([11]) See, for instance, Friedrichs [4], Mikhlin [16], and Riesz and Sz.-Nagy [21].

$[u,v] = 0$ for all $u \in D(A)$. Since $D(A)$ is dense in H_K, $v = 0$. Thus, $R(S)$ is dense and $R(\bar{S}) = H_K$.

To construct L_1 we define the generalized solution of the equation

$$Lu = f \tag{1.15}$$

for a given f in H to be an element u in $D(\bar{S})$ such that for all v in H_K

$$(f,Kv) = [\bar{S}u, v]. \tag{1.16}$$

Note that every ordinary solution of eq. (1.15) is also its generalized solution, but the converse is not necessarily true. Moreover, for any f in H there exists a unique generalized solution of eq. (1.15). To show this note that for each f in H, (f, Kv) is a bounded conjugate linear functional of v in H_K. Hence there exists a bounded linear operator G of H into H_K such that for all v in H_K

$$(f,Kv) = [Gf,v], \quad |G| \leqq \frac{1}{\gamma_1}. \tag{1.17}$$

Clearly G^{-1} exists. From (1.16) and (1.17) we get the following relation

$$[\bar{S}u,v] = [Gf,v], \quad v \in H_K. \tag{1.18}$$

This implies that the seeking solution of (1.15) must satisfy the equation

$$\bar{S}u = Gf, \tag{1.19}$$

whence,

$$u = \bar{S}^{-1}Gf, \quad G^{-1}\bar{S}u = f. \tag{1.20}$$

It is evident that u determined by (1.20) is the unique generalized solution of eq. (1.15). The operator $L_1 = G^{-1}\bar{S}$ is well defined and is an extension of L since for any u in $D(A)$ and all v in H_K

$$(Lu,Kv) = [\bar{S}u,v] = [GLu,v].$$

This shows that $\bar{S}u = GLu$, i.e., $\bar{S}u$ belongs to $D(G^{-1})$ and $L_1u = G^{-1}\bar{S}u = Lu$. Thus, $D(L_1) \supseteq D(L)$. The inverse of L_1 is $L_1^{-1} = \bar{S}^{-1}G$. Since G is everywhere defined and continuous as a mapping from H to H_K and $\bar{S}^{-1}$ is everywhere defined and continuous in H_K, L_1^{-1} is everywhere defined and continuous as a mapping from H to H_K and, from (1.6), as a mapping from H into H. Hence L_1 is closed and $R(L_1) = H$, as was to be shown.

REMARK 3. Theorem 1.5 remains valid if instead of (1.13) we assume that (Lu,Ku) is real and L satisfies the left-hand inequality in (1.13) for all u and v in $D(A)$. In fact, if these conditions are satisfied, then S determined by (1.14), though no longer bounded in H_K, still has a bounded inverse S^{-1} on $R(S)$ which is dense in H_K. Moreover, S is closeable, i.e., if $\{u_n\}$ is a sequence

in $D(S)$ and w an element in H_K such that $|u_n - 0| \to 0$ and $|Su_n - w| \to 0$ as $n \to \infty$, then $w = 0$. To see this note that by (1.5) and (1.6), $Ku_n \to 0$ and if v is any element in $D(A)$, then, since (Lu,Ku) is real for all u in $D(A)$, the relation (1.14) and Lemma 1.1(b) imply that

$$[w,v] = \lim_n [Su_n,v] = \lim_n (Lu_n,Kv) = \lim_n (Ku_n,Lv) = 0.$$

Since $D(A)$ is dense in H_K, $w = 0$. Thus, S is closeable and has a bounded inverse on $R(S)$. Hence S has a closure $\bar{S}$ in H_K, $R(\bar{S}) = H_K$, and $\bar{S}^{-1}$ is bounded on H_K. The rest of the proof is the same as above.

THEOREM 1.6. *If A is K-p.d. and $L = A + B$, where B is a linear operator over $D(A)$ such that for all u in $D(A)$*

(a) $(Lu,Ku) \geqq \eta_3(Au,Ku)$, $\eta_3 > 0$,

(b) $\|Bu\|^2 \leqq \eta_4(Au,Ku)$, $\eta_4 > 0$,

then $D(L_1) = D(A_1)$ and $L_1 = A_1 + \bar{B}$, where $\bar{B}$ is the extension of B to $D(A_1)$.

Proof. (a) and (b) imply that L is K-p.d. and that the norm induced by $[u,u]' = (Lu,Ku)$ is equivalent to the norm $[u,u] = (Au,Ku)$ for all u in $D(A)$. Therefore, H_K' associated with L is the same as H_K associated with A. Furthermore, by Theorem 1.4, L has a solvable generalized Friedrichs extension L_1'. (b) implies that B can be extended to all of H_K as a bounded operator $\bar{B}$ from H_K to H.

To show that $L_1' = A_1 + \bar{B}$ note that since $H_K' = H_K$ the Remark 2 implies that $L_1' = L_1 = G^{-1}\bar{S}$, where L_1 is the solvable generalized Friedrichs extension of L constructed by Theorem 1.5 and Remark 3. Furthermore, since $D(A)$ is dense in H_K the linear operator $\bar{S}$ in H_K has the property that for all $u \in D(L_1)$ and $v \in H_K$

$$[u,v]' = [u,v] + (\bar{B}u,Ku) = (L_1u,Kv)$$

$$= [GL_1u,v] = (G \cdot G^{-1}\bar{S}u,v] = [\bar{S}u,v],$$

$\bar{S}$ has a bounded inverse $\bar{S}^{-1}$ on $R(\bar{S})$, and $R(\bar{S}) = H_K$. The only statement that is not immediate is that $R(\bar{S}) = H_K$ or equivalently that if $v \in H_K$ and $[Su,v] = 0$ for all u in $D(S)$, then $v = 0$. This follows, however, from (1.14) and the equivalence of H_K- and H_K'-norms. Indeed, if v is in H_K, then for all u in $D(A)$ we have

$$[u,v]' = (Lu,Kv) = [Su,v] = 0$$

and since $D(A)$ is dense in $H_K' = H_K$, we must have $v = 0$.

Now if $f \in H$, $L_1^{-1}f$ satisfies the following equality for all v in H_K

$$[L_1^{-1}f,v] + (\bar{B}L_1^{-1}f,Kv) = [\bar{S}L_1^{-1}f,v] = [\bar{S} \cdot \bar{S}^{-1}Gf,v] = [Gf,v] = (f,Kv).$$

Since A_1 is a solvable generalized Friedrichs extension, then for all v in H_K

$$[L_1^{-1}f,v] = (f - \bar{B}L_1^{-1}f,Kv) = [A_1^{-1}(f - \bar{B}L_1^{-1}f),v].$$

The last equality, being valid for all v in H_K, implies that

$$L_1^{-1} f = A_1^{-1}(f - \bar{B}L_1^{-1}f).$$

This shows that $L_1^{-1}f \in D(A_1)$, i.e., $D(L_1) \subset D(A_1)$ and that

$$L_1 u = A_1 u + \bar{B}u$$

holds for all u in $D(L_1)$, where we put $u = L_1^{-1}f$. Interchanging the roles of A and L we get $D(A_1) \subset D(L_1)$. Consequently $D(A_1) = D(L_1)$ and $L_1 = A_1 + \bar{B}$ on $D(A_1)$, as was to be shown.

THEOREM 1.7. *If A is K-p.d. and $L = A + B$, where B is a linear operator over $D(A)$ such that for all u in $D(A)$*

(a) $(Lu,Ku) \geqq \eta_3(Au,Ku)$, $\eta_3 > 0$,

(c) $|(Bu,Kv)| \leqq \eta_5 |u|\,|v|$, $\eta_5 > 0$,

then

$$L_1 = A_1(I + T), \tag{1.21}$$

where T is the extension of the operator $A_1^{-1}B$ to H_K.

Proof. In proving Theorem 1.7 we shall follow the procedure of [14]. (a) and (c) assures us that the spaces H_K associated with the norms $[u,u]'$ and $[u,u]$ are the same and that L_1 exists; but $D(L_1)$ and $D(A_1)$ need not be the same. Also, the operator $A_1^{-1}B$ defined over $D(A)$ is bounded in the H_K-norm. In fact, if we put $v = A_1^{-1}Bu$, then by (c)

$$|v|^2 = |A_1^{-1}Bu|^2 = (Bu,Kv) \leqq \eta_5 |u| \cdot |v|.$$

This shows that $A_1^{-1}B$ is bounded. Let T denote its extension to H_K.

To show (1.21) observe that if u is an element in H_K then there exists a sequence $\{u_i\}$ in $D(A)$ such that for all w in H_K

$$[u,w]' = \lim_i [u_i,w]' = \lim_i (Lu_i, Kw).$$

Since A_1 is a solvable generalized Friedrichs extension of A,

$$(Lu_i,Kw) = (A_1 u_i,Kw) + (A_1 A_1^{-1} Bu_i,Kw) = [(I + T)u_i,w].$$

Taking the limit as $i \to \infty$ and using the fact that $(I + T)$ is bounded in H_K, we get $[u,w]' = [(I + T)u,w]$. With u in $D(L_1)$, $(L_1 u,Kw) = [u,w]'$; whence, by definition of A_1, $(I + T)u$ belongs to $D(A_1)$, i.e., $D(L_1) \subset D(A_1(I + T))$, and $L_1 u = A_1(I + T)u$. This is precisely (1.21). With u in $D(A_1(I + T))$, $[u,w]' = [(I + T)u,w] = (A_1(I + T)u,Kw)$, for all w in H_K, whence, by definition of L_1, $u \in D(L_1)$; i.e., $D(A_1(I + T)) \subset D(L_1)$, as was to be shown.

REMARK 4. In applications it is L_1^{-1} rather than L_1 that we want to compute since L_1^{-1} relates the solution of the problem to the given data. Therein lies the usefulness of (1.21) for, if we can choose A so that A_1^{-1} and $(I + T)^{-1}$ can be relat-

ively easily computed, then L_1^{-1} (being equal to $(I+T)^{-1}A_1^{-1}$) can also be computed. The generalized variational principle enables us to compute A_1^{-1} and thus $T = A_1^{-1}B$.

REMARK 5. In Chapters II and III whenever we consider the unbounded K-p.d. operators it will be assumed that these operators have been already extended and consequently by the solution of a given equation we shall mean the solution of the generalized equation.

1.4. **Additional lemmas.** In this section we are concerned with further properties of K-p.d. operators which we shall use in Chapters II and III.

LEMMA 1.3. *If A is K-p.d. and $D(A) = D(K)$, then AK^{-1} is bounded and*

$$\theta^2 |u|_1^2 \geqq \|Au\|^2 \geqq \alpha |u|^2 \geqq \alpha^2 |u|_1^2, \quad u \in D(A), \tag{1.22}$$

where $|u|_1 = \|Ku\|$ is the norm in H_1, i.e., $H_K = H_1 = D(K)$.

Proof. Since $D(A) = D(K)$ the inequality (1.3) implies that $\|AK^{-1}u\| \leqq \theta \|KK^{-1}u\| = \theta \|u\|$ for all u in H, showing that AK^{-1} is bounded.

To prove (1.22) note that from (1.2) and Schwarz inequality we get $\|Au\| \geqq \alpha \|Ku\|$ and $\|Au\|^2 \geqq \alpha |u|^2 \geqq \alpha^2 \|Ku\|^2$. The last inequality and (1.3) give us (1.22).

