The fundamental properties of symmetry and positive definiteness of linear operators of mechanics have been seen in Chapter 9 to yield desirable properties of eigenfunctions. These properties make eigenfunction expansion methods broadly applicable. In this chapter, even more broadly applicable methods of treating elliptic boundary-value problems are introduced, called variational methods. These methods are shown to yield theoretical and computational results that form the foundation for most modern computational methods in solid mechanics.

10.1 ENERGY CONVERGENCE

In this section, a new measure of closeness of two functions is introduced. Convergence associated with this measure of closeness plays an important role in the variational theory developed in this chapter.

Consider a positive definite linear operator \( A \), with domain of definition \( D_A \). If \( A \) is a differential operator, then \( D_A \) will be a dense set of functions in \( L_2 \) that are continuous and have continuous derivatives in a closed set \( \Omega = \Omega + \Gamma \). If \( A \) is a differential operator of order \( k \), then \( D_A \) will consist of functions that, together with their derivatives of order \( k - 1 \), are continuous in \( \Omega \) and whose derivatives of order \( k \) exist, except at a finite number of points, curves, or surfaces, and have finite norm in the open domain \( \Omega \). It follows that the functions \( Au \) for \( u \) in \( D_A \) also have finite norm. In addition to the above continuity assumptions, the functions in \( D_A \) must satisfy certain boundary conditions, which for the present will be assumed homogeneous.

### Energy Scalar Product and Norm

**Definition 10.1.1.** If \( A \) is a positive definite linear operator, the quantity \( (Au, v) \) defines a scalar product on \( D_A \), called the **energy scalar product**, denoted by

\[
[ u, v ]_A \equiv (Au, v) = \int_\Omega vAu \, d\Omega
\]  

(10.1.1)
To show that Eq. 10.1.1 defines a scalar product, note first that for any real $\alpha$,

$$[\alpha u, v]_A = (\alpha u, v) = (\alpha u, v) = \alpha [u, v]_A.$$  

Next, since $A$ is symmetric, 

$$[u, v]_A = (Au, v) = (Av, u) = [v, u]_A.$$  

Finally, since $A$ is positive definite, 

$$[u, u]_A = (Au, u) > 0$$  

if $u \neq 0$.

If $D_A$ is interpreted as a space of admissible displacements, then for $u$ and $v$ in $D_A$, $Au$ may be interpreted as the force that is required to produce the displacement $u$ and the scalar product $(Au, v)$ becomes the work done by the force $Au$ acting through the displacement $v$. The quantity $(Au, u)$ is proportional to the strain energy associated with the displacement $u$.

**Definition 10.1.2.** The quantity $\|u\|_A = (Au, u)^{1/2}$ is called the energy norm of the function $u$.

Let $u$ and $v$ be two functions in $D_A$. A measure of their closeness is the square root of the energy of their difference; i.e.,

$$\|u - v\|_A = \sqrt{(A(u - v), u - v)}$$  

(10.1.2)

### Energy Convergence

The formulation of the energy measure of Eq. 10.1.2 of the distance between two functions in $D_A$ leads to a new concept of convergence of functions in $D_A$.

**Definition 10.1.3.** Let $u$ and $u_n$ be in $D_A$, $n = 1, 2, \ldots$, for a positive definite operator $A$. The sequence $u_n$ is said to **converge in energy** to $u$ if

$$\lim_{n \to \infty} \|u_n - u\|_A = 0.$$  

Symbolically, convergence in energy will be denoted by $u_n(x) \overset{E}{\to} u(x)$.

Note that the concept of energy convergence can be defined only after the corresponding positive definite operator has been defined.

**Theorem 10.1.1.** If the operator $A$ is positive bounded below and $u_n \to^E u$, then

$$\lim_{n \to \infty} \|u_n - u\| = 0,$$  

denoted $u_n \to^\text{mean} u$.

To prove this theorem, note that since $A$ is positive bounded below, there exists a positive constant $\gamma$ such that $(Au, u) \geq \gamma \|u\|^2$, for all $u$ in $D_A$. It follows therefore that 

$$\|u\|^2 \leq (1/\gamma) \|u\|_A^2$$  

and 

$$\|u_n - u\|^2 \leq (1/\gamma) \|u_n - u\|_A^2.$$  

Thus, if $\|u_n - u\|_A \to 0$, then 

$$\|u_n - u\| \to 0.$$  

Example 10.1.1

Consider the operator defined by
\[
Au = - \frac{d^2u}{dx^2}
\]  
(10.1.3)

\[D_A = \{ u \in C^2(0, 1): u(0) = u(1) = 0 \}\]

Then,
\[
\| u \|_A^2 = (Au, u) = - \int_0^1 \left( \frac{d^2u}{dx^2} \right) u \, dx = \int_0^1 \left( \frac{du}{dx} \right)^2 \, dx \quad (10.1.4)
\]

The statement \( u_n \to u \) is thus equivalent to
\[
\lim_{n \to \infty} \int_0^1 \left[ u_n'(x) - u'(x) \right]^2 \, dx = 0
\]

It was shown in Example 9.2.8 that \( A \) is positive bounded below, so it follows that, for all \( u \) and \( u_n \) in \( D_A \),
\[
\lim_{n \to \infty} \int_0^1 \left[ u_n'(x) - u'(x) \right]^2 \, dx = 0
\]

implies that
\[
\lim_{n \to \infty} \int_0^1 \left[ u_n(x) - u(x) \right]^2 \, dx = 0
\]

Example 10.1.2

Consider the Laplace operator \( Au = - \nabla^2 u \) with domain \( D_A = \{ u \in C^2(\Omega): u = 0 \text{ on } \Gamma \} \). It was shown in Example 9.2.8 that this operator is positive bounded below. The energy of a function \( u \) in \( D_A \) is thus
\[
\| u \|_A^2 = ( - \nabla^2 u, u ) = \iint_{\Omega} \sum_{i=1}^m \left( \frac{\partial u}{\partial x_i} \right)^2 \, d\Omega
\]

If \( u_n \to u \), for this operator,
\[
\lim_{n \to \infty} \iint_{\Omega} \sum_{i=1}^m \left( \frac{\partial u_n}{\partial x_i} - \frac{\partial u}{\partial x_i} \right)^2 \, d\Omega = 0
\]
Hence, it follows that

\[
\lim_{n \to \infty} \iint_{\Omega} \left( \frac{\partial u_n}{\partial x_i} - \frac{\partial u}{\partial x_i} \right)^2 \, d\Omega = 0, \quad i = 1, 2, \ldots, m
\]

Thus, energy convergence implies that each of the first derivatives of the functions \( u_n \) converge in the \( L_2 \) norm to the first derivative of the function \( u(x) \). It also implies that \( u_n \) converges to \( u \) in the mean.

**Definition 10.1.4.** A sequence of functions \( \{ \phi_i \} \) in \( D_A \) is said to be **energy orthogonal** if \( \{ \phi_i, \phi_j \}_A = 0, \ i \neq j \), and energy orthonormal if \( \{ \phi_i, \phi_j \}_A = \delta_{ij} \). An energy orthonormal sequence of functions \( \{ \phi_i \} \) is said to be **complete in energy** if the **Parseval relation**

\[
\| u \|_A^2 = \sum_{i=1}^{\infty} [\phi_i, u]_A^2 \tag{10.1.5}
\]

holds for every function \( u \) in \( D_A \). This is equivalent to the property that given any \( \epsilon > 0 \), there exists an integer \( N \) and constants \( \alpha_1, \alpha_2, \ldots, \alpha_N \) such that

\[
\left\| u - \sum_{i=1}^{N} \alpha_i \phi_i \right\|_A < \epsilon \tag{10.1.6}
\]

**Example 10.1.3**

Let \( A \) be the operator of Example 10.1.1 on \( D_A = \{ u \in C^2(0, \pi) : u(0) = u(\pi) = 0 \} \). Then, for \( u \) and \( v \) in \( D_A \),

\[
[ u, v ]_A = -\int_{0}^{\pi} v \frac{d^2 u}{dx^2} \, dx = \int_{0}^{\pi} \frac{du}{dx} \frac{dv}{dx} \, dx \tag{10.1.7}
\]

and

\[
\| u \|_A^2 = \int_{0}^{\pi} \left( \frac{du}{dx} \right)^2 \, dx \tag{10.1.8}
\]

Energy orthogonality of the functions in this example is thus equivalent to \( L_2 \) orthogonality of their first derivatives.

