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

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. 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 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
target 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 exten-
sion, which slightly generalize the results of Lax and Milgram [14] are dis-
cussed. 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 investiga-
tion 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 Professor 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-Szego, 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

\begin{equation}
Au = f, \quad f \in H,
\end{equation}

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

\begin{equation}
(Au, Ku) \geq \alpha \| Ku \|^2, \quad u \in D(A).
\end{equation}

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

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

(d) $A$ is closeable.

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

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

$$(A(u + v), K(u + v)) - (A(u - v), K(u - v)) + i(A(u + iv), K(u + iv)) - i(A(u - iv), K(u - iv)) = 4 (Au,Kv)$$

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 $\lambda$

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

is a polynomial of $\lambda$ 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 $\|Au\| \geq \alpha \|Ku\|$ for all $u$ in $D(A)$. Hence
the convergence of $\{Au_n\}$ implies the convergence of $\{Ku_n\}$ and since $\nu_n \to 0$ 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),

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

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 $D(K)$

$$\|Au\| \leq \theta \|Ku\|.$$ 

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

(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 $H$ 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

$$\| Au \| \leq \theta \| u \|_1 = \theta \| Ku \|$$

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

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

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

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

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

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

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, \( \frac{dF(u + tv)}{dt} \big|_{t=0} = 0 \). Using Lemma 1.1 (b) we get \( \frac{dF(w + tv)}{dt} \big|_{t=0} = 2 \Re(Aw - f, Kv) = 0 \). Replacing \( v \) by \( iv \) we obtain \( \Im(Aw - f, Kv) = 0 \). 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 \( \langle u, v \rangle = (Au, Kv) \). 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 \(|Ku| \geq \beta |u|\)

\begin{align}
|u| \geq \gamma_1 |Ku|, & \quad \gamma_1 = \alpha^{1/2} > 0, \ u \in D(A), \\
|u| \geq \gamma_2 |u|, & \quad \gamma_2 = \gamma_1 \beta > 0, \ u \in D(A).
\end{align}

**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 corresponding elements of \(H\).

(c) Since for all \(u \in D(A) \subseteq H_K\) the operator \(K\) satisfies the inequality \(\|Ku\| \leq (1/\gamma) |u|\) 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.

\[
(1.7) \quad w = \sum_{i=1}^{\infty} (f, K\phi_i)\phi_i
\]

converging in \(H_K\) (and in \(H\)).

**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 \(H_K\)

\[
(1.8) \quad (f, Ku) = [w, u].
\]

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

\[
(1.9) \quad F(u) = |u|^2 - [u, w] - [w, u]
\]

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

\[
(1.10) \quad w = \sum_{i=1}^{\infty} [w_1 \phi_i] \phi_i
\]

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

\[
(1.11) \quad w = \sum_{i=1}^{\infty} (f, K\phi_i)\phi_i.
\]

\(\text{(10)}\) 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

$$(1.12) \quad (f, Kv) = [Gf, v]$$

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

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

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), $\|G\| \leq 1/\sqrt{\beta}$.

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), $(A_1u, Kh) = [GA_1u, h] = [u, h] = 0$ 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)

$$(A_1w, Kw) = [w, w] \geq \alpha \|Kw\|^2$$

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 $G_A u = u$ 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 $A_1 \supseteq A$. To prove it, let $v^1$ be an element in $D(A_1)$ and $u$ an element in $D(A)$; then, by (1.12), $[u, G A_1 v^1] = (K u, A_1 v^1) = (A_1 u, K v^1)$ and, since $(A u, K v) = [u, v]$ is valid for all $u \in D(A)$ and $v$ in $H_K$, we have $[u, G A_1 v^1] = [u, v^1]$. This implies that $v^1 = G A_1 v^1$. Thus, $v^1 \in D(A_1)$ and $A_1 v^1 = A v^1$, i.e., $A_1 \supseteq A$.

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 $\eta_1$ and $\eta_2$ two positive constants such that for all $u$ and $v$ in $D(A)$

\[
(1.13) \quad \eta_1 |(Au, Kv)| \leq |(Lu, Kv)| \leq \eta_2 |(Au, Kv)|.
\]

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 $D(A)$

\[
(1.14) \quad (Lu, Kv) = [Su, v].
\]

By the right-hand inequality (1.13) and Lemma 1.1 (c), $[Su, v] \leq \eta_2 |u| |v|$ for all $u \in D(S)$ and $v \in H_K$. This shows that $S$ is bounded. By the left-hand inequality (1.13), $|[Su, u]| \geq \eta_1 |u, u|$. Hence $S$ has also a bounded inverse $S^{-1}$. Therefore, $S$ has a closure $\mathcal{S}$ in $H_K$ such that $R(S)$ is closed, $R(S) = R(\mathcal{S})$, $\mathcal{S}$ is bounded in $H_K$ and has a bounded inverse $\mathcal{S}^{-1}$ on $R(\mathcal{S})$. 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),

\[\text{See, for instance, Friedrichs [4], Mikhlin [16], and Riesz and Sz.-Nagy [21].}\]
\[ [u, v] = 0 \text{ for all } u \in D(A). \text{ Since } D(A) \text{ is dense in } H_K, v = 0. \text{ Thus, } R(S) \text{ is dense and } R(S) = H_K. \]

To construct \( L_1 \) we define the generalized solution of the equation

(1.15) \[ Lu = f \]

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

(1.16) \[ (f, Kv) = [Su, v] \]