In the lemma below we consider the question of completeness of a sequence of linearly independent elements $\{\phi_i\}$ in H_K raised in Remark 1.

LEMMA 1.4. *If A is K-p.d. and $D(A) = D(K)$, then the following statements are equivalent*

(a) *The sequence $\{\phi_i\}$, $\phi_i \in D(A)$, is complete in H_K.*

(b) *The sequence $\{K\phi_i\}$ is complete in H.*

(c) *The sequence $\{A\phi_i\}$ is complete in H.*

Proof. (a) $\rightarrow$ (b). Assuming (a) it is sufficient to show that for any v in H, $(v, K\phi_i) = 0$, $i = 1, 2, \ldots$, implies $v = 0$. Let $g \in H$ and w be the solution of $Kw = g$. Then, w being in $D(K) = H_K$, can be approximated arbitrarily closely by linear combinations of the form $w_n = \sum_{i=1}^{m} c_{ni}\phi_i$ in the H_K-norm. (1.5) and (1.6) imply that $\{w_n\}$ and $\{Kw_n\}$ converge in H to w and Kw respectively. But for any w_n we have $(V, Kw_n) = 0$. Passing to the limit, as $n \rightarrow \infty$, we get $(v, Kw) = (v, g) = 0$ for any g in H. This shows that $v = 0$.

(b) $\rightarrow$ (c). If $g \in H$, then Theorem 1.1 and our hypothesis imply that $KA^{-1}g$ exists and belongs to H. By (b) there exists n and numbers $\alpha_1, \alpha_2, \ldots, \alpha_n$ such that $\|KA^{-1}g - Kv_n\| < \varepsilon / \|AK^{-1}\|$, where $\varepsilon > 0$, $v_n = \sum_{i=1}^{n} \alpha_i \phi_i$ and AK^{-1} is bounded by Lemma 1.3. Thus,

$$\|Av_n - g\| = \|AK^{-1}Kv_n - AK^{-1}KA^{-1}g\| \leqq \|AK^{-1}\| \, \|Kv_n - KA^{-1}g\| < \varepsilon.$$

(c) $\rightarrow$ (a). Let $w \in H_K$. By Lemma 1.3 $w \in D(K)$ and hence by (c) there exist n

and $\alpha_1, \ldots, \alpha_n$ such that $\| Aw - \sum_{i=1}^{n} \alpha_i A\phi_i \| < \gamma_1 \varepsilon$. This and (1.22) imply that

$$\left| w - \sum_{i=1}^{n} \alpha_i \phi_i \right| \leqq \frac{1}{\gamma_1} \left\| Aw - \sum_{i=1}^{n} \alpha_i A\phi_i \right\| < \varepsilon.$$

This shows that $\{\phi_i\}$ is complete in H_K and completes the proof of Lemma 1.4.

REMARK 6. Due to a wide freedom in the choice of K, Lemma 1.4 makes it, in general, easier to examine the conditions under which $\{\phi_i\}$ is complete in H_K or $\{A\phi_i\}$ is complete in H.

In Chapter III we make use of the following result. Let K be a bounded linear and positive definite operator in H, i.e., there exists a constant $\beta_1 > 0$ such that

$$(Ku,u) \geqq \beta_1 \| u \|^2, \qquad u \in H. \tag{1.23}$$

Introduce in H a new scalar product and the associated norm by

$$[u,v]' = (Ku,v), \qquad |u|' = \{[u,u]'\}^{1/2}, \qquad u,v \in H. \tag{1.24}$$

Because of (1.23) H becomes in this metric a new Hilbert space denoted by H'.

LEMMA 1.5. *If A is bounded and K-p.d. and K satisfies* (1.23), *then*

(a) *$R(A)=H$, A^{-1} exists and is bounded, and the equation $Au=f$ has a unique solution w for every f in H.*

(b) *A and A^{-1} are symmetrizable by K.*

(c) *Considered as operators in H', A and A^{-1} are symmetric and bounded, $|A|' \leqq \| A \|$, and $|A^{-1}|' \leqq 1/\alpha\beta_1$, where $|A|'$ and $|A^{-1}|'$ denote the H'-norm of A and A^{-1} respectively.*

Proof. (a) follows from Lemma 1.1 and Theorem 1.1.

(b) By Lemma 1.1 (b) $(KAu,v) = (Au,Ku) = (Ku,Av) = (u,KAv)$ for all u and v in H. This shows that A is symmetrizable by K. By (a) and the first part of (b)

$$(A^{-1}u,Kv) = (A^{-1}u,KA(KA)^{-1}Kv) = (Ku,A^{-1}v)$$

which shows that A^{-1} is also symmetrizable by K.

(c) The symmetry of A and A^{-1} as operators in H' follows from (b). Using the inequality of Reid([12]) and considering A as an operator in H' we get

$$[Au,u]' = (KAu,u) \leqq \| A \| (Ku,u) = \| A \| \cdot |u|'^2, \qquad u \in H'. \tag{1.25}$$

This shows that $|A|' \leqq \| A \|$. From (1.23) and the Schwarz inequality $\| Ku \| \geqq \beta_1 \| u \|$, whence $\beta_1(Ku,u) \leqq \| Ku \|^2$. This and (1.2), considered for our present case, imply that

$$[Au,u]' = (Au,Ku) \geqq \alpha\beta_1 |u|'^2, \qquad u \in H',$$

which shows that A^{-1}, as an operator in H', is symmetric and $|A|' \leqq 1/\alpha\beta_1$.

([12]) It was shown by Reid [20] that if A is symmetrizable by a positive operator K, then $|(KAu, u)| \leqq \| A \| (Ku, u)$.

II. Direct methods

2.1. **Generalized Ritz Method.** In Theorem 1.3 we have constructed the exact solution of the minimum problem (1.4) which, however, is not very useful for practical computation. Instead, using an approach based on the variational principle, the following procedure can be used to construct an approximate solution w_n of eq. (1.1):

If $\{\phi_i\}$, $i=1,2\ldots$, is a sequence of linearly independent elements in $D(A)$ which is complete in H_K and w_n is taken in the form $w_n = \sum_{i=1}^{n} a_i^n \phi_i$, then the coefficients $a_1^n, \ldots, a_n^n$ are determined from the condition of the minimum of the function $G(a_1^n, \ldots, a_n^n) = F(w_n)$. This condition yields the following linear algebraic system

$$\sum_{i=1}^{n} (A\phi_i, K\phi_j) a_i^n = (f, K\phi_j), \qquad j = 1,2,\ldots,n. \tag{2.1}$$

We shall call this procedure the *generalized Ritz method.* Its justification is based on Theorems 1.2, 1.3, and the lemmas:

Lemma 2.1. *Every minimizing sequence of $F(u)$ converges in H_K (and in H) to an element realizing the minimum of $F(u)$.*

Lemma 2.2. *The approximation w_n constructed by the generalized Ritz method forms a minimizing sequence of $F(u)$.*

We omit the proofs of these lemmas since by a geometrical approach used below we will obtain an immediate proof of the convergence of the method and also the convergence of Aw_n to f.

Geometrical approach. In our investigation of the generalized Ritz method we do not use the variational principle. Our argument is of a purely geometrical nature. Let us recall that by Lemma 1.2(c) the product (f,Ku) is a bounded conjugate linear functional of u in H_K for each fixed f in H. Hence by the theorem of Fréchet-Riesz there exists a unique element w in H_K such that

$$(f,Ku) = [w,u] \tag{2.2}$$

for all u in H_K. By the Theorem 1.4 the element w belongs to $D(A)$, where A is a solvable generalized Friedrichs extension, and $Aw = f$.

To construct an approximate solution of eq. (1.1) we take a system of linearly independent elements $\phi_i \in D(A)$ which is complete in H_K. Let H_n denote the linear space spanned by $\phi_1, \ldots, \phi_n$ and P_n the orthogonal projection on H_n in the space H_K. We take the approximate solution w_n in the form

$$w_n = \sum_{i=1}^{n} a_i^n \phi_i \tag{2.3}$$

and determine the coefficients $a_1^n, \ldots, a_n^n$ from the condition that

$$w_n = P_n w \tag{2.4}$$

which, because $w - P_n w$ and $\phi_1, \ldots, \phi_n$ are orthogonal in the H_K-metric, leads to the algebraic system of linear equations

$$\sum_{i=1}^{n} (A\phi_i, K\phi_j) a_i^n = (f, K\phi_j), \qquad j = 1, 2, \ldots, n. \tag{2.5}$$

Since the systems (2.1) and (2.5) are identical, w_n is Ritz's approximate solution and for that reason we call the present procedure the generalized Ritz method. Note that because ϕ_i are linearly independent in H_K the determinant $|a_{ij}| = |[\phi_i, \phi_j]| = |(A\phi_i, K\phi_j)| \neq 0$. Hence the system $a_1^n, \ldots, a_n^n$ and consequently w_n is uniquely determined for each n and a given f. It is clear from the way w_n is constructed that $|w_n - w| \to 0$ and because of (1.6) also $\|w_n - w\| \to 0$ as $n \to \infty$. This shows the convergence of the method.

If in addition to the condition (1.2) we also assume that $D(K) = D(A)$, then $\|Aw_n - f\| \to 0$ as $n \to \infty$. In fact, the inequality (1.5) and the closedness of K imply that $Kw_n \to Kw$ in the norm of H. This and the inequality (1.3) imply that

$$\|Aw_n - Aw\| \leqq \theta \|Kw_n - Kw\| \to 0 \text{ as } n \to \infty$$

and thus completes the proof of our theorem.

The above discussion may be summarized in the following

THEOREM 2.1. *If A is K-p.d. and $\{\phi_i\}$, $\phi_i \in D(A)$, is complete in H_K, then*

(a) *w_n is uniquely determined for each n by the generalized Ritz method* (2.3)-(2.5).

(b) *w_n converges in H_K (and in H) to the exact solution w of eq.* (1.1).

(c) *Aw_n converges to f in H provided $D(K) = D(A)$.*

SPECIAL CASES[13]. (i) If $K = A$, then the operator A is simply invertible and the generalized Ritz method reduces in this case to the well known method of *least squares*. Theorem 2.1 gives the convergence: $w_n \to w$ and $Aw_n \to f$, as $n \to \infty$, and the geometrical basis for this method.

(ii) If $K = I$, then A is positive definite and the generalized Ritz method reduces in this case to an *ordinary Ritz method* for self-adjoint positive definite operators A. Theorem 2.1 (b) gives the convergence of this method and its geometrical meaning. It is known [16] that in this case $Aw_n \nrightarrow f$ if A is unbounded. However, it has been proved recently by Mikhlin [17] that if A and some other operator B are self-adjoint and positive definite and such that $D(A) = D(B)$, $D(A^{1/2}) = D(B^{1/2})$, the spectrum of B is discrete, and $\{\phi_i\}$ is a complete orthonormal basis of its eigenvectors, then $Aw_n \to f$. Thus, we see that our Theorem 2.1 (c) gives a similar but a more general result for the generalized Ritz method.

REMARK 7. In the next chapter we shall present a simple iterative procedure

---

[13] For the discussion and extensive literature on the method of least squares and Ritz see [8; 11; 12; 16].

and a compact computational scheme for obtaining the approximate solution w_n without solving (2.5).

2.2. **The generalized method of moments.** In this section we shall consider the problem of approximate solution of a general linear equation of the form

$$Lu = Au + Bu = f, \quad f \in H, \tag{2.6}$$

where A is K-p.d., B is a linear unbounded operator such that $D(B) \supseteq D(A)$.