It can be shown that the sequence of functions \( \phi_n(x) = \frac{1}{n} \sqrt{\frac{2}{\pi}} \sin nx, \ n = 1, 2, \ldots, \) is energy orthonormal and complete in \( D_A \). Orthonormality follows from
For energy completeness, since the sequence 1, cos x, cos 2x, ... is complete in the sense of convergence in $L_2(0, \pi)$, any function $\psi(x)$ with finite norm can be expressed as

$$\psi(x) = b_0 + \sum_{n=1}^{\infty} b_n \cos nx$$

which converges in the $L_2$ norm. Setting $\psi(x) = u'(x)$, for $u$ in $D_A$,

$$u(0) = u(\pi) = 0$$

Thus,

$$b_0 = \frac{1}{\pi} \int_0^\pi \psi(x) \, dx = \frac{1}{\pi} \int_0^\pi u'(x) \, dx = \frac{1}{\pi} [u(\pi) - u(0)] = 0$$

Hence,

$$u'(x) = \sum_{n=1}^{\infty} b_n \cos nx = \sum_{n=1}^{\infty} b_n^* \phi_n'(x)$$

where $b_n^* = \sqrt{\frac{\pi}{2}} b_n$ and $\phi_n'(x) = \sqrt{\frac{2}{\pi}} \cos nx$. This shows that the derivative of any function in $D_A$ can be approximated in the mean with any degree of accuracy by linear combinations of the derivatives of the $\phi_n(x)$. This result and Example 10.1.1 show that the set $\phi_n(x) = \frac{1}{n} \sqrt{\frac{2}{\pi}} \sin nx$ is complete in energy.

EXERCISES 10.1

1. Construct the energy norm associated with the following positive definite operators, using integration by parts to minimize the highest order derivatives that appear:

(a) $A_1 u = -\frac{d}{dx} \left( a(x) \frac{du}{dx} \right)$, $a(x) > 0$ on $(0, \ell)$

$D_{A_1} = \{ u \in C^2(0, \ell): u(0) = u(\ell) = 0 \}$
438 Chap. 10 Variational Methods for Boundary-Value Problems

(b) \[ A_2 u = \frac{d^2}{dx^2} \left[ EI(x) \frac{d^2 u}{dx^2} \right], \quad EI(x) > 0 \text{ on } (0, \ell) \]

\[ D_{A_2} = \{ u \in C^4(0, \ell): u(0) = u'(0) = u(\ell) = u'(-\ell) = 0 \} \]

(c) \[ A_3 u = \nabla^2 (b(x) \nabla^2 u), \quad b(x) > 0 \text{ on } \Omega \]

\[ D_{A_3} = \left\{ u \in C^4(\Omega): u = \frac{\partial u}{\partial n} = 0 \text{ on } \Gamma \right\} \]

2. The fourth order operator \( A_2 \) of Exercise 1 (b) can be reformulated as the following system of two second order equations:

\[ Lz = \begin{bmatrix} -z_2'' & z_1'' \frac{z_2}{EI} \\ -z_1'' - \frac{z_2}{EI} & 0 \end{bmatrix} = \begin{bmatrix} q(x) \\ 0 \end{bmatrix} \]

where \( z_1 \equiv u \) and \( z_2 = -EIz_1'' \). For a simply supported beam, the domain of this operator is

\[ D_L = \{ z_1, z_2 \in C^2(0, \ell): z_1(0) = z_2(0) = z_1(\ell) = z_2(\ell) = 0 \} \]

Show that this operator is symmetric, but not positive definite.

10.2 THE MINIMUM FUNCTIONAL THEOREM

The concepts of positive definite operators and energy norms were used in the 1940's, primarily by Soviet mathematicians, to create the modern foundation for variational methods in mechanics and the mathematical theory of functional analysis. The penetrating and inspiring treatment of this subject in Ref. 21 remains as a landmark in the field.

Let \( A \) be a positive definite linear operator and consider the problem of finding a solution \( u \) of

\[ Au = f \quad (10.2.1) \]

By this is meant, find a function \( u \) in \( D_A \) that satisfies Eq. 10.2.1. The fact that \( u \) is in \( D_A \) implies that it satisfies the boundary conditions of the problem.

Minimum Principles for Operator Equations

Theorem 10.2.1 (Minimum Functional Theorem). Let \( A \) be a positive definite linear operator with domain \( D_A \) that is dense in \( L^2(\Omega) \) and let Eq. 10.2.1 have a solution. Then the solution is unique and minimizes the energy functional

\[ F(u) = (Au, u) - 2(u, f) = \int_\Omega (uAu - 2uf) \, d\Omega \quad (10.2.2) \]
over all $u$ in $D_A$. Conversely, if there exists a function $u$ in $D_A$ that minimizes the functional of Eq. 10.2.2, then it is the unique solution of Eq. 10.2.1.

To show that a solution of Eq. 10.2.1 is unique, assume there are two solutions $u$ and $v$ in $D_A$. Then, $u - v$ satisfies $A(u - v) = 0$. Thus,

$$ (A(u - v), (u - v)) = 0 $$

Since $A$ is positive definite, $u - v = 0$. Thus, if a solution of Eq. 10.2.1 exists, it is unique.

Next let $u_0$ in $D_A$ be the solution to Eq. 10.2.1, so that $Au_0 = f$. Making this substitution for $f$ in Eq. 10.2.2 yields

$$ F(u) = (Au, u) - 2(Au_0, u) $$

$$ = [u, u]_A - 2[u_0, u]_A $$

$$ = [u - u_0, u - u_0]_A - [u_0, u_0]_A $$

$$ = \|u - u_0\|^2_A - \|u_0\|^2_A $$

It is clear that $F(u)$ assumes its minimum value if and only if $u = u_0$, proving the first part of Theorem 10.2.1.

Now suppose that there exists a function $u_0$ in $D_A$ that minimizes the functional $F$ of Eq. 10.2.2. Let $v(x)$ be an arbitrary function from $D_A$ and let $\alpha$ be an arbitrary real number. Then, $F(u_0 + \alpha v) - F(u_0) \geq 0$. Using symmetry of the operator $A$, the function

$$ F(u_0 + \alpha v) - F(u_0) = 2\alpha(Au_0 - f, v) + \alpha^2(Av, v) \geq 0 $$

of $\alpha$ takes on its minimum value of zero at $\alpha = 0$. Thus, its derivative with respect to $\alpha$ at $\alpha = 0$ must be zero; i.e., $2(Au_0 - f, v) = 0$, for all $v$ in $D_A$. Since $D_A$ is dense in $L^2(\Omega)$, it follows that $Au_0 - f = 0$.

As will become evident from later examples, the functional $F(u)$ is proportional to the total potential energy of the system under consideration. In such cases, Theorem 10.2.1 provides a rigorous proof of the principle of minimum total potential energy and allows the problem of integrating a differential equation, under specified boundary conditions, to be replaced by the problem of seeking a function that minimizes the functional of Eq. 10.2.2. The direct methods that will be discussed later allow approximate solutions of the later problem to be found by relatively simple means.

Methods of solving problems of continuum mechanics that, instead of integrating the differential equation under specified boundary conditions, involve minimization of the functional of Eq. 10.2.2 are referred to as energy methods, or variational methods.
Example 10.2.1

Consider the equation \( Au = -\frac{d^2 u}{dx^2} = 2 \), subject to boundary conditions \( u'(0) = 0 \) and \( u'(1) + u(1) = 0 \), which has the solution \( u = 3 - x^2 \). The operator \( A \) is positive definite on

\[
D_A = \{ u \in C^2(0, 1): u'(0) = u'(1) + u(1) = 0 \}
\]

To see this, form

\[
(Au, v) = \int_0^1 -u''v \, dx = \int_0^1 u'v' \, dx - u'v\big|_0^1
\]

\[
= \int_0^1 u'v' \, dx + u(1)v(1) = (u, Av)
\]

Thus \( A \) is symmetric. Further, with \( u = v \),

\[
(Au, u) = \int_0^1 (u')^2 \, dx + u^2(1) \geq 0
\]

and \( (Au, u) = 0 \) implies \( u' = 0 \) for \( 0 \leq x \leq 1 \) and \( u(1) = 0 \). Thus, \( u(x) = 0, 0 \leq x \leq 1 \), and \( A \) is positive definite.

By the minimum functional theorem, the function \( 3 - x^2 \) gives a minimum value to the functional

\[
F(u) = \int_0^1 \left[ (u')^2(x) - 4u \right] dx + u^2(1) \quad (10.2.3)
\]

in the class of functions that satisfy the given boundary conditions.