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 \( H_K \)

(1.17) \[ (f, Kv) = [Gf, v], \quad |G| \leq \frac{1}{l_1}. \]

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

(1.18) \[ [Su, v] = [Gf, v], \quad v \in H_K. \]

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

(1.19) \[ Su = Gf, \]

whence,

(1.20) \[ u = S^{-1}Gf, \quad G^{-1}Su = f. \]

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

\[ (Lu, Kv) = [Su, v] = [GLu, v]. \]

This shows that \( Su = GLu \), i.e., \( Su \) belongs to \( D(G^{-1}) \) and \( L_1u = G^{-1}Su = Lu \). Thus, \( D(L_1) \supseteq D(L) \). The inverse of \( L_1 \) is \( L_1^{-1} = S^{-1}G \). Since \( G \) is everywhere defined and continuous as a mapping from \( H \) to \( H_K \) and \( S^{-1} \) 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 \( \tilde{S} \) in \( H_K \), \( R(\tilde{S}) = H_K \), and \( \tilde{S}^{-1} \) 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) \( (Lu,Ku) \geq \eta_3(Au,Ku) \), \( \eta_3 > 0 \),
(b) \( \|Bu\|^2 \leq \eta_4(Au,Ku) \), \( \eta_4 > 0 \),

then \( D(L_1) = D(A_1) \) and \( L_1 = A_1 + \tilde{B} \), where \( \tilde{B} \) 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 \( \tilde{B} \) from \( H_K \) to \( H \).

To show that \( L_1' = A_1 + \tilde{B} \) note that since \( H_K' = H_K \) the Remark 2 implies that \( L_1' = L_1 = G^{-1}S \), 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 \( S \) in \( H_K \) has the property that for all \( u \in D(L_1) \) and \( v \in H_K \)

\[
[u,v]' = [u,v] + (\tilde{B}u,Kv) = (L_1u,Kv)
= [GL_1u,v] = (G \cdot G^{-1} Su,v] = [Su,v],
\]

\( S \) has a bounded inverse \( S^{-1} \) on \( R(S) \), and \( R(\tilde{S}) = H_K \). The only statement that is not immediate is that \( R(S) = H_K \) 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 \( H_K \)

\[
[L_1^{-1}f,v] + (\tilde{B}L_1^{-1}f,Kv) = [SL_1^{-1}f,v] = [S \cdot S^{-1}Gf,v] = [Gf,v] = (f,Kv).
\]

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

\[
[L_1^{-1}f,v] = (f - \tilde{B}L_1^{-1}f,Kv) = [A_1^{-1}(f - \tilde{B}L_1^{-1}f),v].
\]
The last equality, being valid for all \( v \) in \( H_K \), implies that
\[
L_1^{-1} f = A_1^{-1}(f - BL_1^{-1} f).
\]
This shows that \( L_1^{-1} f \in D(A_1) \), i.e., \( D(L_1) \subseteq D(A_1) \) and that
\[
L_1u = A_1u + Bu
\]
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 \( D(A_1) \subseteq D(L_1) \). Consequently \( D(A_1) = D(L_1) \) and \( L_1 = A_1 + B \) 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 \( D(A) \)
\[
\begin{align*}
(a) & \quad (Lu, Ku) \geq \eta_3(Au, Ku), \quad \eta_3 > 0, \\
(c) & \quad |(Bu, Kv)| \leq \eta_5 |u| \cdot |v|, \quad \eta_5 > 0,
\end{align*}
\]
then
\[
(1.21) \quad L_1 = A_1(I + T),
\]
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)
\[
|v|^2 = |A_1^{-1}Bu|^2 = (Bu, Kv) \leq \eta_5 |u| \cdot |v|.
\]
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 \( H_K \)
\[
[u, w]' = \lim_i [u_i, w]' = \lim_i (L_{u_i}, Kw).
\]
Since \( A_1 \) is a solvable generalized Friedrichs extension of \( A \),
\[
(Lu_i, Kw) = (A_1u_i, Kw) + (A_1A_1^{-1}Bu_i, Kw) = [(I + T)u_i, w].
\]
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 \( D(L_1) \), \( (L_1u, Kw) = [u, w]' \); whence, by definition of \( A_1 \), \( (I + T)u \) belongs to \( D(A_1) \), i.e., \( D(L_1) \subseteq D(A_1(I + T)) \), and \( L_1u = A_1(I + T)u \). 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., \( D(A_1(I + T)) \subseteq D(L_1) \), 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

\[
0^2 \leq |u|^2 \leq \|Au\|^2 \leq \alpha |u|^2 \leq \alpha^2 |u|^2, \quad u \in D(A),
\]

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 \( \|AK^{-1}u\| \leq \theta \|K^{-1}u\| = \theta \|u\| \) 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

\[
\|Au\| \geq \alpha \|Ku\| \quad \text{and} \quad \|Au\|^2 \geq \alpha |u|^2 \geq \alpha^2 \|Ku\|^2.
\]

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 \( H, (v,K\phi_i) = 0, \ i = 1,2, ..., \) 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 \( w_n = \sum_{i=1}^n c_i \phi_i \) 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 \( n \to \infty \), 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 \( \alpha_1, \alpha_2, ..., \alpha_n \) such that

\[
\|KA^{-1}g - Kw_n\| < \varepsilon/\|AK^{-1}\|,
\]

where \( \varepsilon > 0, v_n = \sum_{i=1}^n \alpha_i \phi_i \) and \( AK^{-1} \) is bounded by Lemma 1.3. Thus,

\[
\|Av_n - g\| = \|AK^{-1}Kv_n - AK^{-1}KA^{-1}g\| \leq \|AK^{-1}\| \|Kv_n - KA^{-1}g\| < \varepsilon.
\]

(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
\[
\left| w - \sum_{i=1}^{n} \alpha_i \phi_i \right| \leq \frac{1}{\gamma_1} \left| Aw - \sum_{i=1}^{n} \alpha_i A \phi_i \right| < \varepsilon.
\]
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
(1.23) \( (Ku, u) \geq \beta_1 \| u \|^2 \), \( u \in H \).