The essence of the generalized method of moments consists in the following: We choose a system of linearly independent elements $\phi_i \in D(A)$ which is complete in H_K and construct another system $K\phi_i$, $i = 1, 2, \ldots, n$. The approximate solution w_n of eq. (2.6) is taken in the form

$$w_n = \sum_{i=1}^{n} a_i^n \phi_i, \tag{2.7}$$

where the constants $a_1^n, \ldots, a_n^n$ are determined from the condition that $Lw_n - f$ be orthogonal to $K\phi_i$ for $i = 1, 2, \ldots, n$. This condition gives the algebraic system of linear equations for the determination of $a_1^n, \ldots, a_n^n$

$$\sum_{i=1}^{n} \{(A\phi_i, K\phi_j) + (B\phi_i, K\phi_j)\} a_i^n = (f, K\phi_j), \qquad j = 1, 2, \ldots, n. \tag{2.8}$$

The following theorem gives the justification of this method:

THEOREM 2.2. *If eq.* (2.6) *has at most one solution and the operator* $T = A^{-1}B$ *can be extended to a completely continuous operator in* H_K, *then*

(a) *There exists an integer* n_0 *such that for all* $n \geqq n_0$ *the system* (2.8) *has a unique solution* $\{a_1^n, \ldots, a_n^n\}$.

(b) w_n *determined by the generalized method of moments converges in* H_K (*and in* H) *to the exact ordinary or generalized solution of eq.* (2.6).

(c) *If* L *has a closed extension and* $D(L) = D(K)$, *then* $Lw_n \to f$ *as* $n \to \infty$.

Proof. Applying A^{-1} to both sides of eq. (2.6) we get an equation

$$u + Tu = f_1, \; f_1 = A^{-1}f \tag{2.6$_1$}$$

to which we can apply the Fredholm's alternative. (2.6$_1$) is uniquely solvable since otherwise the equation $u + Tu = 0$ would have a nonzero solution which is impossible in view of our hypothesis. This implies that for a given f the eq. (2.6) has also a unique ordinary or *generalized* solution. The latter is defined as an element u in H_K which satisfies eq. (2.6$_1$). Since f_1 belongs to H_K and T is completely continuous in H_K then the solution w of eq. (2.6$_1$) belongs to H_K. If $\{\tilde{\phi}_i\}$ denote the system obtained from the orthonormalization of $\{\phi_i\}$ in the H_K-metric, then w has representation

$$w = \sum_{i=1}^{\infty} c_i \tilde{\phi}_i, \quad c_i = [w, \tilde{\phi}_i], \tag{2.9}$$

converging in H_K whose coefficients c_i satisfy the algebraic system

$$0 = [w + Tw, \tilde{\phi}_j] - (f, K\tilde{\phi}_j) = c_j + \sum_{i=1}^{\infty} [T\tilde{\phi}_i, \tilde{\phi}_j]\, c_i - (f, K\tilde{\phi}_j), \text{ i.e.,}$$

the system

$$c_j + \sum_{i=1}^{\infty} c_i \gamma_{ij} = b_j, \tag{2.10}$$

where $\gamma_{ij} = [T\tilde{\phi}_i, \tilde{\phi}_j]$ and $b_j = (f, K\tilde{\phi}_j)$. The complete continuity of T in H_K justifies the derivation of (2.10) and guarantees the existence of an integer n_0 such that each n-segment of the system (2.10)

$$c_j^n + \sum_{i=1}^{n} c_i^n \gamma_{ij} = b_j, \qquad j = 1, 2, \ldots, n \tag{2.11}$$

is uniquely solvable for every $n \geqq n_0$ and its solution $\{c_1^n, \ldots, c_n^n\}$ converges in the Hilbert space l_2 to the solution $\{c_1, c_2, \ldots\}$ of (2.10) ([14]). This implies that $w^{(n)} = \sum_{i=1}^{n} c_i^n \tilde{\phi}_i$ converges to w in the H_K-norm. Furthermore $w^{(n)}$ coincides with w_n, where w_n is determined by the generalized method of moments. In fact, since the sequences $\{\phi_i\}$ and $\{\tilde{\phi}_i\}, \{K\phi_i\}$ and $\{K\tilde{\phi}_i\}$ are connected by a nondegenerate matrix transformation $\tilde{\phi}_i = \sum_{j=1}^{i} \beta_{ij}\phi_j, K\tilde{\phi}_i = \sum_{j=1}^{i} \beta_{ij} K\phi_j, \beta_{ii} \neq 0$, it is easy to see that taking w_n in the form (2.7) or in the form $w_n = \sum_{i=1}^{n} d_i^n \tilde{\phi}_i$ leads to one and the same result. However, if we take w_n in the latter form, the system (2.8) is then replaced by its equivalent

$$d_j^n + \sum_{i=1}^{n} d_i^n \gamma_{ij} = b_j, \quad j = 1, \ldots, n, \tag{2.8$_1$}$$

which, in its turn, is identical with the system (2.11). Hence $w_n = w^{(n)}$. This proves parts (a) and (b) of our theorem.

To prove (c) note that since L admits a closed extension and $D(L) = D(K)$, one can easily show that LK^{-1} is bounded in H. Furthermore, (1.5) and (1.6) and the closedness of K imply that $w_n \to w$ and $Kw_n \to Kw$ and consequently

$$\begin{aligned} \| Lw_n - f \| &= \| L(w_n - w) \| = \| LK^{-1} \cdot K(w_n - w) \| \\ &\leqq \| LK^{-1} \| \cdot \| Kw_n - Kw \| \to 0 \end{aligned}$$

as $n \to \infty$. This completes the proof of our theorem.

REMARK 8. Suppose B has also the property that for all $u, v \in D(L)$

$$|(Bu, Kv)| \leqq \eta_6 |(Au, Kv)|, \quad 0 < \eta_6 < 1. \tag{2.12}$$

Then L satisfies the condition (1.13) of Theorem 1.5 with $\eta_1 = 1 - \eta_6$ and hence L possesses a solvable generalized Friedrichs extension which we shall also denote by L. Furthermore, the inequalities (1.5), (1.6), and (2.12) imply that for each f in H

([14]) The details of the proof of the last statements may be found in the treatise of Mikhlin [16].

$$\|L^{-1}f\| \leqq \frac{\|f\|}{(1-\eta_6)\alpha\beta}. \tag{2.13}$$

If $T = A^{-1}B$ can be extended to a completely continuous operator in H_K, $D(L) = D(K)$, w is the exact solution of eq. (2.6), and w_n the approximate solution determined by the generalized method of moments, then from Theorem 2.2 and inequality of (2.13) we obtain the convergence $w_n \to w$, $Lw_n \to f$, and the following *estimate of error*

$$\|w_n - w\| \leqq \frac{1}{(1-\eta_6)\alpha\beta} \|Lw_n - f\| \tag{2.14}$$

which can be made as small as we please.

2.3. **Special cases.** In this section we shall show that the most important direct methods used in the approximate solution of linear operator equations can be deduced from the generalized method of moments given above.

(i) GALERKIN METHOD. If $K = I$, then (1.2) shows that A is a self-adjoint positive definite operator on $D(A), H_K$ is the space H_I with the inner product $[u,v] = (Au,v)$, and the system (2.8) reduces to the system of the well-known Galerkin equations

$$\sum_{i=1}^{n} \{(A\phi_i, \phi_j) + (B\phi_i, \phi_j)\} a_i^n = (f, \phi_j), \quad j = 1,2,\ldots,n.$$

Thus, in this case the generalized method of moments reduces to the Galerkin method which has been extensively studied and used in applications [7;9;13;16].

(ii) THE METHOD OF MOMENTS. If $K = A$, then A is invertible, H_K is the space H_A with the inner product $[u,v] = (Au, Av)$, and (2.8) reduces to the system of the method of moments

$$\sum_{i=1}^{n} \{(A\phi_i, A\phi_j) + (B\phi_i, A\phi_j)\} a_i^n = (f, A\phi_j), \quad j = 1,2,\ldots,n.$$

Theorem 2.2 establishes in this case the convergence $w_n \to w$ and $Lw_n \to f$ of the method of moments. In case A and B are differential operators the method was investigated by Kravchuk [10] and Zdanow [22]. In spite of its generality and a great possibility for applications the method has not been thoroughly investigated from a general Hilbert space point of view.

(iii) GENERALIZED RITZ METHOD. If $B = 0$, then the generalized method of moments reduces to the generalized Ritz method investigated in §1, Chapter II, of this paper.

(iv) THE METHOD OF LEAST SQUARES. If $B = 0$ and $K = A$, then our method reduces to the method of least squares with the well known equations

$$\sum_{i=1}^{n} \{(A\phi_i, A\phi_j)\} a_i^n = (f, A\phi_j), \quad j = 1,2,\ldots,n.$$

In this case we have the convergence: $w_n \to w$, $Aw_n \to f$, and the estimate of error $\| w_n - w \| \leqq (1/\gamma) \| Aw_n - f \|$.

(v) RITZ METHOD. If $B = 0$ and $K = I$, then our method reduces to the ordinary Ritz method which has been extensively studied and used in various applications.

We shall complete this section with the following remark:

ADVANTAGES OF THE GENERALIZED METHOD OF MOMENTS. Let us first observe that due to a wide freedom in the choice of the operator K subject only to the inequality (1.2) and the generality of the operator B the generalized method of moments can be applied to a much larger class of problems than any of its special cases mentioned above. At the same time using the geometrical approach we were able to present in a unified manner methods which seemed to be different. When applied to the differential boundary-value problems the generalized method of moments will give a better character of convergence than the methods of Galerkin or Ritz. Finally, a great possibility in the variation of the operator K indicates its usefulness and convenience from a strictly computational point of view.

2.4. **The generalized Murray method.** In this section we describe a geometrical procedure which gives the necessary and sufficient conditions for solving an operator equation in H and which for a proper choice of the coordinate elements reduces in its hypothesis and conclusions to the method proposed by Professor Francis J. Murray [18]. For that reason we call it the *generalized Murray method.*

Let $\{\psi_i\}$ be a complete linearly independent set in H; consider an equation

$$Au = f, \qquad f \in H, \tag{2.15}$$

where A is assumed to be a bounded operator in H. If A^* denotes the adjoint of A, π the projection of H on $Z(A)$, and P the projection of H on $\overline{R(A^*)}$, then $H = Z(A) \oplus \overline{R(A^*)}$ and for each u in H we have $u = \pi u + Pu$, $PA^*u = A^*u$, $APu = Au$, and $\pi A^*u = 0$. Thus, if for a given f there is a solution u such that $Au = f$, then there is also a u^* in $\overline{R(A^*)}$, $u^* = Pu$, such that $Au^* = f$.