The Space of Functions with Finite Energy

In Section 10.1, it was observed that if \( A \) is a positive bounded below operator, defined on a dense subspace \( D_A \) of a Hilbert space \( H \), then a new energy scalar product \([u, v]_A = (Au, v)\) can be defined on \( D_A \), with the property that convergence in the energy norm implies convergence in the original norm of \( H \); i.e., for a sequence \( \{u_n\} \) in \( D_A \),

\[
\lim_{n,m \to \infty} \|u_n - u_m\|_A = 0 \implies \lim_{n,m \to \infty} \|u_n - u_m\| = 0 \quad (10.2.4)
\]

Hence, a Cauchy sequence \( \{u_n\} \) in \( \|\cdot\|_A \) is also a Cauchy sequence in the \( H \)-norm. Since \( H \) is complete, there is a \( u_0 \) in \( H \) such that \( \lim_{n \to \infty} u_n = u_0 \). While \( u_0 \) may not be in \( D_A \), all such limit functions may be added to \( D_A \) to obtain a larger space of functions.
Definition 10.2.1. The space $H_A$ of generalized functions associated with a positive bounded below operator $A$ consists of those elements $u$ in the Hilbert space $H$ for which there exists a sequence of elements $\{ u_n \}$ in $D_A$ such that

$$\lim_{n,m \to \infty} \| u_n - u_m \|_A = 0 \quad \text{and} \quad \lim_{n \to \infty} \| u_n - u \| = 0$$

(10.2.5)

i.e.,

$$H_A = \{ u \in H : \text{there is a sequence } \{ u_n \} \text{ in } D_A \text{ that is Cauchy in } \| \cdot \|_A \text{ and converges in the norm of } H \text{ to } u \}$$

(10.2.6)

The elements of $H_A$ are said to be functions of finite energy.

Example 10.2.2

To understand how the space $H_A$ of generalized functions is constructed, consider the string problem with static load,

$$Au = -u_{xx} = f_n(x)$$

(10.2.7)

for $0 \leq x \leq 1$, with boundary conditions

$$u(0) = u(1) = 0$$

as shown in Fig. 10.2.1, where $n = 1, 2, \ldots$. Consider a point force at $x = \xi$, $0 < \xi < 1$, as the limiting case of continuous forces $f_n(x)$ that vanish in $(0, 1)$ if $|x - \xi| > 1/2n$, but for which the total intensity is

$$\int_0^1 f_n(x) \, dx = 1$$

Figure 10.2.1 String with Distributed Load $f_n(x)$

For each $f_n(x)$, let the solution of Eq. 10.2.7 be denoted as $u_n(x)$. Then $u_n(x)$ is in $C^2(0, 1)$ and satisfies the boundary conditions, so $u_n(x)$ is in $D_A$ for each $n = 1, 2, \ldots$. However,
\[
\lim_{n \to \infty} u_n(x) = u_0(x) = \begin{cases} 
(\xi - 1)x, & 0 \leq x \leq \xi \\
\xi x - \xi, & \xi \leq x \leq 1 
\end{cases}
\] (10.2.8)

and \(u_0(x)\) does not belong to \(D_A\). On the other hand, \(u_0(x)\) belongs to \(H_A\). The energy norm of \(u_0(x)\) is

\[
\|u_0\|_A^2 = \int_0^1 (u')^2 \, dx = \int_0^\xi (\xi - x)^2 \, dx + \int_\xi^1 \xi^2 \, dx = \xi - \xi^2
\]

Thus, \(\|u_0\|_A = (\xi - \xi^2)^{1/2}\), which is finite.

**Theorem 10.2.2.** Let \(u\) and \(v\) be in \(H_A\) for a positive bounded below operator \(A\); i.e., there are Cauchy sequences \(\{u_n\}\) and \(\{v_n\}\) in \(H_A\), \(u_n\) and \(v_n\) are in \(D_A\), and

\[
\lim_{n \to \infty} \|u - u_n\| = \lim_{n \to \infty} \|v - v_n\| = 0.
\]

Then,

\[
\|u\|_A = \lim_{n \to \infty} \|u_n\|_A
\]

\[
[u, v]_A = \lim_{n \to \infty} [u_n, v_n]_A
\]

are finite and define a norm and the associated scalar product on \(H_A\).

To prove this result, note first that for any \(x\), \(y\), and \(z\) in \(D_A\),

\[
\|x + y\|_A \leq \|x\|_A + \|y\|_A
\]

since \(\|\cdot\|_A\) is a norm on \(D_A\). With \(z = x + y\), this is

\[
\|z\|_A - \|y\|_A \leq \|z - y\|_A
\]

Interchanging \(z\) and \(y\), this shows that

\[
\|z\|_A - \|y\|_A \leq \|z - y\|_A
\]

(10.2.11)

Thus,

\[
\|u_n\|_A - \|u_m\|_A \leq \|u_n - u_m\|_A
\]

Hence, \(\{\|u_n\|_A\}\) is a Cauchy sequence of real numbers, so it converges to a finite real number; i.e.,

\[
\lim_{n \to \infty} \|u_n\|_A = \|u\|_A < \infty
\]
Next, note that

\[
\left| \left[ u_n, v_n \right]_A - \left[ u_m, v_m \right]_A \right| = \left| \left[ u_n, v_n - v_m \right]_A + \left[ u_n - u_m, v_m \right]_A \right| \\
\leq \left| \left[ u_n, v_n - v_m \right]_A \right| + \left| \left[ u_n - u_m, v_m \right]_A \right| \\
\leq \| u_n \|_A \| v_n - v_m \|_A + \| v_m \|_A \| u_n - u_m \|_A
\]

Since \{ \| u_n \|_A \} and \{ \| v_n \|_A \} converge to finite values, this shows that \{ \left[ u_n, v_n \right]_A \} is a Cauchy sequence of real numbers, so it is convergent to a finite real number; i.e.,

\[
\lim_{n \to \infty} \left[ u_n, v_n \right]_A = \left[ u, v \right]_A < \infty
\]

Since \[ \bullet, \bullet \]_A is a scalar product on \( D_A \) and the limit in Eq. 10.2.10 is finite,

\[
\left[ u, v \right]_A = \lim_{n \to \infty} \left[ u_n, v_n \right]_A = \lim_{n \to \infty} \left[ v_n, u_n \right]_A = \left[ v, u \right]_A
\]

and

\[
\left[ \alpha u, v \right]_A = \lim_{n \to \infty} \left[ \alpha u_n, v_n \right]_A = \lim_{n \to \infty} \alpha \left[ u_n, v_n \right]_A \\
= \alpha \lim_{n \to \infty} \left[ u_n, v_n \right]_A = \alpha \left[ u, v \right]_A
\]

Since \[ u_n, u_n \]_A \geq 0,

\[
\left[ u, u \right]_A = \lim_{n \to \infty} \left[ u_n, u_n \right]_A \geq 0
\]

Finally, if

\[
\left[ u, u \right]_A = \lim_{n \to \infty} \left[ u_n, u_n \right]_A = 0
\]

then, since \[ u_n, u_n \]_A \geq \gamma \| u_n \|^2,

\[
\lim_{n \to \infty} \| u_n \|^2 = 0
\]

and \( u = 0 \).

To see that the norm \( \| \bullet \|_A \) is generated by the scalar product \[ \bullet, \bullet \]_A, note that

\[
\| u \|^2_A = \lim_{n \to \infty} \| u_n \|^2_A \\
= \lim_{n \to \infty} \left[ u_n, u_n \right]_A \\
= \left[ u, u \right]_A
\]

\[ \blacksquare \]
These results permit the proof of an important theorem of modern mechanics and applied mathematics.

**Theorem 10.2.3.** For a positive bounded below operator $A$, $H_A$ with the norm and scalar product of Eqs. 10.2.9 and 10.2.10 is a Hilbert space. That is, $H_A$ is complete.