Introduce in \( H \) a new scalar product and the associated norm by
(1.24) \( [u, v]' = (Ku, v), \quad |u|' = \left( [u, u]' \right)^{1/2}, \quad u, v \in H \).

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, \( |A'| \leq |A| \), and \( |A^{-1}'| \leq 1/\alpha \beta_1 \), 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) \( (KAu, v) = (Au, Ku) = (Ku, Av) = (u, KA v) \) 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
(1.25) \( [Au, u]' = (KAu, u) \leq \| A \| (Ku, u) = \| A \| \cdot |u|'^2, \quad u \in H' \).

This shows that \( |A'| \leq |A| \). From (1.23) and the Schwarz inequality \( \| Ku \| \geq \beta_1 \| u \| \), whence \( \beta_1(Ku, u) \leq \| Ku \|^2 \). This and (1.2), considered for our present case, imply that
\[
[Au, u]' = (Au, Ku) \geq \beta_1 |u|'^2, \quad u \in H',
\]
which shows that \( A^{-1} \), as an operator in \( H' \), is symmetric and \( |A'| \leq 1/\alpha \beta_1 \).

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

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

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

$$
(f, Ku) = [w, u]
$$

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

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

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

$$
w_n = P_n w
$$
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

\[
\sum_{i=1}^{n} (A\phi_i, K\phi_j)a^n_i = (f, K\phi_j), \quad j = 1, 2, \ldots, n.
\]

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 \( \phi_i \) are linearly independent in \( H_K \) the determinant \( |a_{ij}| = \left| \left[ (A\phi_i, K\phi_j) \right] \right| = \left| (A\phi_i, K\phi_j) \right| \neq 0 \). 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

\[
\|Aw_n - f\| \leq \theta \|Kw_n - Kw\| \to 0 \quad \text{as} \quad n \to \infty
\]

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 \( Aw_n \to f \) 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

\[ (2.6) \quad Lu = Au + Bu = f, \quad f \in H, \]

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

\[ (2.7) \quad w_n = \sum_{i=1}^{n} a_i^n \phi_i, \]

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 \( a_1^n, \ldots, a_n^n \)

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

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 \( n \geq n_0 \) 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

\[ (2.6_1) \quad u + Tu = f_1, \quad f_1 = A^{-1}f \]

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

\[ (2.9) \quad w = \sum_{i=1}^{\infty} c_i \tilde{\phi}_i, \quad c_i = [w, \tilde{\phi}_i], \]
converging in $H_K$ whose coefficients $c_i$ satisfy the algebraic system

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

i.e., the system

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

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)

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

is uniquely solvable for every $n \geq n_0$ and its solution $\{c_1, c_2, \ldots\}$ converges in the Hilbert space $l_2$ to the solution $\{c_1, c_2, \ldots\}$ of (2.10) (14). This implies that $w^{(n)} = \sum_{i=1}^{n} c_i^\ast \tilde{\phi}_i$ 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 $\{\tilde{\phi}_i\}$ and $\{\tilde{\phi}_j\}$, $\{K\phi_i\}$ and $\{K\phi_j\}$ are connected by a nondegenerate matrix transformation $\phi_i = \sum_{j=1}^{l_i} \beta_{ij} \phi_j$, $\tilde{K}\phi_i = \sum_{j=1}^{l_i} \beta_{ij} K\phi_j$, $\beta_{ij} \neq 0$, it is easy to see that taking $w_n$ in the form (2.7) or in the form $w_n = \sum_{i=1}^{n} d^*_i \phi_i$ 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

$$(2.8_1) \quad d^*_j + \sum_{i=1}^{n} d^*_i \gamma_{ij} = b_j, \quad j = 1, \ldots, n,$$

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

$$\| Lw_n - f \| = \| L(w_n - w) \| = \| LK^{-1} \cdot (w_n - w) \| \leq \| LK^{-1} \| \cdot \| Kw_n - Kw \| \to 0$$

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

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

$$(2.12) \quad \|(Bu, Kv)\| \leq \eta_6 \| (Au, Kv) \|, \quad 0 < \eta_6 < 1.$$ 

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].
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

\[
\|w_n - w\| \leq \frac{1}{(1 - \eta_6)\alpha \beta} \|Lw_n - f\|,
\]

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 \( \langle u, v \rangle = \langle Au, v \rangle \), and the system (2.8) reduces to the system of the well-known Galerkin equations

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

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 \( \langle u, v \rangle = \langle Au, Av \rangle \), and (2.8) reduces to the system of the method of moments

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

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 \( \beta = 0 \), 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

\[
\sum_{i=1}^{n} \{(A\phi_i, A\phi_j)\} a_i^j = (f, A\phi_j), \quad j = 1, 2, \ldots, n.
\]
In this case we have the convergence: $w_n \to w$, $Aw_n \to f$, and the estimate of error $\| w_n - w \| \leq \frac{1}{\gamma} \| Aw_n - f \|$. 