Consider the set $\{A^*\psi_i\}$ and assume that $A^*\psi_i$ are linearly independent. Observe that eq. (2.15) is valid if and only if

$$(f,\psi_i) = (u,A^*\psi_i), \qquad i = 1,2,\ldots. \tag{2.16}$$

Let H_n denote the subspace of H spanned by $A^*\psi_1,\ldots,A^*\psi_n$ and P_n the projection of H on H_n. We shall construct the approximate solution u_n in H_n to the exact solution u, which we assume to exist for a given f in $R(A)$, in the form

$$u_n = \sum_{i=1}^{n} a_i^n A^*\psi_i, \tag{2.17}$$

where $a_1^n,\ldots,a_n^n$ are determined from the condition that

$$u_n = P_n u. \tag{2.18}$$

By the equivalence of eq. (2.15) and relations (2.16) the projection $u_n = P_n u$ is defined by the linear algebraic equations

$$\sum_{i=1}^{n} (A^*\psi_i, A^*\psi_j) a_i^n = a_j, \qquad j = 1, \ldots, n, \tag{2.19}$$

where $a_j = (f, \psi_j)$. Note that since $A^*\psi_i$ are linearly independent the system (2.19) is uniquely solvable and so u_n is the unique element in H_n with the property that

$$\delta_n^2 = \min_{v \in H_n} \| u - v \|^2 = \| u - u_n \|^2 = \| u \|^2 - (u_n, u) = \| u \|^2 - \sum_{i=1}^{n} a_i^n \bar{a}_i \geqq 0 \tag{2.20}$$

with equality sign holding in (2.20) if and only if $\| u - \sum_{i=1}^{n} a_i^n A^*\psi_i \| = 0$.

It is clear from the way u_n was constructed that $u_n \to u^* = Pu$ and $Au^* = APu = Au = f$ and $\delta^2 = \| u - u^* \|^2$, where $\delta^2 = \lim_n \delta_n^2$ as $n \to \infty$. Moreover, if $u \in \overline{R(A^*)}$, then $u^* = u$, $\delta = 0$, and $\lim_n \sum_{i=1}^{n} a_i^n \bar{a}_i = \| u \|^2$.

If u is any other solution of eq. (2.15), then $A(u - u^*) = 0$ and so any other solution of eq. (2.15) is of the form $u^* + h$ with $h \in Z(A)$. This shows that of all solutions u^* has the least norm. We thus proved the following

LEMMA 2.3. *For any n the approximate solution u_n can be uniquely constructed by* (2.19). *u_n converges monotonically to the minimal solution u^* in $\overline{R(A^*)}$ of eq.* (2.15) *provided that* (2.15) *has a solution.*

We shall now prove the converse of the above lemma.

LEMMA 2.4. *Let $f \in \overline{R(A^*)}$ and consider a set of numbers $a_i = (f, \psi_i)$, $i = 1, 2, \ldots$. Determine the system of numbers $a_1^n, \ldots, a_n^n$ from the algebraic system* (2.19). *Assume that there exists a constant $M > 0$ such that for each n*

$$G_n = \sum_{i=1}^{n} a_i^n \bar{a}_i \leqq M. \tag{2.21}$$

Then there exists a u^* in $\overline{R(A^*)}$ such that the sequence $\{u_n\}$ constructed by means of the formula (2.17) *converges to u^* and $Au^* = f$.*

Proof. Since the numbers $a_1^n, \ldots, a_n^n$ and $a_1, \ldots, a_n$ satisfy the system (2.19) one easily finds that $G_n = \| u_n \|^2$ and that $\| u_m - u_n \|^2 = G_m - G_n > 0$ whenever $m > n$. This shows that the sequence $\{G_n\}$ is monotonically increasing. By (2.21) $\{G_n\}$ is bounded. Consequently $\{G_n\}$ converges. This implies that $\| u_m - u_n \| \to 0$ as $n, m \to \infty$. Since $\overline{R(A^*)}$ is closed there exists $u^* \in \overline{R(A^*)}$ such that $u_n \to u^*$ as $n \to \infty$. By the continuity of the inner product and (2.19)

$$(u^*, A^*\psi_j) = \lim_n (u_n, A^*\psi_j) = \lim_n \sum_{i=1}^{n} a_i^n (A^*\psi_i, A^*\psi_j) = a_j = (f, \psi_j)$$

implying that $Au^* = f$, as was to be proved.

The above discussion and Lemmas 2.3 and 2.4 give

THEOREM 2.3. *A necessary and sufficient condition that there exists a u such that* $Au = f$ *is that*

(a) f *belong to* $\overline{R(A)}$.

(b) *The sequence* $\{G_n\} = \{\sum_{i=1}^{n} a_i^n \bar{a}_i\}$, *where* a_i^n *and* a_i *are related by equations* (2.19), *be bounded by some constant* $M > 0$ *independent of n. If* (a) *and* (b) *are satisfied, then the approximate solution* u_n *determined by* (2.17)-(2.19) *converges monotonically to the minimal solution of eq.* (2.15).

REMARK 9. If $\{\psi_i\}$ is so chosen that $\{A^*\psi_i\}$ is orthonormal, then the system (2.19) has a simple form $a_j^n = a_j = (f, \psi_j)$ so that $u_n = \sum_{j=1}^{n} (f, \psi_j) A^*\psi_j$, i.e., u_n coincides with the nth section of the Fourier series. In this case the method (2.17)-(2.19) reduces to the method proposed by Professor F. J. Murray [18].

REMARK 10. In the discussion of the generalized Murray method we have assumed that A is bounded. However, with due precaution the method is also applicable to *densely defined linear closed* operators for f in $\overline{R(A)}$, provided that the complete linearly independent system $\{\psi_i\}, \psi_i \in D(A^*)$, $i = 1, 2, \ldots$, has the property that the set $\{A^*\psi_i\}$, $i = 1, 2, \ldots$, determines the range space $\overline{R(A^*)}$.

The method is also applicable to a still more general class of operator equations $Au = f$, where A is *not closed but has a densely defined adjoint operator* A^* over its domain $D(A)$ (and the operators occuring in most applications do have such adjoints). However, in this case, as is easily seen from relations (2.16), the method yields only a generalized or also called *weak solution* of the equation $Au = f$. The latter is defined as an element u in H such that for every v in $D(A^*)$ the relation $(f, v) = (u, A^*v)$ is true.

2.5. **Connection between the generalized Murray method and direct methods.** In this section we show that in case A in (2.15) satisfies some additional conditions customarily assumed in the investigation of the methods of Galerkin, Ritz, and least squares, the generalized Murray method, after the proper renorming of H, reduces to the above methods. This enables us to derive the necessary and sufficient conditions for the existence of the solution and the convergence of the above direct methods for a larger class of operators. It seems th t in this sense the generalized Murray method is superior to the above methods.

GALERKIN METHOD. If we denote $\phi_i = A^*\psi_i$, $i = 1, 2, \ldots$, then the equations (2.19) of the generalized Murray method may be written in the form

$$\sum_{i=1}^{n} (A\phi_i, \psi_j) a_i^n = (f, \psi_j), \quad j = 1, 2, \ldots, n, \tag{2.19$_1$}$$

which are the equations of the Galerkin method [19] for eq. (2.15) with u_n taken in the form $u_n = \sum_{i=1}^{n} a_i^n \phi_i$. In this case we obtain the necessary and sufficient conditions for the convergence of the Galerkin method subject to the above

choice of ϕ_i. This is a much sharper result for the above choice of $\{\phi_i\}$ than the one obtained in [16; 19], where the study is restricted to establishing only the sufficient condition for the convergence of the Galerkin method under the assumption that $A = 1 + T, T$ is completely continuous, and eq. (2.15) has a solution.

RITZ METHOD. Let $A = A^*$ and $M > m > 0$ be constants such that

$$m(u,u) \leqq (Au,u) \leqq M(u,u), \qquad u \in H. \tag{2.22}$$

Let H' denote the space H with respect to the new metric

$$[u,v] = (Au,v), \qquad |u| = [u,u]^{1/2}. \tag{2.23}$$

By (2.22) the norms of H' and H are equivalent and so H' and H are the same.

We shall show that in H' the generalized Murray method reduces to the Ritz method. Indeed, if we denote by H_n' the subspace generated by $\phi_i = A\psi_i$, $i = 1, \ldots, n$ and by P_n' the orthogonal projection on H_n' in H', then u_n, according to the generalized Murray method, is determined from the condition that $u_n = p_n' u = \sum_{i=1}^n a_i^n A\psi_i$ which gives the system

$$\sum_{i=1}^{n} [A\psi_i, A\psi_j]\, a_i^n = [f,\psi_j], \qquad j = 1,2,\ldots,n. \tag{2.19$_2$}$$

In view of (2.23), the system (2.19$_2$) can be put in the form

$$\sum_{i=1}^{n} (A\phi_i,\phi_j)\, a_i^n = (f,\phi_j), \qquad j = 1,2,\ldots,n \tag{2.24}$$

which are the equations of the Ritz method.

THE METHOD OF LEAST SQUARES. In accordance with this method assume that there exists a constant $c > 0$ such that for all u in H $\|Au\| \geqq c\|u\|$. Let H_A denote the space H with respect to the new metric

$$[u,v]_A = (Au,Av), \qquad |u|_A = [u,u]_A^{1/2}; \tag{2.25}$$

let H_n'' be the subspace spanned by $\phi_i = A^*\psi_i$, $i = 1,\ldots,n$, and P_n'' the projection on H_n'' in H_A. The approximate solution $u_n = \sum_{i=1}^n a_i^n A^*\psi_i$, according to the generalized Murray method, is determined from the condition that $u_n = P_n'' u$ which yields the algebraic system

$$\sum_{i=1}^{n} (A\phi_i, A\phi_j)\, a_i^n = (f, A\phi_j), \qquad j = 1,2,\ldots,n. \tag{2.19$_3$}$$

The latter are the well-known equations of the method of least squares. In this case, Theorem 2.3 also gives the necessary and sufficient condition for the convergence of the method of least squares.

III. Iterative methods

The purpose of this chapter is to present a fairly general iteration method by means of a geometrical approach based on the notion of projection which in a certain sense unifies and extends the results of [6; 7; 10] to a larger class of operators and which furnishes a basis for comparison of its special cases and for the discovery of a very effective procedure, called here the *method with minimal errors*, that appears to have escaped the notice of various investigators in this field.

3.1. **General inequality.** For subsequent use we derive in this section an inequality which gives an easy but nevertheless substantial generalization of the inequalities derived recently by Greub and Rheinboldt [5]. It turns out that the well-known inequalities of Pólya-Szegö [5], Kantorovich [7], and Krasnoselsky-Krein [10] are special cases of our inequality and, moreover, all three are equivalent.

Let A be a bounded K^*-p.d. operator in H, where K^* is the adjoint of a continuously invertible and bounded operator K, i.e., there exists a constant $c > 0$ such that for all u in H

$$(Au,K^*u) = (KAu,u) \geqq c \| u \|^2. \tag{3.1}$$

Assume also that there exists a second bounded linear operator B and a constant $d > 0$ such that for all u in H

$$(KAu,BAu) \geqq d(KAu,u). \tag{3.2}$$

Let H' denote the space H with respect to the new metric

$$[u,v] = (KAu,v), \qquad |u| = [u,u]^{1/2}. \tag{3.3}$$

In view of (3.1) the norms in both spaces are equivalent and therefore these spaces are the same.

Lemma 3.1. *If the operators A, K and B satisfy* (3.1) *and* (3.2) *and D is the H'-norm of the operator BA, then for all $u \neq 0$*

$$\frac{(KABAu,BAu)(KAu,u)}{(KAu,BAu)^2} \leqq \frac{(D + d)^2}{4D \cdot d}. \tag{3.4}$$

Proof. By (3.1) the operator $S = KA$ is positive definite in H and by (3.2) the operator $L = BA$ is symmetrizable by S. By Lemma 1.5 the operator L, considered as an operator in the space H', is bounded and symmetric. Let D denote the norm of L as an operator in H'. Then by (3.2) and the definition of D we have for all u

$$d|u|^2 \leqq [Lu,u] \leqq D|u|^2. \tag{3.5}$$

Since, in view of the definition of the metric in H',

$$\frac{(KABAu,BAu)(KAu,u)}{(KAu,BAu)^2} = \frac{[Lu,Lu][u,u]}{[u,Lu]^2}$$

the inequality (3.4) follows immediately from (3.5) and Theorem 2 in [5]([15]).