To prove this result, first let $\{u_n\}$ in $D_A$ be a Cauchy sequence in the energy norm. Then, $w_k = u - u_k$, where $\lim_{n \to \infty} \|u - u_n\| = 0$, is in $H_A$. For fixed $k$, the sequence $\{w_{kn}\} = \{u_n - u_k\}$ is in $D_A$ and is Cauchy, since

$$\|w_{kn} - w_{km}\|_A = \|u_n - u_m\|_A$$

Thus, by Eq. 10.2.9,

$$\|u - u_k\|_A = \|w_k\|_A = \lim_{n \to \infty} \|w_{kn}\|_A = \lim_{n \to \infty} \|u_n - u_k\|_A$$

Since $\{u_n\}$ is Cauchy in the norm $\|\cdot\|_A$,

$$\lim_{k \to \infty} \|u - u_k\|_A = \lim_{k \to \infty} \lim_{n \to \infty} \|u_n - u_k\|_A = 0 \quad (10.2.12)$$

This shows that for any $u$ in $H_A$, there is a $u_k$ in $D_A$ that is arbitrarily close to $u$ in the energy norm. This result is of significant value in its own right.

To complete the proof of Theorem 10.2.3, let $\{u_n\}$ be a Cauchy sequence in $H_A$. Thus, for any $\varepsilon > 0$, there is an integer $N_1$ such that

$$\|u_n - u_m\|_A \leq \frac{\varepsilon}{2} \quad (10.2.13)$$

for $n, m > N_1$. As a result of Eq. 10.2.12, for each $u_n$ in $H_A$, there is a $\phi_n$ in $D_A$ such that

$$\|u_n - \phi_n\|_A \leq \frac{\varepsilon}{4}$$

Thus,

$$\|\phi_n - \phi_m - (u_n - u_m)\|_A \leq \|\phi_n - u_n\|_A + \|\phi_m - u_m\|_A \leq \frac{\varepsilon}{2}$$

and by Eq. 10.2.11,

$$\|\phi_n - \phi_m\|_A + \|u_n - u_m\|_A \leq \|\phi_n - \phi_m - (u_n - u_m)\|_A \leq \frac{\varepsilon}{2}$$
For \( n, m > N_1 \), this result and Eq. 10.2.13 yield
\[
\| \phi_n - \phi_m \|_A \leq \epsilon
\]
so \( \{ \phi_n \} \) in \( D_A \) is a Cauchy sequence in the energy norm. There is thus a \( u \) in \( H_A \) such that, by Eq. 10.2.12,
\[
\lim_{n \to \infty} \| u - \phi_n \|_A = 0
\]
Thus, there is an integer \( N_2 \) such that
\[
\| u - \phi_n \|_A < \frac{\epsilon}{2}
\]
if \( n > N_2 \). Finally,
\[
\| u_n - u \|_A = \| u_n - \phi_n + (\phi_n - u) \|_A \\
\leq \| u_n - \phi_n \|_A + \| \phi_n - u \|_A \\
\leq \frac{\epsilon}{4} + \frac{\epsilon}{2} < \epsilon
\]
for \( n > \max(N_1, N_2) \). Therefore,
\[
\lim_{n \to \infty} \| u_n - u \|_A = 0
\]
and \( H_A \) is complete in the energy norm. Taken with Theorem 10.2.2, this shows that \( H_A \) is a Hilbert space.

**Example 10.2.3**

Consider again the operator \( A = -\frac{d^2}{dx^2} \) of Example 10.1.1. The energy space \( H_A \) of this operator is the space of functions on \( (0, 1) \) that are equal to zero at \( x = 0 \) and \( x = 1 \) and whose first derivatives are in \( L^2(0, 1) \). Note that the first derivative of the solution \( u_0(x) \) in Eq. 10.2.8 indeed belongs to \( L^2(0, 1) \). This result is a special case of an embedding theorem [18, 22] that guarantees roughly that if \( A \) is a differential operator of order \( 2m \), then \( H_A \) consists of functions that have continuous derivatives of order \( m - 1 \), derivatives of order \( m \) in \( H \), and satisfy the boundary conditions.

To prove this, first let \( u \) be an element of \( H_A \). Then there exists a sequence \( \{ u_n \} \) in \( D_A \) such that \( \lim_{n,m \to \infty} \| u_n - u_m \|_A = 0 \) and \( \lim_{n \to \infty} \| u_n - u \|_A = 0 \); i.e.,
\[
\lim_{n,m \to \infty} \int_0^1 [u_n'(x) - u_m'(x)]^2 \, dx = 0
\]
Since $L_2(0, 1)$ is complete, there exists a $w$ in $L_2(0, 1)$ such that \( \lim_{n \to \infty} \| u_n' - w \| = 0 \).

Define $u(x) = \int_0^x w(t) \, dt$. Then, $u' = w$ is in $L_2(0, 1)$ and it remains to show that $u(0) = u(1) = 0$. Since $u_n$ is in $D_A$, it follows that $u_n(0) = u_n(1) = 0$ and

$$u_n(x) = u_n(0) + \int_0^x u_n'(t) \, dt = \int_0^x u_n(t) \, dt$$

To see that $u_n(x) = \int_0^x u_n'(t) \, dt$ converges uniformly to $\int_0^x w(t) \, dt$ on the interval $[0, 1]$, note that

$$\left[ \int_0^x u_n'(t) \, dt - \int_0^x w(t) \, dt \right]^2 = \left[ \int_0^x [ u_n'(t) - w(t) ] \, dt \right]^2 \leq x \int_0^x [ u_n'(t) - w(t) ]^2 \, dt \leq \int_0^1 [ u_n'(t) - w(t) ]^2 \, dt = \| u_n' - w \|^2$$

Thus, $u(x) = \int_0^x w(t) \, dt$, for all $x$ in $[0, 1]$, giving $u(0) = 0$. Similarly, using the formula

$$u_n(x) = u_n(1) - \int_x^1 u_n'(t) \, dt = - \int_x^1 u_n'(t) \, dt$$

$$u(x) = - \int_0^1 w(t) \, dt \text{ and } u(1) = 0.$$  

Conversely, let $u$ be a function with $u'$ in $L_2(0, 1)$ and $u(0) = u(1) = 0$. To show that $u$ is in the space $H_A$, a sequence \( \{ u_n \} \) in $D_A$ must be found such that $\lim_{n,m \to \infty} \| u_n - u_m \| = 0$ and $\lim_{n \to \infty} \| u_n - u \| = 0$. Since $u'$ is in $L_2(0, 1)$, it can be expanded in a cosine series,

$$u'(x) = \sum_{k=0}^\infty a_k \cos k\pi x = \sum_{k=1}^\infty a_k \cos k\pi x$$

where
Sec. 10.2 The Minimum Functional Theorem

\[ a_0 = \int_0^1 u'(x) \, dx = u(1) - u(0) = 0 \]

Integrating this series term by term and taking into account \( u(0) = 0 \),

\[ u(x) = \sum_{k=1}^{\infty} b_k \sin k\pi x \]

where

\[ b_k = \frac{a_k}{k\pi} \]

Consider the sequence \( \{ u_n \} \), where \( u_n(x) = \sum_{k=1}^{n} b_k \sin k\pi x \). Clearly \( \{ u_n \} \) is in \( D_A \) and \( \lim_{n \to \infty} \| u_n - u \| = 0 \). To show that \( \lim_{n,m \to \infty} \| u_n - u_m \|_A = 0 \), assume without loss of generality that \( n > m \). Then, \( u_n(x) - u_m(x) = \sum_{k=m+1}^{n} b_k \sin k\pi x \) and

\[ \| u_n - u_m \|_A^2 = \int_0^1 \left[ \sum_{k=m+1}^{n} a_k \cos k\pi x \right]^2 \, dx = \frac{1}{2} \sum_{k=m+1}^{n} a_k^2 \]

which approaches zero as \( m, n \to \infty \). Thus, \( u \) is in \( H_A \).

**Generalized Solutions of Operator Equations**

For a positive bounded below operator \( A \) on a domain \( D_A \), the energy functional \( F(u) = (Au, u) - 2(u, f) \) can be extended to all of the energy space \( H_A \), as \( F(u) = \| u \|_A^2 - 2(u, f) \). The following important existence theorem will now be proved.

**Theorem 10.2.4.** In the energy space \( H_A \) of a positive bounded below operator \( A \), for each \( f \) in \( H \), there exists one and only one element for which the energy functional attains a minimum.