(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

$$Au = f, \quad f \in H,$$

where $A$ is assumed to be a bounded operator in $H$. If $A^*$ denotes the adjoint of $A$, $\pi$ the projection of $H$ on $Z(A)$, and $P$ the projection of $H$ on $R(A^*)$, then $H = Z(A) \oplus R(A^*)$ 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 $R(A^*)$, $u^* = Pu$, such that $A u^* = f$.

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

$$\langle f, \psi_i \rangle = \langle u, A^*\psi_i \rangle, \quad i = 1, 2, \ldots .$$

Let $H_n$ denote the subspace of $H$ spanned by $A^*\psi_i, \ldots , A^*\psi_n$ 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

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

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

$$u_n = P_n u.$$
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

\[
\sum_{i=1}^{n} (A^*\psi_i, A^*\psi_j) a_i^n = a_j, \quad j = 1, \ldots, n,
\]

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

\[
\delta^2 = \min_{v \in H_n} \| u - v \|^2 = \| u - u_n \|^2 = \| u \|^2 - (u_n, u) = \| u \|^2 - \sum_{i=1}^{n} a_i^* a_i \geq 0
\]

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

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 \( \delta^2 = \lim_{n \to \infty} \delta_n^2 \) as \( n \to \infty \). Moreover, if \( u \in \overline{R}(A^*) \), then \( u^* = u, \delta = 0, \) and \( \lim_{n \to \infty} \sum_{i=1}^{n} a_i^* a_i = \| u \|^2 \).

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 \( a_1^n, a_2^n, \ldots, a_n^n \) from the algebraic system (2.19). Assume that there exists a constant \( M > 0 \) such that for each \( n \)

\[
G_n = \sum_{i=1}^{n} a_i^n a_i \leq M.
\]

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 \( a_1^n, a_2^n \) 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.
Theorem 2.3. A necessary and sufficient condition that there exists a \( u \) such that \( Au = f \) is that
(a) \( f \) belong to \( R(A) \).
(b) The sequence \( \{G_n\} = \{ \sum_{i=1}^{n} a_i^* \bar{a}_i \} \), where \( a_i^* \) 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 \( a_i^* = a_j = (f, \psi_j) \) so that \( u_n = \sum_{j=1}^{n} (f, \psi_j) A^* \psi_j \), i.e., \( u_n \) coincides with the \( n \)th 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 \( R(A) \), 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 \( R(A^*) \).

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 occurring 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 that 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

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

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

\[
m(u, u) \leq (Au, u) \leq M(u, u), \quad u \in H.
\]

Let \( H' \) denote the space \( H \) with respect to the new metric

\[
[u, v] = (Au, v), \quad |u| = [u, u]^{1/2}.
\]

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

\[
\sum_{i=1}^{n} [A\psi_i, A\psi_j] a_i^n = [f, \psi_j], \quad j = 1, 2, \ldots, n.
\]

In view of (2.23), the system (2.19₂) can be put in the form

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

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 \( H \)

\[
||Au|| \geq c ||u||.
\]

Let \( H_A \) denote the space \( H \) with respect to the new metric

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

let \( H''_A \) be the subspace spanned by \( \phi_i = A^*\psi_i, \ i = 1, \ldots, n \), and \( P_n'' \) the projection on \( H''_A \). 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

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

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 $H$

$$\langle Au, K^*u \rangle = \langle KAu, u \rangle = c \|u\|^2. \tag{3.1}$$

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

$$\langle KAu, BAu \rangle \geq d \langle KAu, u \rangle. \tag{3.2}$$

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

$$[u, v] = \langle KAu, v \rangle, \quad |u| = [u, u]^{1/2}. \tag{3.3}$$

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 $u \neq 0$

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

**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 $u$

$$d |u|^2 \leq [Lu, u] \leq D |u|^2. \tag{3.5}$$
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 $0 < mI \leq T \leq MI$, then
\[
(u,Tu)(T^{-1}u,u) \leq \frac{(M + m)^2}{4Mm}, \quad u \in H, \ u \neq 0.
\]
To derive (3.41) 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 (3.41).

(ii) **Krasnoselsky-Krein inequality**. If $T$ is self-adjoint and $0 < mI \leq T \leq MI$, then
\[
\frac{(Tu,Tu)}{(Tu,u)^2} \leq \frac{(M + m)^2}{4Mm}, \quad u \in H, \ u \neq 0.
\]
The inequality (3.42) follows immediately from (3.4) if in (3.4) we set $A = I$, $K = I$, and $B = T$ with $c \leq 1$, $d = m$, and $D = M$.

(iii) **Generalized Polya-Szegö inequality**. If $T$ and $S$ are self-adjoint and commutative such that $0 < m_1I \leq T \leq M_1I$ and $0 < m_2I \leq S \leq M_2I$, then
\[
\frac{(Tu,Tu)(Su,Su)}{(Tu,u)^2} \leq \frac{(M_1M_2 + m_1m_2)^2}{4m_1m_2M_1M_2}, \quad u \in H, \ u \neq 0.
\]
To obtain (3.43) 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 (3.43):
\[
\frac{(Tu,Tu)(Su,Su)}{(Tu,u)^2} \leq \frac{(M_1/m_2 + m_1/M_2)^2}{4M_1/m_2 \cdot m_1/M_2} = \frac{(M_1M_2 + m_1m_2)^2}{4m_1m_2M_1M_2}.
\]
We conclude this section by showing that all three special inequalities are equivalent. The equivalence of (3.41) and (3.42) was shown in [5]. Hence, it is sufficient to show, for example, that (3.42) and (3.43) are equivalent.