Special cases. (i) *Kantorovich inequality.* If T is self-adjoint and $0 < mI \leqq T \leqq MI$, then

$$\frac{(u,Tu)\,(T^{-1}u,u)}{(u,u)^2} \leqq \frac{(M+m)^2}{4Mm}, \qquad u \in H,\ u \neq 0. \tag{3.4$_1$}$$

To derive (3.4$_1$) from (3.4) put $A = I$, $B = T$, and $K = T^{-1}$ and note that A, K, and B so chosen satisfy all the hypotheses used in deriving (3.4) with $c = 1/M$, $d = m$, and $D = M$. It is obvious that for such a choice of A, K, and B (3.4) reduces to (3.4$_1$).

(ii) *Krasnoselsky-Krein inequality*([16]). If T is self-adjoint and $0 < mI \leqq T \leqq MI$, then

$$\frac{(Tu,Tu)\ (u,u)}{(Tu,u)^2} \leqq \frac{(M+m)^2}{4Mm}, \qquad u \in H,\ u \neq 0. \tag{3.4$_2$}$$

The inequality (3.4$_2$) follows immediately from (3.4) if in (3.4) we set $A = I$, $K = I$, and $B = T$ with $c \leqq 1$, $d = m$, and $D = M$.

(iii) *Generalized Pólya-Szegö inequality*([17]). If T and S are self-adjoint and commutative such that $0 < m_1 I \leqq T \leqq M_1 I$ and $0 < m_2 I \leqq S \leqq M_2 I$, then

$$\frac{(Tu,Tu)(Su,Su)}{(Tu,Su)^2} \leqq \frac{(M_1M_2 + m_1m_2)^2}{4m_1m_2M_1M_2}, \qquad u \in H,\ u \neq 0. \tag{3.4$_3$}$$

To obtain (3.4$_3$) from (3.4) set $A = S$, $K = S$, and $B = S^{-1}TS^{-1}$. Using the properties of T and S, it can be easily shown that A,K,B so chosen satisfy our hypothesis with $c = m_2^2$, $d = m_1/M_2$, and $D = M_1/m_2$. Thus, setting them into (3.4) we get (3.4$_3$):

$$\frac{(Tu,Tu)(Su,Su)}{(Tu,Su)^2} \leqq \frac{(M_1/m_2 + m_1/M_2)^2}{4M_1/m_2 \cdot m_1/M_2} = \frac{(M_1M_2 + m_1m_2)^2}{4m_1m_2M_1M_2}.$$

We conclude this section by showing that all three special inequalities are equivalent. The equivalence of (3.4$_1$) and (3.4$_3$) was shown in [5]. Hence, it is sufficient to show, for example, that (3.4$_2$) and (3.4$_3$) are equivalent.

(3.4$_3$) $\rightarrow$ (3.4$_2$). This case is trivial since (3.4$_3$) reduces to (3.4$_2$) when in (3.4$_3$) we set $S = I$ and $m_2 = M_2 = 1$.

([15]) For the inequality given in Theorem 2 in [5] see (3.4$_3$) below.

([16]) In case T is a finite matrix, (3.4$_2$) was obtained from the spectral resolution by Krasnoselsky and Krein [10].

([17]) (3.4$_3$), in the finite case, when expressed in terms of the spectral decomposition was stated by Pólya and Szegö [5] and in the general form (3.4$_3$) was derived in [5].

$(3.4_2) \rightarrow (3.4_3)$. If $N = TS^{-1}$, then $0 < m_1/M_2 I \leqq N \leqq M_1/m_2 I$ and by (3.4_2)

$$\frac{(Nu, Nu)(u, u)}{(Nu, u)^2} \leqq \frac{(M_1/m_2 + m_1/M_2)^2}{4M_1/m_2 + m_1/M_2} = \frac{(M_1M_2 + m_1m_2)^2}{4m_1m_2M_1M_2}.$$

Using the self-adjointness of the operators involved and the fact that S maps H onto H we obtain, for $u = Sv$, $(Nu, Nu) = (TS^{-1}Sv, TS^{-1}Sv) = (Tv, Tv)$, $(u,u) = (Sv, Sv)$, and $(Nu, u) = (TS^{-1}Sv, Sv) = (Tv, Sv)$. Substituting these relations into the last inequality we obtain (3.4_3).

3.2. **The iterative method with relative minimal errors.** In this section we discuss an iteration method for finding the solution u of a linear operator equation in a complex Hilbert space H

$$Au = f, \qquad f \in H \tag{3.5}$$

where A is bounded and K^*-p.d. in H. Observe that by Theorem 1.1 eq. (3.5) has a unique solution u for every f in H. Assume also that there exists an operator B satisfying (3.2).

If u_i is any approximation to the solution u of (3.5), then $r_i = Au_i - f = Aw_i$ denotes the *residual* and $w_i = u_i - u$ the *error* vector at u_i. The iterative process for solving eq. (3.5) is defined as follows:

Let u_0 be some given initial approximation to u and u_n the approximation to u obtained at the nth step of our procress. Then the succeeding appoximation u_{n+1} is taken in the form

$$u_{n+1} = u_n - t_n Br_n, \qquad n = 0, 1, 2, \ldots. \tag{3.6}$$

From (3.6) we see that the character of the iteration at each step is then completely determined by the choice of the scalars t_n. Our choice of t_n will be governed by the condition that the error vector

$$w_{n+1} = w_n - t_n Br_n, \qquad n = 0, 1, 2, \ldots \tag{3.7}$$

be orthogonal to the vector Br_n in the sense of the H'-metric, i.e., $[w_n - t_nBr_n, Br_n] = 0$. This condition yields the following values for t_n

$$t_n = \frac{[w_n, Br_n]}{[Br_n, Br_n]} = \frac{(Kr_n, Br_n)}{(KABr_n, Br_n)}, \qquad n = 0, 1, 2, \ldots. \tag{3.8}$$

Note that for this value of t_n the vector t_nBr_n is the orthogonal projection in the sense of H'-metric of the error vector w_n on Br_n. Consequently w_{n+1} is obtained from w_n by subtracting from it the projection of w_n on Br_n and the error $|w_{n+1}|$ assumes its minimum for this choice of t_n. For that reason we call the present iterative procedure *the method with relative minimal errors* (RME-method). Observe also that the assumed properties of K and B imply that $t_n \neq 0$ whenever $r_n \neq 0$. Thus, for u_{n+1} we obtain the iterative formula

$$u_{n+1} = u_n - \frac{(Kr_n, Br_n)}{(KABr_n, Br_n)} Br_n , \qquad n = 0,1,2,\dots . \tag{3.9}$$

The corresponding error and residual vectors are given by the formulas

$$w_{n+1} = w_n - \frac{(Kr_n, Br_n)}{(KABr_n, Br_n)} Br_n , \qquad n = 0,1,2,\dots, \tag{3.10}$$

$$r_{n+1} = r_n - \frac{(Kr_n, Br_n)}{(KABr_n, Br_n)} ABr_n , \qquad n = 0,1,\dots . \tag{3.11}$$

To obtain the convergence of this process and the estimate of error note first that, in view of (3.10),

$$|w_{n+1}| = \left(|w_n|^2 - \frac{[w_n, Br_n]^2}{|Br_n|^2}\right)^{1/2} . \tag{3.12}$$

(3.3) and the fact that $r_n = Aw_n$ imply that (3.12) can be put in the form to which the inequality (3.4) is applicable. In fact,

$$|w_{n+1}| = \left(1 - \frac{[w_n, BAw_n]^2}{[BAw_n, BAw_n][w_n, w_n]}\right)^{1/2} |w_n|$$

$$= \left(1 - \frac{(KAw_n, BAw_n)^2}{(KABAw_n, BAw_n)(KAw_n, w_n)}\right)^{1/2} |w_n| .$$

Applying the inequality (3.4) we obtain

$$|w_{n+1}| \leqq \left(1 - \frac{4dD}{(D+d)^2}\right)^{1/2} |w_n| = \frac{D-d}{D+d} |w_n| , \tag{3.13}$$

whence we get the following error estimate in the H'-norm

$$|w_{n+1}| \leqq \left(\frac{D-d}{D+d}\right)^n |w_0| . \tag{3.14}$$

This implies the convergence of the RME-method in the H'-norm and because of (3.1) also in the H-norm. Moreover, (3.12) shows that the convergence is monotonic in the sense that $|w_{n+1}| \to 0$ monotonically as $n \to \infty$.

The above discussion may be summarized in the following

THEOREM 3.1. *If the operators K and B satisfy respectively the conditions* (3.1) *and* (3.2), *then the* RME-*method is monotonically convergent. The rate of convergence is characterized by the inequality* (3.14).

3.3. **Special cases.** The general RME-method described in the last section is not precise until the choice of the operators K and B has been made. In this section we derive three of its important special cases.

GRADIENT METHOD. (i) *The positive definite case.* If we let $K = B = I$, then (3.1) and (3.2) reduce to

$$(3.1_1) \qquad (Au,u) \geqq c_1 \| u \|^2 , \qquad c_1 > 0$$

$$(3.2_1) \qquad (Au,Au) \geqq d_1 (Au,u) , \quad d_1 > 0$$

showing that A is positive definite. The metric (3.3) becomes

$$(3.3_1) \qquad [u,v] = (Au,v) , \qquad |u| = [u,u]^{1/2} .$$

Starting with $u_0, r_0 = Au_0 - f$, the iteration process (3.9) becomes

$$(3.9_1) \qquad u_{n+1} = u_n - \frac{(r_n,r_n)}{(Ar_n,r_n)} r_n , \qquad n = 0,1,2,\ldots .$$

Theorem 3.1. shows that u_{n+1} determined by (3.9_1) converges monotonically to the exact solution u of eq. (3.5) in the space H'_1 with metric (3.3_1). In view of (3.1_1), u_{n+1} converges also in the H-metric. If D_1 denotes the norm of A as an operator in H'_1 and $\theta_1 = |w_0|$, then by (3.1_1) and (3.14) the error estimate in H is given by

$$(3.14_1) \qquad \| u_{n+1} - u \| \leqq \frac{\theta_1}{c_1} \left(\frac{D_1 - d_1}{D_1 + d_1} \right)^n .$$

Let us note that the convergence is also monotonic in the H-norm. This follows from the inequality

$$(3.15) \qquad (r_n,w_n) - \frac{(r_n,r_n)^2}{(Ar_n,r_n)} = |w_n|^2 - \frac{[w_n,r_n]^2}{|r_n|^2} \geqq 0$$

and the fact obtained from (3.9_1):

$$\| w_{n+1} \|^2 = \| w_n \|^2 - 2 \frac{(r_n,r_n)}{(Ar_n,r_n)} \left[(r_n,w_n) - \frac{1}{2} \frac{(r_n, r_n)^2}{(Ar_n, r_n)} \right].$$

It is known that for a positive definite A the problem of solving eq. (3.5) is equivalent to the problem of minimizing $F(u) = (Au,u) - (f,u) - (u,f)$. If the gradient method is applied to the latter problem, then the minimizing sequence $\{u_{n+1}\}$ of $F(u)$ is given by (3.9_1) [7,8]. Thus, in case $K = B = I$, the RME-method reduces to the gradient method and gives the latter the geometrical basis.