To prove this theorem, consider the linear functional \( \mathcal{L}(u) = (u, f) \), defined on \( H_A \). Then, \( |\mathcal{L}(u)| = |(u, f)| \leq \| u \| \| f \| \leq (\| f \|/\gamma) \| u \|_A \), which implies that \( \mathcal{L} \) is a continuous linear functional on \( H_A \). By the Riesz Representation Theorem, Theorem 9.1.3, there exists a unique element \( u_0 \) in \( H_A \) such that \( (u, f) = [u, u_0]_A \), for all \( u \) in \( H_A \). The energy functional \( F \) can now be written as

\[ F(u) = \| u \|_A^2 - 2[u, u_0]_A = \| u - u_0 \|_A^2 - \| u_0 \|_A^2 \]

from which it follows that the minimum of \( F \) is attained at \( u_0 \), proving the theorem.
Definition 10.2.2. For a positive bounded below operator $A$ and a function $f$ in $H$, the element $u_0$ in $H_A$ that yields a minimum for the functional $F$ is called the \textit{generalized solution} of the equation $Au = f$. If $u_0$ is in $D_A$, it is called an \textit{ordinary solution}.

EXERCISES 10.2

1. Show that the operator

$$Au = -\frac{d^2u}{dx^2}$$

with domain

$$D_A = \{ u \in C^2(0, 1): u'(0) = u'(1) + u(1) = 0 \}$$

is positive definite.

2. Consider the problem of integrating the equation

$$-\frac{du^2}{dx^2} = 2$$

subject to boundary conditions $u(0) = u(1) = 0$. Determine the functional whose minimum provides the solution to the above equation and determine the domain of this functional; i.e., the space of functions for which it is defined. The solution of the differential equation is $u_0 = x(1 - x)$. Compute the value of the functional for $u_0 = x(1 - x)$ and $u_1 = x(2 - x)$. Is the result a contradiction of the minimum functional theorem? Why?

10.3 \textbf{CALCULUS OF VARIATIONS}

The Minimum Functional Theorem of Section 10.2 motivates a brief introduction to the Calculus of Variations. Only necessary conditions for minimization of an integral functional are treated here, which is adequate for the purposes of this text.

Variations of Functionals

\textbf{Definition 10.3.1.} Let $H$ be a Hilbert space (normally $L_2$) and $T$ a functional (possibly nonlinear) that is defined on a subset $D_T$ (not necessarily a linear subspace) of $H$. Let $u$ be in $D_T$ and $\eta$ be in a linear subspace $D_T$ of $H$, such that $u + \alpha \eta$ is in $D_T$ for $\alpha$ sufficiently small. If the limit

$$\delta T(u; \eta) = \lim_{\alpha \to 0} \frac{1}{\alpha} \left[ T(u + \alpha \eta) - T(u) \right]$$

$$= \frac{d}{d\alpha} T(u + \alpha \eta) \bigg|_{\alpha=0}$$

(10.3.1)
exists, it is called the **first variation** of $T$ at $u$ in the direction $\eta$. If this limit exists for every $\eta$ in $\tilde{D}_T$, $T$ is said to be a **differentiable functional** at $u$.

**Example 10.3.1**

Let $f(x, u, u')$ be a continuous, real valued function for $x$ in $[a, b]$, $-\infty < u < +\infty$ and $-\infty < u' < +\infty$, and let the partial derivatives $f_u$ and $f_{u'}$ be continuous. Consider the integral functional

$$ T(u) = \int_a^b f(x, u(x), u'(x)) \, dx \quad (10.3.2) $$

whose domain is $D_T = \{ u \in C^1(a, b); u(a) = d, u(b) = e \}$ and let $\tilde{D}_T = \{ \eta \in C^1(a, b); \eta(a) = \eta(b) = 0 \}$. The first variation of $T$ takes the form

$$ \delta T(u; \eta) = \frac{d}{dx} T(u + \alpha \eta) \bigg|_{\alpha=0} = \frac{d}{d\alpha} \int_a^b f(x, u + \alpha \eta, u' + \alpha \eta') \, dx \bigg|_{\alpha=0} = \int_a^b \left[ f_u(x, u, u') \eta + f_{u'}(x, u, u') \eta' \right] \, dx \quad (10.3.3) $$

If $u(x)$ is such that $f_{u'}(x, u, u')$ is continuously differentiable, Eq. 10.3.3 can be integrated by parts to obtain

$$ \delta T(u; \eta) = \int_a^b \left[ f_u - \frac{d}{dx} f_{u'} \right] \eta \, dx \quad (10.3.4) $$

**Example 10.3.2**

Consider a region $\Omega$ in $\mathbb{R}^m$, with a boundary $\Gamma$ that consists of a finite number of piecewise smooth $(m-1)$-dimensional surfaces. Let the real valued function $f(x, u, z_1, z_2, \ldots, z_m)$ be twice continuously differentiable in all its arguments; i.e., $x, u,$ and $z_i$. Consider the functional

$$ T(u) = \int_{\Omega} f \left( x, u, \frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, \ldots, \frac{\partial u}{\partial x_m} \right) \, dx \quad (10.3.5) $$

defined on $D_T = \{ u \in C^1(\Omega); u(x) = g(x) \text{ on } \Gamma \}$ and let $\tilde{D}_T = \{ \eta \in C^1(\Omega); \eta(x) = 0 \text{ on } \Gamma \}$. The first variation $\delta T(u; \eta)$ is then
where \( u_k = \frac{\partial u}{\partial x_k} \) and \( \eta_k = \frac{\partial \eta}{\partial x_k} \). Assume now that \( \frac{\partial}{\partial x_k} \frac{\partial f}{\partial u} \) exists and is in \( L_2(a, b) \).

Integration by parts yields

\[
\delta T(u; \eta) = \int_\Omega \left[ \frac{\partial f}{\partial u} \eta + \sum_{k=1}^{m} \frac{\partial f}{\partial u_k} \eta_k \right] dx
\]

(10.3.7)

**Minimization of Functionals**

Given that a functional has a first variation, quantitative criteria can be defined for its minimization. The focus here is on necessary conditions for extrema. Presume \( \hat{u} \) in \( D_T \) is such that

\[
T(\hat{u}) \leq T(u)
\]

(10.3.8)

for all \( u \) in \( D_T \). Then, \( \hat{u} \) is said to minimize \( T \) over \( D_T \). If Eq. 10.3.8 holds for all \( u \) in \( D_T \) that satisfy \( \| u - \hat{u} \| \leq \delta \), for some \( \delta > 0 \), \( T \) is said to have a relative minimum at \( \hat{u} \). A change in the sense of the inequality of Eq. 10.3.8 leads to maximization. Results will be stated here only for minimization.

Employing the notation and conventions of Definition 10.3.1, observe that for any fixed \( \eta \) in \( D_T \) and for any \( \alpha \) sufficiently small, if \( T \) has a relative minimum at \( \hat{u} \), then

\[
T(\hat{u}) = \min_{\alpha} T(\hat{u} + \alpha \eta) = T(\hat{u} + \alpha \eta) \bigg|_{\alpha=0}
\]

That is, for \( \hat{u} \) and \( \eta \) fixed, the real valued function \( T(\hat{u} + \alpha \eta) \) of the real parameter \( \alpha \) is a minimum at \( \alpha = 0 \). If the functional has a first variation, then \( T(\hat{u} + \alpha \eta) \) is a differentiable function of \( \alpha \) and a necessary condition for a minimum of \( T \) at \( \hat{u} \) is

\[
\delta T(\hat{u}; \eta) = \frac{d}{d\alpha} T(\hat{u} + \alpha \eta) \bigg|_{\alpha=0} = 0
\]

(10.3.9)

for all \( \eta \) in \( D_T \). For cases such as Examples 10.3.1 and 10.3.2, Eqs. 10.3.4 and 10.3.7 are of the form,
\[ \delta T(\hat{u}; \eta) = \int_{\Omega} E(\hat{u}, x) \eta(x) \, d\Omega = 0 \]  

(10.3.10)

for all \( \eta \) in \( \tilde{D}_T \).

If the subspace \( \tilde{D}_T \) is dense in the underlying space \( H \), then Eq. 10.3.10 is equivalent to requiring that the function \( E(\hat{u}(x), x) \) be orthogonal to a complete set of functions in \( H \); i.e., \( D_T \) contains a basis for \( H \). Thus,

\[ E(\hat{u}, x) = 0 \]

(10.3.11)

is a necessary condition for minimization of the functional \( T \) at \( \hat{u} \). The function spaces \( \tilde{D}_T \) treated herein are large enough to contain the space \( C_0^\infty(a, b) \) of functions in \( C^\infty(a, b) \) that, with all their derivatives, vanish at \( a \) and \( b \), which is itself dense in \( L_2(a, b) \) [22], so they are dense in \( L_2(a, b) \) and the condition of Eq. 10.3.11 holds.