(3.43) $\rightarrow$ (3.42). This case is trivial since (3.43) reduces to (3.42) when in (3.43) we set $S = I$ and $m_2 = M_2 = 1$.

(15) For the inequality given in Theorem 2 in [5] see (3.43) below.

(16) In case $T$ is a finite matrix, (3.42) was obtained from the spectral resolution by Krasnoselsky and Krein [10].

(17) (3.43), in the finite case, when expressed in terms of the spectral decomposition was stated by Pólya and Szegö [5] and in the general form (3.43) was derived in [5].
If $N = TS^{-1}$, then $0 < m_1/M_2 \leq N \leq M_1/m_2 I$ and by (3.42)
\[
\frac{(Nu, Nu)(u, u)}{(Nu, u)^2} \leq \frac{(M_1/m_2 + m_1/M_2)^2}{4M_1/m_2 + m_1/M_2} = \frac{(M_1M_2 + m_1m_2)^2}{4m_1m_2M_1M_2}.
\]
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 (3.43).

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 $H$

(3.5)
\[
Au = f, \quad f \in H
\]
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 $n$th step of our process. Then the succeeding approximation $u^{n+1}$ is taken in the form

(3.6)
\[
u^{n+1} = u_n - t_n Br_n, \quad n = 0, 1, 2, \ldots.
\]
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

(3.7)
\[
w_{n+1} = w_n - t_n Br_n, \quad n = 0, 1, 2, \ldots
\]
be orthogonal to the vector $Br_n$ in the sense of the $H'$-metric, i.e., $[w_n - t_n Br_n, Br_n] = 0$. This condition yields the following values for $t_n$

(3.8)
\[
t_n = \frac{[w_n, Br_n]}{[Br_n, Br_n]} = \frac{(Kr_n, Br_n)}{(KABr_n, Br_n)}, \quad n = 0, 1, 2, \ldots.
\]
Note that for this value of $t_n$ the vector $t_n Br_n$ 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
The corresponding error and residual vectors are given by the formulas

\[ w_{n+1} = w_n - \frac{(Kr_n, Br_n)}{(KABr_n, Br_n)} Br_n, \quad n = 0, 1, 2, \ldots, \]

\[ r_{n+1} = r_n - \frac{(Kr_n, Br_n)}{(KABr_n, Br_n)} ABr_n, \quad n = 0, 1, \ldots. \]

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

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

(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

\[ |w_{n+1}| \leq \left( 1 - \frac{4dD}{(D + d)^2} \right)^{1/2} |w_n| = \frac{D - d}{D + d} |w_n|, \]

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

\[ |w_{n+1}| \leq \left( \frac{D - d}{D + d} \right)^n |w_0|. \]

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 \( |w_{n+1}| \rightarrow 0 \) monotonically as \( n \rightarrow \infty \).

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

\[(3.1)\]
\[(Au, u) \geq c_1 \| u \|^2 , \quad c_1 > 0 \]

\[(3.2)\]
\[(Au, Au) \geq d_1 (Au, u) , \quad d_1 > 0 \]

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

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

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

\[(3.9)\]
\[u_{n+1} = u_n - \frac{(r_n, r_n)}{(Ar_n, r_n)} r_n , \quad n = 0, 1, 2, \ldots .\]

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

\[(3.14)\]
\[\| u_{n+1} - u \| \leq \frac{\theta_1}{c_1} \left( \frac{D_1 - d_1}{D_1 + d_1} \right)^n .\]

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

\[(3.15)\]
\[(r_n, w_n) - \frac{(r_n, r_n)^2}{(Ar_n, r_n)} = \left| w_n \right|^2 - \frac{[w_n, r_n]^2}{r_n^2} \geq 0\]

and the fact obtained from (3.9):

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

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) [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) 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 \( c_1 \| u \|^2 \leq (Au, u) \leq C_1 \| u \|^2 \), then Kantorovich [7], using the spectral resolution of \( A \), derived the estimate

\[\| u_{n+1} - u \| \leq \frac{\theta_1}{c_1} \left( \frac{C_1 - c_1}{C_1 + c_1} \right)^n .\]
considered the best available in case \( u_{n+1} \) is given by (3.9). To prove our statement observe that from the definition of \( d_i \) in (3.24) it follows easily that \( c_i \leq d_i \).

On the other hand, by Lemma 1.5, \( D_i \leq C_i \). This implies that \( C_i/c_i \geq D_i/d_i > 1 \). Since the function \( g(x) = x/(x^2 + 1) \) is monotonically decreasing for \( x \geq 1 \) and \( C_i/c_i \geq D_i/d_i > 1 \), it follows that

\[
\frac{C_1 c_1}{C_1^2 + c_1^2} \leq \frac{D_1 d_1}{D_1^2 + d_1^2}
\]

from which one derives that

\[
(3.16) \quad \frac{D_1 - d}{D_1 + d} \leq \frac{C_1 - c_1}{C_1 + c_1}.
\]