We conclude this case by showing that our error estimate (3.14_1) is at least as good as the best available. In fact, if the constants $C_1 > c_1 > 0$ are such that for all $u \in H$ we have $c_1 \| u \|^2 \leqq (Au,u) \leqq C_1 \| u \|^2$, then Kantorovich [7], using the spectral resolution of A, derived the estimate

$$\| u_{n+1} - u \| \leqq \frac{\theta_1}{c_1} \left(\frac{C_1 - c_1}{C_1 + c_1} \right)^n$$

considered the best available in case u_{n+1} is given by (3.9_1). To prove our statement observe that from the definition of d_1 in (3.2_1) it follows easily that $c_1 \leqq d_1$. On the other hand, by Lemma 1.5, $D_1 \leqq C_1$. This implies that $C_1/c_1 \geqq D_1/d_1 > 1$. Since the function $g(x) = x/(x^2+1)$ is monotonically decreasing for $x \geqq 1$ and $C_1/c_1 \geqq D_1/d_1 > 1$, it follows that

$$\frac{C_1c_1}{C_1^2+c_1^2} \leqq \frac{D_1d_1}{D_1^2+d_1^2}$$

from which one derives that

$$\frac{D_1 - d}{D_1 + d} \leqq \frac{C_1 - c_1}{C_1 + c_1}. \tag{3.16}$$

In fact,

$$\frac{C_1c_1}{C_1^2+c_1^2} \leqq \frac{D_1d_1}{D_1^2+d_1^2}$$

implies that

$$2 + \left(\frac{C_1^2+c_1^2}{C_1c_1}\right) \geqq 2 + \left(\frac{D_1^2+d_1^2}{D_1d_1}\right),$$

or equivalently that

$$\frac{(C_1 + c_1)^2}{C_1c_1} \geqq \frac{(D_1 + d_1)^2}{D_1d_1}.$$

This in turn implies that

$$1 - \frac{4C_1c_1}{(C_1 + c_1)^2} \geqq 1 - \frac{4D_1d_1}{(D_1 + d_1)^2} \quad \text{or} \quad \frac{(C_1 - c_1)^2}{(C_1 + c_1)^2} \geqq \frac{(D_1 - d_1)^2}{(D_1 + d_1)^2}$$

and thus gives the inequality (3.16). This proves our statement and completes the discussion of this case.

(ii) *The non-Hermitian case.* If we let $K = B = A^*$, then (3.1) and (3.2) give

$$\| Au \|^2 \geqq c_2 \| u \|^2, \tag{3.1$_2$}$$

$$\| A^*Au \|^2 \geqq d_2 \| Au \|^2, \tag{3.2$_2$}$$

showing that A is continuously invertible. The metric (3.3) reduces to

$$[u,v] = (Au,Av)\ , \qquad |u| = \| Au \|, \tag{3.3$_2$}$$

and the iteration formula (3.9) reduces to the formula

$$u_{n+1} = u_n - \frac{\| A^*r_n \|^2}{\| AA^*r_n \|^2} \cdot A^*r_n\ , \qquad n = 0,1,\ldots \tag{3.9$_2$}$$

which, as is known [7], is exactly the same as one used for the determination of the minimizing sequence $\{u_{n+1}\}$ when the gradient method is applied to the minimum problem of the functional $F(u) = (A^*Au,u) - (u,A^*f) - (A^*f,u) + (f,f)$, the latter being equivalent to the problem of solving eq. (3.5) when A is non-symmetric. Thus, the RME-method reduces in this event to the gradient method. Theorem 3.1 gives its convergence and estimate of error. As in (i) one also proves that the convergence is monotonic in the H-norm.

THE METHOD WITH MINIMAL RESIDUALS. In case A is positive definite the choice of $K = A$ and $B = I$ leads to a very effective iteration process. (3.1) and (3.2) reduce in this event to

$$\|Au\|^2 \geqq c_3 \|u\|^2, \tag{3.1$_3$}$$

$$(A^2u,Au) \geqq d_3 \|Au\|^2 \tag{3.2$_3$}$$

and the space H' becomes the space H'_3 with the metric

$$[u,v] = (Au,Av)\ , \qquad |u| = \|Au\|. \tag{3.3$_3$}$$

The RME-method (3.9) takes in this event the form

$$u_{n+1} = u_n - \frac{(Ar_n,r_n)}{(Ar_n,Ar_n)} r_n\ , \qquad n = 0,1,2,\ldots, \tag{3.9$_3$}$$

while (3.12) reduces to

$$|w_{n+1}|^2 = \|r_{n+1}\|^2 = \|r_n\|^2 - \frac{(Ar_n,r_n)^2}{\|Ar_n\|^2} \tag{3.12$_3$}$$

showing that the residual r_{n+1} is obtained from r_n by subtracting from it the projection of r_n on Ar_n. (3.12$_3$) shows that in the algorithm (3.9$_3$) the squared residual diminishes at each step of the process and for the value of t_n in this case $\|r_{n+1}\|^2$ has the minimum magnitude. Theorem 3.1 gives the convergence and the estimate of error of this method. Moreover, $\{u_{n+1}\}$ determined by (3.9$_3$) converges also monotonically in the H-norm. In fact, by (3.9$_3$)

$$\|w_{n+1}\|^2 = \|w_n\|^2 - 2\frac{(Ar_n,r_n)}{(Ar_n,Ar_n)}\left[(r_n,w_n) - \frac{1}{2}\frac{(Ar_n,r_n)}{(Ar_n,Ar_n)}(r_n,r_n)\right].$$

Thus, the proof of monotonic convergence is equivalent to the proof that

$$(A^{-1}r_n,r_n)(Ar_n,Ar_n) \geqq (r_n,r_n)(Ar_n,r_n). \tag{3.17}$$

The latter follows from Schwarz inequality

$$\begin{aligned}(Ar_n,r_n)^2(r_n,r_n)^2 &= (Ar_n,r_n)^2(A^{1/2}r_n,(A^{-1})^{1/2}r_n)^2\\ &\leqq \|Ar_n\|^2\|r_n\|^2\|A^{1/2}r_n\|^2\|(A^{-1})^{1/2}r_n\|^2\\ &= (Ar_n,Ar_n)(r_n,r_n)(Ar_n,r_n)(A^{-1}r_n,r_n).\end{aligned}$$

In case A is a symmetric and positive definite matrix of *finite order* the method of this section was extensively studied by Krasnoselsky and Krein [10] which they called the *iteration method with minimal residuals.* Thus, our Theorem 3.1 extends their results to bounded linear and positive definite operators in Hilbert space H and gives it a geometrical basis. Furthermore, a very useful and economical computational scheme devised by them can also be extended to operators considered here.

THE METHOD WITH MINIMAL ERRORS. The third special case of the RME-method that appears to be very useful for computational purposes is a new method obtained by selecting $K = A^{-1}$ and $B = A^*$. In this event (3.1) becomes an identity while (3.2) reduces to

$$\| Au \|^2 \geqq d_4 \| u \|^2. \tag{3.2$_4$}$$

The metric (3.3) in this case is identical with the metric of H and (3.9) reduces to

$$u_{n+1} = u_n - \frac{\| r_n \|^2}{\| A^* r_n \|^2} A^* r_n , \qquad n = 0,1,2,\ldots, \tag{3.9$_4$}$$

while (3.12) becomes

$$\| w_{n+1} \|^2 = \| w_n \|^2 - \frac{\| r_n \|^4}{\| A^* r_n \|^2}. \tag{3.12$_4$}$$

The formula (3.9$_4$) shows that the error vector w_{n+1} is obtained by subtracting from w_n the projection of w_n on A^*Aw_n. By (3.12$_4$) the squared error $\| w_{n+1} \|^2$ diminishes at each step of this process in such a way that its magnitude is minimum. For that reason we call it *the method with minimal errors.* The Theorem 3.1 establishes the monotonic convergence of the method with minimal errors and gives the error estimate

$$\| u_{n+1} - u \| \leqq \theta_4 \left(\frac{D_4 - d_4}{D_4 + d_4} \right)^n , \qquad \theta_4 = \| u_0 - u \|, \tag{3.14$_4$}$$

where D_4 denotes the norm of A^*A. Furthermore, we shall prove in the next section that from the point of view of the decrease in the magnitude of error $\| w_{n+1} \|$ at each step of the iteration the present method is better than the gr adient method considered above.

3.4. **Comparison of the special cases of the** RME**-method.** In this section we compare the three special methods from the point of view of the decrease in the magnitude of error and the decrease in the magnitude of residual at each step of the iteration.

LEMMA 3.2. *If A is positive definite, then*

(a) *From the point of view of the decrease in the magnitude of error $\| w_{n+1} \|$ at each step of the iteration the gradient method is better than the method with minimal residuals.*

(b) *From the point of view of the decrease in the magnitude of residual* $\| r_{n+1} \|$ *at each step of the iteration the method with minimal residuals is better than the gradient method.*

Proof. (a) If $w_n = u_n - u$ is the error vector at the nth step of the iteration and w_{n+1} and $\tilde{w}_{n+1}$ are determined at the succeeding step by the method with minimal residuals and the gradient method respectively, then by (3.9_3) and (3.9_1) we get the structually similar formulas

$$w_{n+1} = w_n - \frac{(Ar_n,r_n)}{(Ar_n,Ar_n)} r_n , \qquad n = 0,1,\ldots \tag{3.18}$$

$$\tilde{w}_{n+1} = w_n - \frac{(r_n,r_n)}{(Ar_n, r_n)} r_n , \qquad n = 0,1,\ldots . \tag{3.19}$$

The relations (3.18) and (3.19) imply that for each n

$$\| w_{n+1} \|^2 - \| \tilde{w}_{n+1} \|^2$$
$$= \left[\frac{(Ar_n r_n)^2 (r_n,r_n)}{(Ar_n,Ar_n)^2} - 2\frac{(Ar_n,r_n)\ (r_n,w_n)}{(Ar_n,Ar_n)}\right] - \left[\frac{(r_n,r_n)^3}{(Ar_n,r_n)^2} - 2\frac{(r_n,r_n)\,(r_n,w_n)}{(Ar_n r_n)}\right] .$$

The last equality can be written in the following convenient form

$$\| w_{n+1} \|^2 - \| \tilde{w}_{n+1} \|^2$$
$$= \left[\left(\frac{(Ar_n,r_n)}{\| Ar_n \|^2} - \frac{(r_n,w_n)}{\| r_n \|^2}\right)^2 - \left(\frac{\| r_n \|^2}{(Ar_n,r_n)} - \frac{(r_n,w_n)}{\| r_n \|^2}\right)^2\right] \| r_n \|^2. \tag{3.20}$$

In view of (3.15) and (3.17) and the readily derived inequality

$$\frac{(r_n,r_n)}{(Ar_n,r_n)} \geqq \frac{(Ar_n,r_n)}{\| Ar_n \|^2}$$

(3.20) implies that $\| w_{n+1} \|^2 - \| \tilde{w}_{n+1} \| \geqq 0$. This proves (a).