**Necessary Conditions for Minimization of Functionals**

**Theorem 10.3.1.** The results of Examples 10.3.1 and 10.3.2 yield the following **Euler - Lagrange equations** as necessary conditions for minimization of the associated functionals:

(a) From Eq. 10.3.2,

\[ \frac{\partial f}{\partial u} - \frac{d}{dx} \left( \frac{\partial f}{\partial u'} \right) = 0, \quad a \leq x \leq b \]

(10.3.12)

for minimization of the functional

\[ T(u) = \int_{a}^{b} f(x, u, u') \, dx \]

(10.3.13)

(b) From Eq. 10.3.5,

\[ \frac{\partial f}{\partial u} - \sum_{k=1}^{m} \frac{\partial}{\partial x} \left( \frac{\partial f}{\partial u_k} \right) = 0, \quad \text{in } \Omega \]

(10.3.14)

for minimization of the functional

\[ T(u) = \int_{\Omega} f(x, u, u_1, u_2, \ldots, u_m) \, d\Omega \]

(10.3.15)

The integration by parts in Eqs. 10.3.3 and 10.3.6 resulted in no terms from the boundary conditions, because of the nature of \( D_T \) and \( \tilde{D}_T \) in those problems. In many practical problems, however, boundary terms do arise.
Example 10.3.3

Consider minimization of the functional of Eq. 10.3.2,

\[ T(u) = \int_a^b f(x, u, u') \, dx \]

but with \( D_T = \tilde{D}_T = C^2(a, b) \). Then,

\[ \delta T(u; \eta) = \int_a^b \left( f_u \eta + f_u' \right) \, dx \]

Integration by parts yields

\[ \delta T(u; \eta) = \int_a^b \left[ f_u - \frac{d}{dx} f_u' \right] \eta \, dx + f_u(x, \hat{u}(x), \hat{u}'(x)) \eta(x) \bigg|_a^b = 0 \]

(10.3.16)

for all \( \eta \) in \( \tilde{D}_T \). The function \( \eta \) can be selected arbitrarily at the endpoints of the interval and in the interior of the interval, so both the integral and the boundary term must vanish. Thus, in addition to the necessary condition of Eq. 10.3.12, the transversality condition

\[ f_u(x, \hat{u}(x), \hat{u}'(x)) \eta(x) \bigg|_a^b = 0 \]

(10.3.17)

must hold for all \( \eta \) in \( \tilde{D}_T \). Since \( \tilde{D}_T = C^2(a, b) \), \( \eta(a) \) and \( \eta(b) \) are arbitrary. This means that the boundary conditions

\[ f_u(a, \hat{u}(a), \hat{u}'(a)) = 0 \]

(10.3.18)

\[ f_u(b, \hat{u}(b), \hat{u}'(b)) = 0 \]

are also necessary conditions for a minimizing function \( \hat{u} \).

To illustrate methods of the Calculus of Variations, it is instructive to briefly study special cases and applications of the fundamental problem of minimizing

\[ T(u) = \int_a^b f(x, u, u') \, dx \]

subject to a variety of boundary conditions.

Example 10.3.4

The shortest path between points \((a, u^0)\) and \((b, u^1)\) in the \(x-u\) plane is to be found. As shown in Fig. 10.3.1, the particular path chosen between the points has a length
associated with it. The problem is to choose the curve \( u(x), a \leq x \leq b \), that has the shortest length. For a smooth curve \( u(x) \), the length is given by the functional

\[
T(u) = \int_a^b \left[ 1 + \left( \frac{du}{dx} \right)^2 \right]^{1/2} \, dx \tag{10.3.19}
\]

Similarly, given points \((a, u^0)\) and \((b, u^1)\) in a vertical plane that do not lie on the same vertical line, a curve \( u = \bar{u}(x), a \leq x \leq b \), joining them is to be found so that a particle that moves without friction will start at rest from point \((a, u^0)\) and traverse the curve to \((b, u^1)\) in the shortest possible time. Candidate curves are shown in Fig. 10.3.2. This is called the **Brachistochrone Problem**.

Let \( m \) be the mass of the particle and \( g \) be the acceleration due to gravity. Since the particle starts at rest at \((a, u^0)\) and there is no friction, the kinetic energy at any time is equal to the loss of potential energy; i.e.,

\[
\frac{1}{2}mv^2 = mg (u - u^0) \tag{10.3.20}
\]

where \( v \) is velocity,

\[
v = \left[ \left( \frac{dx}{dt} \right)^2 + \left( \frac{du}{dt} \right)^2 \right]^{1/2} = \left[ 1 + \left( \frac{du}{dx} \right)^2 \right]^{1/2} \frac{dx}{dt} \tag{10.3.21}
\]

and \( t \) is time. Solving Eq. 10.3.20 for \( v \), substituting the result into Eq. 10.3.21, and solving for \( dt \) yields

\[
dt = \frac{\left[ 1 + \left( \frac{du}{dx} \right)^2 \right]^{1/2}}{[2g (u - u^0)]^{1/2}} \, dx
\]
The total time $T$ required for the particle to move from $(a, u_0)$ to $(b, u^1)$ is thus

$$T = T(u) = \int_a^b \left[ \frac{1}{2} \left( \frac{du}{dx} \right)^2 \right]^{1/2} \frac{1}{2g (u - u_0)} \, dx$$  \hspace{1cm} (10.3.22)

This notation makes it clear that $T$ depends on the curve traversed by the particle. The Brachistochrone Problem, therefore, is reduced to finding a curve $u = \hat{u}(x)$, $a \leq x \leq b$, that passes through the given end points and makes $T$ as small as possible.

In many problems, the form of the function $f(x, u, u')$ permits simplification of the Euler - Lagrange equations. In any case, Eq. 10.3.12 may be expanded, using the chain rule of differentiation and the notation

$$f_u = \frac{\partial f}{\partial u}, \quad f_u' = \frac{\partial f}{\partial u'}, \quad f_{ux} = \frac{\partial^2 f}{\partial x \partial u'}, \quad f_{uu} = \frac{\partial^2 f}{\partial u' \partial u}$$

to obtain

$$f_u - f_{ux} - u' f_u u' - u'' f_u u' = 0 \hspace{1cm} (10.3.23)$$

This is simply a second order differential equation for $u(x)$. Several special cases are now considered.

**Case 1.** $f$ does not depend on $u'$; i.e.,

$$f = f(x, u) \hspace{1cm} (10.3.24)$$

Equation 10.3.23, in this case, is

$$f_u(x, u) = 0 \hspace{1cm} (10.3.25)$$

This is simply an algebraic equation in $x$ and $u$. Since there are no constants of integration, it will not generally be possible to pass the resulting curve through particular end points. This means that a solution to such a problem generally will not exist.

**Example 10.3.5**

As a concrete example of Case 1, find $u$ in $C^2(0, 1)$ to minimize

$$\int_0^1 u^2 \, dx$$

with

$$u(0) = 0, \quad u(1) = 1$$

The condition of Eq. 10.3.25 is

$$2u = 0$$
It is impossible to satisfy the condition \( u(1) = 1 \), so the problem has no solution. To get an idea of what has gone wrong, note that since \( u^2(x) \geq 0 \) for each \( x \),

\[
\int_0^1 u^2(x) \, dx \geq 0
\]

for any function \( u(x) \) on \( 0 \leq x \leq 1 \). It is clear that if there were a function that minimized \( \int_0^1 u^2 \, dx \), then the minimum value of the integral would be nonnegative. Even though no minimum exists, consider the following family of functions in \( C^2(0, 1) \):

\[
u_n(x) = x^n
\]

These functions all satisfy the end conditions and

\[
T(u_n) = \int_0^1 x^{2n} \, dx = \frac{1}{2n + 1}
\]

It is possible to choose \( n \) large enough so that \( \int_0^1 u_n^2 \, dx \) is as close as desired to zero. However, the limit of \( u_n(x) \) as \( n \) approaches infinity is the function

\[
u_\infty(x) = \begin{cases} 
0, & x < 1 \\
1, & x = 1
\end{cases}
\]

which is not even continuous, much less in \( C^2(0, 1) \). Graphs of this sequence of functions are shown in Fig. 10.3.3.
In this illustration, a solution of the problem exists in the class of piecewise continuous functions, but not in the class of twice continuously differentiable functions. This problem, therefore, serves as a warning that not all innocent looking calculus of variations problems have solutions.