In fact,

\[
\frac{C_1 c_1}{C_1^2 + c_1^2} \leq \frac{D_1 d_1}{D_1^2 + d_1^2}
\]

implies that

\[
2 + \left( \frac{C_1^2 + c_1^2}{C_1 c_1} \right) \geq 2 + \left( \frac{D_1^2 + d_1^2}{D_1 d_1} \right),
\]

or equivalently that

\[
\frac{(C_1 + c_1)^2}{C_1 c_1} \geq \frac{(D_1 + d_1)^2}{D_1 d_1}.
\]

This in turn implies that

\[
1 - \frac{4C_1 c_1}{(C_1 + c_1)^2} \geq 1 - \frac{4D_1 d_1}{(D_1 + d_1)^2} \quad \text{or} \quad \frac{(C_1 - c_1)^2}{(C_1 + c_1)^2} \geq \frac{(D_1 - d_1)^2}{(D_1 + d_1)^2}
\]

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

\[
(3.12) \quad \| Au \|_2^2 \leq c_2 \| u \|_2^2,
\]

\[
(3.22) \quad \| A^* Au \|_2^2 \leq d_2 \| Au \|_2^2,
\]

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

\[
(3.32) \quad [u,v] = (Au,Av), \quad |u| = \| Au \|,
\]

and the iteration formula (3.9) reduces to the formula

\[
(3.92) \quad u_{n+1} = u_n - \frac{\| A^* r_n \|_2^2}{\| AA^* r_n \|_2^2} \cdot A^* r_n, \quad n = 0,1,\ldots
\]
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

\[
\| Au \|^2 \geq c_3 \| u \|^2,
\]
\[
(A^2 u, Au) \geq d_3 \| Au \|^2
\]
and the space \( H' \) becomes the space \( H'_1 \) with the metric

\[
[u,v] = (Au,Av), \quad |u| = \| Au \|.
\]
The RME-method (3.9) takes in this event the form

\[
u_{n+1} = u_n - \frac{(Ar_n,r_n)}{(Ar_n,Ar_n)} r_n, \quad n = 0, 1, 2, \ldots \]

while (3.12) reduces to

\[
|w_{n+1}|^2 = |r_{n+1}|^2 = |r_n|^2 - \frac{(Ar_n,r_n)^2}{\| Ar_n \|^2}
\]
showing that the residual \( r_{n+1} \) is obtained from \( r_n \) by subtracting from it the
projection of \( r_n \) on \( Ar_n \). (3.123) shows that in the algorithm (3.93) 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 (3.93)
converges also monotonically in the \( H \)-norm. In fact, by (3.93)

\[
\| w_{n+1} \|^2 = \| w_n \|^2 - 2 \frac{(Ar_n,r_n)}{(Ar_n,Ar_n)} \left[ (r_n,w_n) - \frac{1}{2} \frac{(Ar_n,r_n)}{(Ar_n,Ar_n)} (r_n,r_n) \right].
\]
Thus, the proof of monotonic convergence is equivalent to the proof that

\[
(A^{-1}r_n,r_n)(Ar_n,Ar_n) \geq (r_n,r_n)(Ar_n,r_n).
\]
The latter follows from Schwarz inequality

\[
(Ar_n,r_n)^2 (r_n,r_n)^2 = (Ar_n,r_n)^2 (A^{1/2}r_n,(A^{-1})^{1/2}r_n)^2
\leq \| Ar_n \|^2 \| r_n \|^2 \| A^{1/2}r_n \|^2 \| (A^{-1})^{1/2}r_n \|^2
= (Ar_n,Ar_n)(r_n,r_n)(Ar_n,r_n)(A^{-1}r_n,r_n).
\]
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

$$(3.2_1) \quad \| A u \|^2 \geq d_4 \| u \|^2.$$  

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

$$(3.9_1) \quad u_{n+1} = u_n - \frac{\| r_n \|^2}{\| A^* r_n \|^2} A^* r_n, \quad n = 0, 1, 2, \ldots,$$

while (3.12) becomes

$$(3.12_1) \quad \| w_{n+1} \|^2 = \| w_n \|^2 - \frac{\| r_n \|^4}{\| A^* r_n \|^4}.$$  

The formula (3.9_1) shows that the error vector $w_{n+1}$ is obtained by subtracting from $w_n$ the projection of $w_n$ on $A^* A w_n$. By (3.12_1) 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

$$(3.14_1) \quad \| u_{n+1} - u \| \leq \theta_4 \left( \frac{D_4 - d_4}{D_4 + d_4} \right)^n, \quad \theta_4 = \| u_0 - u \|,$$

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 gradient 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.
1962] SOLUTION OF LINEAR OPERATOR EQUATIONS 167

(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 \( n \)th 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) and (3.91) we get the structurally similar formulas

\[
\begin{align*}
\tag{3.18} w_{n+1} &= w_n - \frac{(Ar_n, r_n)}{(Ar_n, Ar_n)} r_n, \quad n = 0, 1, \ldots \\
\tag{3.19} \tilde{w}_{n+1} &= w_n - \frac{(r_n, r_n)}{(Ar_n, r_n)} r_n, \quad n = 0, 1, \ldots .
\end{align*}
\]

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

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

The last equality can be written in the following convenient form

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

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

\[
\frac{(r_n, r_n)}{(Ar_n, r_n)} \leq \frac{(Ar_n, r_n)}{\| Ar_n \|^2}
\]