(b) If $r_n = Aw_n$ is the residual vector at the nth step of the iteration and r_{n+1} and $\tilde{r}_{n+1}$ are determined at the succeeding step by the method with minimal residuals and the gradient method respectively, then from (3.18) and (3.19) we get the structurally similar formulas

$$r_{n+1} = r_n - \frac{(Ar_n,r_n)}{(Ar_n, r_n)} Ar_n , \qquad n = 0,1,\ldots , \tag{3.21}$$

$$\tilde{r}_{n+1} = r_n - \frac{(r_n, r_n)}{(Ar_n, r_n)} Ar_n , \qquad n = 0,1,\ldots . \tag{3.22}$$

Using (3.21) and (3.22) we obtain

$$\| \tilde{r}_{n+1} \|^2 - \| r_{n+1} \|^2 = \frac{\| r_n \|^4}{(Ar_n,r_n)^2} \| Ar_n \|^2 - 2\| r_n \|^2 + \frac{(Ar_n,r_n)^2}{\| Ar_n \|^2} .$$

Denoting the coefficient in (3.21) by t_n and in (3.22) by $\tilde{t}_n$, then from above we get

$$\| \tilde{r}_{n+1} \|^2 - \| r_{n+1} \|^2 = \left(\frac{\| r_n \|^2}{(Ar_n, r_n)} - \frac{(Ar_n, r_n)}{\| Ar_n \|^2} \right)^2 \| Ar_n \|^2 = (\tilde{t}_n - t_n)^2 \| Ar_n \|^2 .$$

The last relation proves (b) and shows at the same time that one step of the iteration by the method with minimal residuals diminishes the norm of the residual $\| r_{n+1} \|^2$ more than the corresponding step of the gradient method by the magnitude $(\tilde{t}_n - t_n)^2 \| Ar_n \|^2$.

LEMMA 3.3. *If A is invertible, then*

(a) *From the point of view of the decrease in the magnitude of error $\| w_{n+1} \|$ at each step of the iteration the method with minimal errors is better than the gradient method applied to non-Hermitian operators.*

(b) *From the point of view of the decrease in the magnitude of residual $\| r_{n+1} \|$ at each step of the iteration the gradient method is better than the method with minimal errors.*

Proof. (a) If w'_{n+1} and $\tilde{w}'_{n+1}$ are determined at the succeeding step by the method with minimal errors and the gradient method respectively, then by (3.9_4) and (3.9_2)

$$w'_{n+1} = w_n - \frac{\| r_n \|^2}{\| A^* r_n \|^2} A^* r_n , \qquad n = 0, 1, \ldots , \tag{3.23}$$

$$\tilde{w}'_{n+1} = w_n - \frac{\| A^* r_n \|^2}{\| AA^* r_n \|^2} A^* r_n , \qquad n = 0, 1, \ldots . \tag{3.24}$$

From (3.23) and (3.24) we immediately obtain

$$\| \tilde{w}'_{n+1} \|^2 - \| w'_{n+1} \|^2 = \frac{\| A^* r_n \|^6}{\| AA^* r_n \|^4} - 2 \frac{\| A^* r_n \|^2 \| r_n \|^2}{\| AA^* r_n \|^2} + \frac{\| r_n \|^4}{\| A^* r_n \|^2} .$$

Denoting the coefficient in (3.23) by t'_n and in (3.24) by $\tilde{t}'_n$, then from above we get

$$\| \tilde{w}'_{n+1} \|^2 - \| w'_{n+1} \|^2 = \left(\frac{\| A^* r_n \|^2}{\| AA^* r_n \|^2} - \frac{\| r_n \|^2}{\| A^* r_n \|^2} \right)^2 \| A^* r_n \|^2 = (\tilde{t}'_n - t'_n)^2 \| A^* r_n \|^2 .$$

Thus we see that one step of the iteration by the method with minimal errors diminishes the norm of the error $\| w_{n+1} \|^2$ more than the corresponding step of the gradient method by the magnitude $(\tilde{t}'_n - t'_n)^2 \| A^* r_n \|^2$.

(b) Using (2.23) and (3.24) we obtain

$$r'_{n+1} = r_n - \frac{\| r_n \|^2}{\| A^* r_n \|^2} AA^* r_n , \qquad n = 0, 1, \ldots , \tag{3.25}$$

$$\tilde{r}'_{n+1} = r_n - \frac{\| A^* r_n \|^2}{\| AA^* r_n \|^2} AA^* r_n , \qquad n = 0, 1, \ldots , \tag{3.25'}$$

from which we derive that

$$\| r'_{n+1} \|^2 - \| \tilde{r}'_{n+1} \|^2 = \frac{\| r_n \|^4}{\| A^* r_n \|^4} \| AA^* r_n \|^2 - 2 \| r_n \|^2 + \frac{\| A^* r_n{}^4 \|}{\| AA^* r_n \|^2}$$

or equivalently

$$\| r'_{n+1} \|^2 - \| \tilde{r}'_{n+1} \|^2 = \left(\frac{\| r_n \|^2}{\| A^* r_n \|^2} - \frac{\| A^* r_n \|^2}{\| AA^* r_n \|^2} \right)^2 \| AA^* r_n \|^2$$

$$= (t'_n - \tilde{t}'_n)^2 \| AA^* r_n \|^2.$$

This proves (b) and shows that one step of the iteration by the gradient method diminishes $\| r_{n+1} \|^2$ more than the corresponding step of the method with minimal errors by the magnitude $(t'_n - \tilde{t}'_n)^2 \| AA^* r_n \|^2$.

3.5. **Generalized gradient method for unbounded K-p.d. operators.** In this section we generalize the gradient method to the solution of eq. (1.1) where A is an *unbounded* K-p.d. operator studied in Chapter I.

Let us assume that on $D(A)$ there is given a "well investigated" K-p.d. operator L and a constant $\xi > 0$ such that for all u in $D(A)$

$$(Lu,Ku) \geqq \xi \| Ku \|^2. \tag{3.26}$$

Denote by H'_K the completion of $D(A)$ in the metric

$$[u,v]' = (Lu,Kv) , \qquad |u|' = \{[u,u]'\}^{1/2}. \tag{3.27}$$

Applying the results of Chapter I we see that K can be extended to a bounded linear operator of all of H'_K to H and L has a solvable generalized Friedrichs extension which is closed and has a bounded inverse on H. We assume that these extensions have been carried out.

Consider now the linear operator equation

$$Au = f, \qquad f \in H, \tag{3.28}$$

where A has the property that there exist constants $M > m > 0$ such that

$$m(Lu,Ku) \leqq (Au,Ku) \leqq M(Lu,Ku) , \qquad u \in D(A). \tag{3.29}$$

The inequalities (3.26) and (3.29) imply that A is K-p.d. and that the norm induced by $[u,u] = (Au,Ku)$ is equivalent to the norm induced by $[u,u]' = (Lu,Ku)$. So the space H_K associated with A is the same as the space H'_K associated with L and A has a solvable generalized Friedrichs extension. This also implies that the inner product (Au,Kv) is a bounded bilinear functional on the dense set $D(A)$ of H'_K and hence can be extended to a bounded bilinear functional in all of H'_K. We shall denote it by $[u,v]$, i.e., $[u,v] = (Au,Kv)$ whenever $u \in D(A)$ and $v \in H'_K$. Consequently by the Fréchet-Riesz theorem, there exists a bounded and everywhere defined linear operator A' of H'_K into H'_K such that for all u, $v \in H'_K$

$$[u,v] = [A'u,v]' , \qquad |A'|' \leqq M. \tag{3.30}$$

If $A'u = 0$, then by (3.30) $[u,v] = 0$ for all v in H'_K. Since H'_K and H_K are the same this implies that $u = 0$, i.e., A'^{-1} exists. By (3.29)

$$m[u,u]' \leqq [A'u,u]' \leqq M[u,u]' , \qquad u \in H'_K. \tag{3.31}$$

This implies that A' is positive definite on H'_K, the inverse A'^{-1} is bounded, and $R(A') = H'_K$.

For a given element w in $D(A)$ consider the equation

$$Lu = Aw. \tag{3.32}$$

If v is any element in H'_K, then obviously eq. (3.32) is equivalent to

$$[u,v]' = [w,v]. \tag{3.33}$$

However, in the form (3.33) the given equation has also sense for u, $w \in H'_K$. An element u in H'_K which for a given w in H'_K satisfies eq. (3.33) for all v in H'_K will be called the *generalized solution* of eq. (3.32). It is not hard to see that for any w in H'_K eq. (3.32) has a unique generalized solution. Indeed, (3.30) and (3.33) imply that the solution sought must satisfy the relation

$$[u,v]' = [A'w,v]' \tag{3.34}$$

for all $v \in H'_K$. This shows that

$$u = A'w, \quad u \in H'_K. \tag{3.35}$$

Let us also observe that the equation

$$Lu = Aw - f, \qquad f \in H, \tag{3.36}$$

reduces to an equation of the form (3.32) since $f_1 = L^{-1}f$ is in $D(L) \subseteq H'_K$ and hence (3.36) can be written in the form $L(u + f_1) = Aw$. Thus, eq. (3.36) has a generalized solution u of the form

$$u = A'w - f_1. \tag{3.37}$$

We can now generalize the gradient method discussed in (i) of § 3.3 to our eq. (3.28). If v is any element in H'_K and $f_1 = L^{-1}f$, then (3.27) and (3.30) imply that eq. (3.28) can be written in the form

$$[A'u,v]' = [f_1,v]', \tag{3.38}$$

which in turn is equivalent to the equation

$$A'u = f_1 , \qquad f_1 \in H'_K. \tag{3.39}$$

Considered as an equation in the space H'_K, (3.39) is the same as the equation discussed in §3.3 (i). The operator A' is a continuous transformation of H'_K onto

H_K' which is positive definite and satisfies the condition (3.31). We may therefore apply to eq. (3.39) the entire procedure of §3.3 (i).

Thus if u_0 is some initial approximation to the solution of eq. (3.39) and $z_0 = A'u_0 - f_1$, then according to (3.9_1) the next iterant u_1 is given by

$$u_1 = u_0 - t_0 z_0, \tag{3.40}$$

where

$$t_0 = \frac{[z_0, z_0]'}{[A'z_0, z_0]'} = \frac{[z_0, z_0]'}{[z_0, z_0]}. \tag{3.41}$$

But now we take u_0 in $D(A)$ and z_0 is in $D(L)$ or at any rate in H_K'. Note that if we let $r_0 = Au_0 - f$, then the comparison with (3.37) shows that z_0 coincides in this case with the generalized solution of eq. (3.36), i.e., z_0 is such that

$$Lz_0 = Au_0 - f = r_0. \tag{3.42}$$

Observe that for a given u_0 in $D(A)$ the above discussion assures the existence of the generalized solution z_0. Furthermore, (3.41) has sense since $z_0 \in H_K'$. By (3.40) u_1 is in H_K' and so we may repeat the process to get the iteration formula

$$u_{n+1} = u_n - t_n z_n\ , \qquad n = 0, 1, 2, \ldots, \tag{3.43}$$

where in accordance with (3.9_1) we take $z_n = A'u_n - f_1$. The comparison with (3.37) shows that z_n coincides with the generalized solution of equation

$$Lz_n = Au_n - f. \tag{3.44}$$

Once z_n is determined we then obtain t_n according to (3.9_1) and (3.30)

$$t_n = \frac{[z_n, z_n]'}{[A'z_n, z_n]'} = \frac{[z_n, z_n]'}{[z_n, z_n]}\ , \qquad n = 0, 1, 2, \ldots. \tag{3.45}$$

The following theorem, which is the consequence of Theorem 3.1 as applied to the gradient method in §3.3 (i), proves that the iteration process (3.43)-(3.45) yields the generalized solution of equation (3.28), i.e., u_{n+1} converges to an element u in H_K' such that $[u, v] = (f, Kv)$ for all v in H_K'.