**Case 2.** \( f \) depends only on \( u' \); i.e.,

\[
f = f(u')
\]

Equation 10.3.23 is, in this case,

\[
f_{u'u'} u'' = 0
\]

**Example 10.3.6**

As a concrete example of Case 2, find the shortest curve in the \( x-u \) plane that passes through points \((0, 0)\) and \((1, 1)\). The function \( f \) from Eq. 10.3.19 is

\[
f = [1 + (u')^2]^{1/2}
\]

The form of the Euler-Lagrange equation in this case is, from Eq. 10.3.27,

\[-[1 + (u')^2]^{-3/2} u'' = 0\]

Since \((u')^2 \geq 0\), \([1 + (u')^2] \neq 0\) and \(u'(x)\) is required to be continuous, so \([1 + (u')^2] \neq \infty\). Therefore, it is required that

\[u''(x) = 0\]

or,

\[u(x) = ax + b\]

where \(a\) and \(b\) are constants. This implies that the shortest path between two points in a plane is a straight line, which is not surprising.

The end conditions yield

\[u(0) = b = 0\]

and

\[u(1) = a = 1\]

Therefore, the solution of the problem is

\[u(x) = x\]

**Case 3.** \( f \) depends only on \( x \) and \( u' \); i.e.,

\[
f = f(x, u')
\]
Equation 10.3.12 is, in this case,
\[ \frac{d}{dx} f_u(x, u') = 0 \]
or,
\[ f_u(x, u') = c \]  \hspace{1cm} (10.3.29)
where \( c \) is an arbitrary constant.

**Case 4.** \( f \) depends only on \( u \) and \( u' \); i.e.,
\[ f = f(u, u') \]  \hspace{1cm} (10.3.30)
Equation 10.3.23 is, in this case,
\[ f_u - f_{u'u} u' - f_{u''} u'' = 0 \]
Multiplying by \( u' \) yields
\[ u' f_u - (u')^2 f_{u'u} - u'u'' f_{u'u'} = 0 \]
This is just
\[ \frac{d}{dx} \left( f - u' f_u \right) = 0 \]
so,
\[ f - u' f_u = c \]  \hspace{1cm} (10.3.31)
where \( c \) is an arbitrary constant.

**Example 10.3.7**

As a concrete example of Case 4, consider the Brachistochrone Problem of Example 10.3.4. The function \( f \) of Eq. 10.3.22 is
\[ f = \left[ \frac{1 + (u')^2}{2gu} \right]^{1/2} \]
Equation 10.3.31 applies in this case and yields
\[ \left[ \frac{1 + (u')^2}{2gu} \right]^{1/2} - \frac{(u')^2}{(2gu)^{1/2} \left[ 1 + (u')^2 \right]^{1/2}} = c \]
This reduces to
\[ 1 = \{ 2gu \{ 1 + (u')^2 \} \}^{1/2} c \]
or,
\[ u \{ 1 + (u')^2 \} = c_1 \]
where \( c_1 \) is a new constant. The solution of this differential equation is a family of cycloids in parametric form; i.e.,
\[ x = c_2 + \frac{c_1}{2} (s - \sin s) \]
\[ u = \frac{c_1}{2} (1 - \cos s) \]
(10.3.32)

The constants \( c_1 \) and \( c_2 \) are to be determined so that the cycloid passes through the given points.

Example 10.3.8

As an example that has more engineering significance, it is instructive to view \( H_A \) as a space of candidate generalized solutions of a boundary-value problem and see what sort of generalized solution is actually generated by Theorem 10.2.4. The differential equation and boundary conditions for displacement of the static, laterally loaded string of Fig. 10.3.4 are

\[ -u'' = -f \]
\[ u(0) = u(\ell) = 0 \]
(10.3.33)

Based on analysis of the operator \( Au = -u'' \) with domain \( D_A = \{ u \in C^2(0, \ell): u(0) = u(\ell) = 0 \} \) in Example 10.2.2, it is known that functions in \( H_A \) are continuous, have square integrable derivatives, and satisfy the boundary conditions; i.e., \( H_A = \{ u \in C^0(0, \ell): u' \in L^2(0, \ell), u(0) = u(\ell) = 0 \} \). Consider first a string shown in Fig. 10.3.5, loaded with a piecewise continuous distributed load

\[ f(x) = \begin{cases} 0, & 0 \leq x \leq a \\ 1, & a \leq x \leq \ell \end{cases} \]
The energy functional $F(x) = \int_0^L (u'^2 + 2fu)\,dx$, in this problem, is

$$F(u) = \int_0^L \left[ (u')^2 + 2fu \right] dx$$

Ideas of the calculus of variations may now be applied, recognizing that $u'$ may be discontinuous at $x = a$, to obtain a function $u$ in $H_A$ that minimizes $F(u)$. Setting the first variation of $F(u)$ equal to zero,

$$\delta F(u; \eta) = \int_0^L \left[ 2u'\eta' + 2fu \right] dx = 0$$

for all $\eta$ in $H_A$. Integration by parts from 0 to $a$ and from $a$ to $L$ yields, noting that $u'$ may be discontinuous at $x = a$,

$$\int_0^L \left[ -2u'' + 2f \right] \eta\,dx + 2u'\eta \bigg|_{a-0}^{a+0} = 0$$

for arbitrary $\eta$ in $H_A$. Thus,

$$-u'' + f = 0$$

except at $x = a$. Since $\eta$ is continuous at $x = a$,

$$u'(a-0) = u'(a+0) = 0$$

so $u'(x)$ is continuous at $x = a$, even though $u''$ may not be defined there.

Solving the differential equation of Eq. 10.3.33,

$$u(x) = \begin{cases} 
\alpha_1 x, & 0 \leq x \leq a \\
\frac{1}{2} \left( x^2 - L^2 \right) + \alpha_2 \left( x - L \right), & a \leq x \leq L
\end{cases}$$

where $\alpha_1$ and $\alpha_2$ are constants of integration. Setting $u(a-0) = u(a+0)$ and $u'(a-0) = u'(a+0)$,
Thus, the generalized solution is in $D^2(0, \ell) \cap C^1(0, \ell) \subset H_a$ and is the physically meaningful solution of this problem.

As a second loading, consider the static string with a point load applied, as shown in Fig. 10.3.6. This application allows consideration of less regular candidate solutions.

**Figure 10.3.6** String with Point Load

In order to write $F(u)$, first form a distributed load $f(x)$ around $x = a$, with resultant $\hat{f}$. Then form the integral $\int_0^\ell f u \, dx$ and take the limit as the interval over which the load is distributed approaches zero, to obtain

$$\int_0^\ell f u \, dx = \hat{f} \, u(a)$$

Thus, the functional $F(u)$ is

$$F(u) = \int_0^\ell (u')^2 \, dx + 2\hat{f} \, u(a)$$

The first variation of $F(u)$ is

$$\delta F(u; \eta) = \int_0^\ell 2u'\eta' \, dx + 2\hat{f} \, \eta(a) = 0$$

for all $\eta$ in $H_a$. Integrating by parts, noting that $u'$ may be discontinuous at $x = a$,

$$-\int_0^\ell 2u''\eta \, dx + 2u'\eta \bigg|_{a-0}^{a+0} + 2\hat{f} \, \eta(a)$$

$$= \int_0^\ell 2u''\eta \, dx + 2 \left[ u'(a-0) - u'(a+0) + \hat{f} \right] \eta(a) = 0$$
since $\eta$ is continuous at $x = a$. The integrand and boundary terms must vanish independently, so

$$u'' = 0$$  \hspace{1cm} (10.3.34)

except at $x = a$, and

$$u'(a^-) - u'(a^+) + \hat{f} = 0$$  \hspace{1cm} (10.3.35)

The solution of the differential equation of Eq. 10.3.34 is

$$u(x) = \begin{cases} 
\alpha_1x, & 0 \leq x \leq a \\
\alpha_2(\ell - x), & a \leq x \leq \ell 
\end{cases}$$

Continuity of $u$ at $x = a$ implies

$$\alpha_1a = \alpha_2(\ell - a)$$

so,

$$u(x) = \begin{cases} 
\alpha_2 \frac{\ell - a}{a} x, & 0 \leq x \leq a \\
\alpha_2(\ell - x), & a \leq x \leq \ell 
\end{cases}$$