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

(b) If \( r_n = Ar_n \) is the residual vector at the \( n \)th 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

\[
\begin{align*}
\tag{3.21} r_{n+1} &= r_n - \frac{(Ar_n, r_n)}{(Ar_n, Ar_n)} Ar_n, \quad n = 0, 1, \ldots , \\
\tag{3.22} \tilde{r}_{n+1} &= r_n - \frac{(r_n, r_n)}{(Ar_n, r_n)} Ar_n, \quad n = 0, 1, \ldots .
\end{align*}
\]

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 $\bar{t}_n$, then from above we get
\[
\|\tilde{r}_{n+1}\|^2 - \|r_{n+1}\|^2 = \left(\frac{\|r_n\|^2}{(\bar{A}r_n, r_n)} - \frac{(\bar{A}r_n, r_n)}{\|A^*r_n\|^2}\right)^2 \|Ar_n\|^2 = (\bar{t}_n - t_n)^2 \|Ar_n\|^2.
\]
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 $(\bar{t}_n - t_n)^2 \|Ar_n\|^2$.

**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 (3.9a) and (3.9)
\[
(3.23) \quad w'_{n+1} = w_n - \frac{\|r_n\|^2}{\|A^*r_n\|^2} A^*r_n, \quad n = 0, 1, \ldots,
\]
\[
(3.24) \quad \tilde{w}'_{n+1} = w_n - \frac{\|A^*r_n\|^2}{\|AA^*r_n\|^2} A^*r_n, \quad n = 0, 1, \ldots.
\]
From (3.23) and (3.24) we immediately obtain
\[
\|\tilde{w}'_{n+1}\|^2 - \|w'_{n+1}\|^2 = \frac{\|A^*r_n\|^6}{\|AA^*r_n\|^4} - 2 \left(\frac{\|A^*r_n\|^2}{\|AA^*r_n\|^2}\right)^2 \|A^*r_n\|^2 + \frac{\|r_n\|^4}{\|A^*r_n\|^2}.
\]
Denoting the coefficient in (3.23) by $t'_n$ and in (3.24) by $\bar{t}'_n$, then from above we get
\[
\|\tilde{w}'_{n+1}\|^2 - \|w'_{n+1}\|^2 = \left(\frac{\|A^*r_n\|^2}{\|AA^*r_n\|^2} - \frac{\|r_n\|^2}{\|A^*r_n\|^2}\right)^2 \|A^*r_n\|^2 = (t'_n - \bar{t}'_n)^2 \|A^*r_n\|^2.
\]
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 $(t'_n - \bar{t}'_n)^2 \|A^*r_n\|^2$.

(b) Using (2.23) and (3.24) we obtain
\[
(3.25) \quad r'_{n+1} = r_n - \frac{\|r_n\|^2}{\|A^*r_n\|^2} AA^*r_n, \quad n = 0, 1, \ldots,
\]
\[
(3.25') \quad \tilde{r}'_{n+1} = r_n - \frac{\|A^*r_n\|^2}{\|AA^*r_n\|^2} AA^*r_n, \quad n = 0, 1, \ldots,
\]
from which we derive that

\[ \| r_{n+1} \|^2 - \| r_n \|^2 = \frac{\| r_n \|^4}{\| A^* r_n \|^4} \| A A^* r_n \|^2 - 2 \| A^* r_n \|^2 + \frac{\| A^* r_n \|^4}{\| A A^* r_n \|^4} \]

or equivalently

\[ \| r_{n+1} \|^2 - \| r_n \|^2 = \left( \frac{\| r_n \|^2}{\| A^* r_n \|^2} - \frac{\| A^* r_n \|^2}{\| A A^* r_n \|^2} \right)^2 \| A A^* r_n \|^2 \]

\[ = (t_n' - t_n)^2 \| A A^* r_n \|^2. \]

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 \( (t_n' - t_n)^2 \| A A^* r_n \|^2 \).

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 \( D(A) \)

\[ (3.26) \quad (L u, K u) \geq \xi \| K u \|^2. \]

Denote by \( H_K' \) the completion of \( D(A) \) in the metric

\[ (3.27) \quad [u,v]' = (L u, K v), \quad \| u \|' = \left\{ [u,u]' \right\}^{1/2}. \]

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

\[ (3.28) \quad A u = f, \quad f \in H, \]

where \( A \) has the property that there exist constants \( M > m > 0 \) such that

\[ (3.29) \quad m(L u, K u) \leq (A u, K u) \leq M(L u, K u), \quad u \in D(A). \]

The inequalities (3.26) and (3.29) imply that \( A \) is K.p.d. and that the norm induced by \( [u,u'] = (A u, K u) \) is equivalent to the norm induced by \( [u,u']' = (L u, K u) \). 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 \( (A u, K v) \) 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., \( [u,v] = (A u, K v) \) 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 \( u, v \in H_K' \)
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)

\[
(3.31) \quad m[u,u]' \leq [A'u,u]' \leq M[u,u]', \quad u \in H'_K.
\]

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

\[
(3.32) \quad Lu = Aw.
\]

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

\[
(3.33) \quad [u,v]' = [w,v].
\]

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

\[
(3.34) \quad [u,v]' = [A'w,v]',
\]

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

\[
(3.35) \quad u = A'w, \quad u \in H'_K.
\]

Let us also observe that the equation

\[
(3.36) \quad Lu = Aw - f, \quad f \in H,
\]

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

\[
(3.37) \quad u = A'w - f_1.
\]

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

\[
(3.38) \quad [A'u,v]' = [f_1,v]',
\]

which in turn is equivalent to the equation

\[
(3.39) \quad A'u = f_1 , \quad f_1 \in H'_K.
\]