THEOREM 3.2 *If a linear operator A defined on a dense domain $D(A)$ in H has the propecty that there exists a K-p.d. operator L on $D(A)$ and two constants $M > m > 0$ such that the condition* (3.29) *is satisfied, then the sequence $\{u_{n+1}\}$ determined by the* generalized gradient method (3.43)–(3.45) *converges in the H_K'-norm (and in H-norm) to the generalized solution of eq.* (3.28) *at least as fast as the geometric progression with ratio $(M - m)/(M + m)$.*

REMARK 11. In case $K = I$ and A is self-adjoint and positive definite on $D(A)$, similar results were obtained by Kantorovich [7] by means of the variational

principle. Thus, Theorem 3.2 extends the gradient method to K-p.d. operators without the use of the variational principle.

3.6. **Compact computational scheme for a simple iterative method associated with the generalized Ritz method.** In this section we shall be concerned with the problem suggested in Remark 7. Let A be K-p.d. and $\{\phi_i\}$, $i = 1, 2, \ldots$, a linearly independent set in $D(A)$ which is complete in H_K. We indicate here a rather simple iterative method for the approximate solution of eq. (1.1) which coincides in principle but not in form with the generalized Ritz method. The method is essentially based on the generalized orthogonalization process and furnishes the approximate solution $w_n = P_n w = \sum_{i=1}^{n} a_i \phi_i$ without solving the algebraic system (2.5). At the same time we give a compact computational scheme based on this method which is very convenient in practice.

We construct a biorthogonal system $\{e_i\}$ and $\{\tilde{e}_i\}$ from $\{\phi_i\}$ in H_n recursively as follows

$$e_1 = A\bar{\phi}_1, \qquad \tilde{e}_1 = K\bar{\phi}_1, \qquad \bar{\phi}_1 = \phi_1. \tag{3.46}$$

If $\{e_i\}$ and $\{\tilde{e}_i\}$ are already determined for $i = 1, \ldots, k-1$, then we put

$$e_k = A\bar{\phi}_k\,, \qquad \tilde{e}_k = K\bar{\phi}_k, \tag{3.47}$$

where

$$\bar{\phi}_k = \phi_k - \sum_{i=1}^{k-1} \frac{1}{(A\bar{\phi}_i, K\bar{\phi}_i)} (A\phi_k, K\bar{\phi}_i)\,\bar{\phi}_i. \tag{3.48}$$

It is clear that the sequence $\{\bar{\phi}_k\}$ defined by the iteration formula (3.48) is orthogonal in H_K and $\{e_k\}$ and $\{\tilde{e}_k\}$ are biorthogonal in H. In view of the formula (2.2), valid for every u in H_K, the approximate solution w_n in (2.3)–(2.4) is given by

$$w_n = \sum_{i=1}^{n} \frac{(f, K\bar{\phi}_i)}{(A\bar{\phi}_i, K\bar{\phi}_i)}\,\bar{\phi}_i. \tag{3.49}$$

If $w_1 = ((f, K\phi_1)/(A\phi_1, K\phi_1))\ \phi_1$ is the first approximate solution to w, then putting

$$w_k = w_{k-1} + \frac{(f, K\bar{\phi}_k)}{(A\bar{\phi}_k, K\bar{\phi}_k)}\,\bar{\phi}_k \tag{3.50}$$

we obtain an iterative process for the approximate solution of eq. (1.1) defined by formulas (3.48) and (3.50) which because of (3.49) converges to the exact solution w of eq. (1.1).

Computational scheme based on (3.48) and (3.50)([18]). If we put

([18]) In case the operator A is self-adjoint and positive definite and $K = A$ the scheme was employed by Altman [1].

$$\alpha_{ki} = (A\phi_k, K\phi_i), \quad k, i = 1, \ldots, n; \quad \beta_{ki} = (A\phi_k, K\bar{\phi}_i), \qquad i < k;$$

(3.51)

$$\beta_{ii} = (A\bar{\phi}_i, K\bar{\phi}_i), \qquad i = 1, \ldots, n; \qquad \gamma_{ki} = \frac{\beta_{ki}}{\beta_{ii}}, \qquad i < k,$$

and observe that $\beta_{k1} = \alpha_{k1}$ for $k = 1, 2, \ldots, n$, then (3.48) can be written as

$$\phi_k = \sum_{i=1}^{k-1} \gamma_{ki}\bar{\phi}_i + \bar{\phi}_k \tag{3.52}$$

and

$$A\phi_k = \sum_{i=1}^{k-1} \gamma_{ki} A\bar{\phi}_i + A\bar{\phi}_k, \tag{3.53}$$

$$K\phi_k = \sum_{i=1}^{k-1} \gamma_{ki} K\bar{\phi}_i + K\bar{\phi}_k. \tag{3.54}$$

To obtain the recurrent formulas for the consecutive computation of the elements of the triangular matrix γ_{ki} and the elements $\beta_k = (f, K\bar{\phi}_k)$ we note that in view of (3.54) and the definition of β_{ki} we get the recurrent formula

$$\beta_{kj} = (A\phi_k, K\bar{\phi}_j) = (A\phi_k, K\phi_j - \sum_{i=1}^{j-1} \gamma_{ji} K\bar{\phi}_i) = \alpha_{kj} - \sum_{i=1}^{j-1} \bar{\gamma}_{ji}\beta_{ki}, \qquad j < k. \tag{3.55}$$

Similarly, in view of the biorthogonality of $\{A\bar{\phi}_i\}$ and $\{K\bar{\phi}_k\}$, Lemma 1.1 (b), and equalities (3.53) and (3.54) we obtain the recurrent formula for β_{kk}

$$\beta_{kk} = \alpha_{kk} - \sum_{i=1}^{k-1} \bar{\gamma}_{ki}\beta_{ki}. \tag{3.56}$$

We compute now in consecutive order the numbers β_{ki} by means of the recurrent formula (3.55) and then γ_{ki}, $i = 1, 2, \ldots, k-1$, by (3.51) for every k.

To compute the numbers $b_k = (f, K\bar{\phi}_k)$ note that by (3.54)

$$b_k = (f, K\bar{\phi}_k) = (f, K\phi_k) - \sum_{i=1}^{k-1} \bar{\gamma}_{ki} b_i, \qquad b_1 = (f, K\phi_1). \tag{3.57}$$

If we put the approximate solution of eq. (1.1) in the form $w_n = \sum_{k=1}^{n} a_k\phi_k$ and recall that $\gamma_{kk} = 1$ for $k = 1, 2, \ldots, n$ and $\gamma_{ki} = 0$ for $i > k$, then multiplying (3.52) by a_k and summing up from $k = 1$ to $k = n$ we obtain

$$w_n = \sum_{i=1}^{n} \left(\sum_{k=1}^{n} \gamma_{ki} a_k \right) \bar{\phi}_k. \tag{3.58}$$

On the other hand, by (3.50),

$$w_n = \sum_{i=1}^{n} \frac{b_i}{\beta_{ii}} \bar{\phi}_i. \tag{3.59}$$

The comparison of the coefficients in (3.58) and (3.59) yields the following algebraic triangular system for the determination of a_i, $i = 1, \ldots, n.$,

$$\sum_{k=1}^{n} \gamma_{ki} a_k = \frac{b_i}{\beta_{ii}}, \qquad i = 1, 2, \ldots, n. \tag{3.60}$$

Setting $i = n$ in (3.60) we obtain

$$a_n = \frac{b_n}{\beta_{nn}} \tag{3.61}$$

and the following recurrent formula for the determination of a_i

$$a_i = \frac{b_i}{\beta_{ii}} - \sum_{k=i+1}^{n} \gamma_{ki} a_k, \qquad i = n-1, n-2, \ldots, 2, 1. \tag{3.62}$$

Once a_i are found we obtain the approximate solution

$$w_n = \sum_{i=1}^{n} a_i \phi_i.$$

REMARK 12. The iterative method and the computational scheme considered in this section are also valid for both special cases, i.e., for the ordinary Ritz method when $K = I$ and for the method of least squares when $K = A$. However, in the latter case when A is also assumed to be self-adjoint and positive definite a stronger result has been obtained [1].

REFERENCES

1. M. Altman, *A simple practical method and a compact computing scheme for the solution of linear equations in Hilbert space*, Bull. Polska Akad. Nauk **5** (1957), 29–34.

2. F. E. Browder, *Functional analysis and partial differential equations. I**, Math. Ann. **138** (1959), 55–79.

3. R. Courant and D. Hilbert, *Methoden der Mathematischen Physik*, 2nd Ed., Springer, Berlin, 1931.

4. K. Friedrichs, *Spektraltheorie halbbeschränkter Operatoren*, Math. Ann. **109** (1934) 465–487, 685–713.

5. W. Greub and W. Rheinboldt, *On a generalization of an inequality of L. V. Kantorovich*, Proc. Amer. Math. Soc. **10** (1959), 407–415.

6. R. M. Hayes, *Iterative methods for solving linear problems on Hilbert space*, Nat. Bur. Standard. Appl. Math. Ser. **39** (1954), 71–103.

7. L. V. Kantorovich, *Functional analysis and applied mathematics*, Uspehi Mat. Nauk (3) **6** (1948), 89–185.

8. ———, *Approximate solution of functional equations*, Uspehi Mat. Nauk (11) **6** (72), (1956), 99–116.

9. M. V. Keldysh, *On the method of Galerkin for the solution of boundary-value problems*, Izv. Akad. Nauk SSSR Ser. Mat. **6** (1942), 307–330.

10. M. A. Krasnoselsky and S. G. Krein, *Iterative method with minimal residuals*, Mat. Sb. **31 (73)** (1952), 315–334.

11. M. Kravchuk, *Application of the method of moments to the solution of linear differential and integral equations*, Ukrain. Akad. Nauk, Kiev (1932).

12. N. M. Krilov, *Les méthodes de solution approchée des problems de la physique mathématique*, Coll. Mem. de Math., No **49** (1931), Paris.

13. A. D. Lashko, *On the convergence of the methods of Galerkin type*, Dokl. Akad. Nauk SSSR **120** (1958), 242–244.

14. P. D. Lax and A. N. Milgram, *Parabolic equations*, Ann. of Math. **33** (1954), 167–190.

15. A.E. Martyniuk, *On some generalization of the variational method*, Dokl. Akad. Nauk SSSR **117** (1957), 374–377.

16. S. G. Mikhlin, *Direct methods in mathematical physics*, Moskva-Leningrad, 1950.

17. ———, *On Ritz method*, Dokl. Akad. Nauk SSSR **106** (1956), 391–394.

18. F. J. Murray, *The solution of linear operator equations*, J. Math. and Phys. **22** (1943), 148–157.

19. N. I. Polsky, *On the convergence of certain approximate methods of analysis*, Ukrain. Math. J. **7** (1955), 56–70.

20. T. Reid, *Symmetrizable completely continuous linear transformations in Hilbert space*, Duke Math. J. **18** (1951), 41-56.

21. F. Riesz and B. Sz.-Nagy, *Functional analysis*, Ungar, New York, 1955.

22. H. A. Zdanov, *On convergence of some variants of the methods of Galerkin*, Dokl. Akad. Nauk SSSR **115** (1957), 223–225.

COLUMBIA UNIVERSITY,
NEW YORK, NEW YORK