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
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 \( z_0 = A'u_0 - f \), then according to (3.9) the next iterant \( u_1 \) is given by

\[
(3.40) \quad u_1 = u_0 - t_0 z_0,
\]

where

\[
(3.41) \quad t_0 = \frac{[z_0, z_0]' - [z_0, z_0]'}{[z_0, z_0]'} = \frac{[z_0, z_0]'}{[z_0, z_0]}.
\]

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

\[
(3.42) \quad Lz_0 = Au_0 - f = r_0.
\]

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

\[
(3.43) \quad u_{n+1} = u_n - t_n z_n, \quad n = 0, 1, 2, \ldots,
\]

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

\[
(3.44) \quad Lz_n = Au_n - f.
\]

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

\[
(3.45) \quad t_n = \frac{[z_n, z_n]' - [z_n, z_n]'}{[z_n, z_n]'} = \frac{[z_n, z_n]'}{[z_n, z_n]}, \quad n = 0, 1, 2, \ldots.
\]

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 property 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

\[
e_1 = A\phi_1, \quad \tilde{e}_1 = K\bar{\phi}_1, \quad \bar{\phi}_1 = \phi_1.
\]

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

\[
e_k = A\phi_k, \quad \tilde{e}_k = K\bar{\phi}_k,
\]

where

\[
\bar{\phi}_k = \phi_k - \sum_{i=1}^{k-1} \frac{1}{(A\phi_i, K\bar{\phi}_i)} (A\phi_k, K\bar{\phi}_i) \bar{\phi}_i.
\]

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

\[
w_n = \sum_{i=1}^{n} (f, K\bar{\phi}_i) \bar{\phi}_i.
\]

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

\[
w_k = w_{k-1} + \frac{(f, K\bar{\phi}_k)}{(A\phi_k, K\bar{\phi}_k)} \bar{\phi}_k
\]

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

\[(18)\text{ In case the operator } A \text{ is self-adjoint and positive definite and } K = A \text{ the scheme was employed by Altman [1].}\]
\[
\alpha_{ki} = (A\phi_k, K\phi_i), \quad k, i = 1, \ldots, n; \quad \beta_{ki} = (A\phi_k, K\phi_i), \quad i < k;
\]
(3.51)
\[
\beta_{ii} = (A\phi_i, K\phi_i), \quad i = 1, \ldots, n; \quad \gamma_{ki} = \frac{\beta_{ki}}{\beta_{ii}}, \quad i < k,
\]
and observe that \(\beta_{ki} = \alpha_{ki}\) for \(k = 1, 2, \ldots, n\), then (3.48) can be written as
\[
\phi_k = \sum_{i=1}^{k-1} \gamma_{ki} \phi_i + \phi_k
\]
(3.52)
and
\[
A\phi_k = \sum_{i=1}^{k-1} \gamma_{ki} A\phi_i + A\phi_k,
\]
(3.53)
\[
K\phi_k = \sum_{i=1}^{k-1} \gamma_{ki} K\phi_i + K\phi_k.
\]
(3.54)

To obtain the recurrent formulas for the consecutive computation of the elements of the triangular matrix \(\gamma_{ki}\) and the elements \(\beta_{ki} = (f, K\phi_k)\) we note that in view of (3.54) and the definition of \(\beta_{ki}\) we get the recurrent formula
\[
\beta_{kj} = (A\phi_k, K\phi_j) = (A\phi_k, K\phi_j - \sum_{i=1}^{j-1} \gamma_{ji} K\phi_i) = \alpha_{kj} - \sum_{i=1}^{j-1} \gamma_{ji} \beta_{ki}, \quad j < k.
\]
(3.55)

Similarly, in view of the biorthogonality of \(\{A\phi_i\}\) and \(\{K\phi_k\}\), Lemma 1.1 (b), and equalities (3.53) and (3.54) we obtain the recurrent formula for \(\beta_{kk}\)
\[
\beta_{kk} = \alpha_{kk} - \sum_{i=1}^{k-1} \gamma_{ki} \beta_{ki}.
\]
(3.56)

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

To compute the numbers \(b_k = (f, K\phi_k)\) note that by (3.54)
\[
b_k = (f, K\phi_k) = (f, K\phi_k) - \sum_{i=1}^{k-1} \gamma_{ki} b_i, \quad b_1 = (f, K\phi_1).
\]
(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
\[
w_n = \sum_{i=1}^{n} \left( \sum_{k=1}^{n} \gamma_{ki} a_k \right) \phi_i.
\]
(3.58)
On the other hand, by (3.50),
\[
w_n = \sum_{i=1}^{n} \frac{b_i}{\beta_{ii}} \phi_i.
\]
(3.59)
The comparison of the coefficients in (3.58) and (3.59) yields the following algebraic triangular system for the determination of $a_i$, $i = 1, \ldots, n$,

\begin{equation}
\sum_{k=1}^{n} \gamma_{ki}a_k = \frac{b_i}{\beta_{ii}}, \quad i = 1, 2, \ldots, n.
\end{equation}

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

\begin{equation}
a_n = \frac{b_n}{\beta_{nn}}
\end{equation}

and the following recurrent formula for the determination of $a_i$

\begin{equation}
a_i = \frac{b_i}{\beta_{ii}} - \sum_{k=i+1}^{n} \gamma_{ki}a_k, \quad i = n - 1, n - 2, \ldots, 2, 1.
\end{equation}

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**