Substitution of $u'(a^-)$ and $u'(a^+)$ from this equation into Eq. 10.3.35,

$$\alpha_2 \left( \frac{\ell - a}{a} \right) + \alpha_2 + \hat{f} = 0$$

or,

$$\alpha_2 = -\frac{\hat{f}a}{\ell}$$

Thus, the generalized solution is

$$u(x) = -\frac{\hat{f}a}{\ell} \begin{cases} 
\frac{\ell - a}{a} x, & 0 \leq x \leq a \\
(\ell - x), & a \leq x \leq \ell 
\end{cases}$$

For this problem, the generalized solution is in $D^1(0, \ell) \cap C^0(0, \ell) \subset H_\Lambda$, which is physically meaningful for this problem. In fact, the condition of Eq. 10.3.35 for slope discontinuity at $x = a$, derived here from the minimum functional theorem, is identical to Eq. 6.3.13, which was derived on purely physical grounds. This result reinforces the applicability and usefulness of Theorem 10.2.4.
EXERCISES 10.3

1. Construct the first variation of the following functionals:

(a) \( T_1(u) = \int_0^1 \left[ 1 + (u')^2 \right]^{1/2} \, dx \)

\( D_{T_1} = \{ u \in C^2(0, 1): u(0) = a, \ u(1) = b \} \)

\( \tilde{D}_{T_1} = \{ \eta \in C^2(0, 1): \eta(0) = \eta(1) = 0 \} \)

(b) \( T_2(u) = \int \Omega \left[ \sum_{k=1}^m \left( \frac{\partial u}{\partial x_k} \right)^2 - 2fu \right] \, d\Omega \)

\( D_{T_2} = \tilde{D}_{T_2} = \{ u \in C^2(\Omega): u = 0 \text{ on } \Gamma \} \)

2. Extend the results given in Eq. 10.3.4 for a functional involving higher order derivatives; i.e., find the first variation of the following functionals:

(a) \( T_1(u) = \int_0^1 f(x, u, u', u'') \, dx \)

\( D_{T_1} = \tilde{D}_{T_1} = \{ u \in C^4(0, 1): u(0) = u'(0) = u(1) = u'(1) = 0 \} \)

(b) \( T_2(u) = \int_0^1 f(x, u, u', u'', \ldots, u^{(n)}) \, dx \)

\( D_{T_2} = \tilde{D}_{T_2} = \{ u \in C^{2n}(0, 1): u(0) = \ldots = u^{(n-1)}(0) \)

\( = u(1) = \ldots = u^{(n-1)}(1) = 0 \} \)

3. Extend the result of Eq. 10.3.7 to the functional

\( T(u) = \int_\Omega f(x, u, u_1, u_2, u_{11}, u_{12}, u_{22}) \, d\Omega \)

where \( \Omega \) is a domain in \( \mathbb{R}^2 \) and

\( D_T = \tilde{D}_T = \{ u \in C^4(\Omega): u(x) = \frac{\partial u}{\partial n}(x) = 0 \text{ in } \Gamma \} \)

4. Let \( u(x) = [ u_1(x), u_2(x), \ldots, u_k(x) ]^T \), where each \( u_i(x) \) is in \( D = \tilde{D} = \{ u_i \in C^2(0, 1): u_i(0) = u_i(1) = 0 \} \). Find the first variation of

\( T(u) = \int_0^1 f(x, u_1, \ldots, u_k, u_1', \ldots, u_k') \, dx \)
5. State necessary conditions for minima of the functionals $T_1$ and $T_2$ of Exercise 1, but with

$$
D_{T_1} = \{ u \in C^2(0, 1): u(0) = a \}
$$

$$
\bar{D}_{T_1} = \{ \eta \in C^2(0, 1): \eta(0) = 0 \}
$$

and

$$
D_{T_2} = \bar{D}_{T_2} = C^2(\Omega)
$$

6. Obtain necessary conditions for the functional of Exercise 2(a), but with

$$
D_{T_1} = \bar{D}_{T_1} = \{ u \in C^4(0, 1): u(0) = u(1) = 0 \}
$$

7. Verify that the cycloids of Eq. 10.3.32 satisfy the Euler-Lagrange equations for the Brachistochrone problem.

### 10.4 Natural Boundary Conditions

In using the variational formulation of boundary-value problems given by Theorems 10.2.1 and 10.2.4, certain conclusions concerning candidate minimizing functions may be drawn from the calculus of variations.

**Definition 10.4.1.** Boundary conditions that must be satisfied by the function that minimizes $F(u) = \| u \|_{H_A}^2 - 2 (u, f)$ are called natural boundary conditions and the remaining conditions specified in the definition of $D_A$ are called principal boundary conditions.

Definition 10.4.1 can also be stated in terms of the energy space $H_A$, as follows [21]: Boundary conditions that must be satisfied by functions in $D_A$, but not necessarily by functions in $H_A$, are known as natural boundary conditions for the given operator.

**Example 10.4.1**

Consider the operator equation of Example 10.2.1, $Au \equiv -d^2u/dx^2 = 2$, with domain $D_A = \{ u \in C^2(0, 1): u'(0) = u'(1) + u(1) = 0 \}$. The energy functional, in this case, is

$$
F(u) = (Au, u) - 2 (u, 2) = \int_0^1 ( (u')^2 - 4u ) \, dx + u^2(1)
$$

Theorem 10.2.1 implies that if $\hat{u}$ minimizes $F(u)$, over all functions in $D_A$, then it
will be the solution of $Au = 2$, $u$ in $D_A$. Consider minimizing $F(u)$, over all $u$ in $C^2(0, 1)$; i.e., disregarding boundary conditions. The first variation of $F$ is

$$
\delta F(u; \eta) = \frac{d}{d\alpha} F(u + \alpha \eta) \bigg|_{\alpha = 0}
$$

$$
= \frac{d}{d\alpha} \int_0^1 \left[ (u' + \alpha \eta')^2 - 4(u + \alpha \eta) \right] dx
+ (u(1) + \alpha \eta(1))^2 \bigg|_{\alpha = 0}
$$

$$
= \int_0^1 \left[ 2u' \eta' - 4 \eta \right] dx + 2u(1) \eta(1)
$$

$$
= 2 \int_0^1 \left[ -u'' - 2 \eta \right] dx + 2\left( u'(1) + u(1) \right) \eta(1)
- 2u'(0) \eta(0)

(10.4.1)
$$

In order that $\delta F(u; \eta) = 0$, for all $\eta$ in $C^2(0, 1)$, the coefficients of $\eta$ in the integrand, of $\eta(1)$, and of $\eta(0)$ must all be zero. This is,

$$
\begin{align*}
-u'' &= 2, \quad 0 < x < 1 \\
u'(0) &= 0, \quad u'(1) + u(1) = 0
\end{align*}

(10.4.2)
$$

By Definition 10.4.1, these are natural boundary conditions for the problem.

In order to determine whether a given boundary condition is natural, the following rule [21] may be applied to a positive bounded below operator of order $2m$: The natural boundary conditions for this class of operators contain derivatives of $u$ of order $m$ and higher. The principal boundary conditions contain derivatives of $u$ up to only order $m - 1$. Thus, for the Laplace operator ($m = 1$), the condition $u |_{\Gamma} = 0$ is principal and the condition $(\partial u / \partial n + \sigma u) |_{\Gamma} = 0$ is natural. For the biharmonic operator ($m = 2$),

$$
\nabla^4 u = \frac{\partial^4 u}{\partial x^4} + 2 \frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4}
$$

which occurs in the theories of elasticity and thin plates [23, 24], the conditions for rigid fixing of the edge, $u |_{\Gamma} = 0$ and $(\partial u / \partial n) |_{\Gamma} = 0$, are both principal. In the conditions for a simply supported edge, $u |_{\Gamma} = 0$ and $\frac{\partial^2 u}{\partial n^2} + \gamma \left( \frac{\partial^2 u}{\partial \tau^2} + \frac{1}{\rho} \frac{\partial u}{\partial n} \right) = 0$ ($\rho$ is the radius of curvature of the edge), the first is principal and the second is natural.

A procedure that can be used to rigorously verify whether given boundary conditions are natural is the following:
Sec. 10.4 Natural Boundary Conditions

(1) Transform the expression \( (Au, v) = [u, v]_A \), where \( u \) and \( v \) in \( D_A \); i.e., \( u \) and \( v \) satisfy the complete set of boundary conditions, to a symmetric form in \( u \) and \( v \), using integration by parts and the boundary conditions.

(2) Substitute the symmetric expression for \( [u, u]_A \) into \( F(u) \).

(3) Derive necessary conditions of the calculus of variations for minimization of \( F(u) \), without regard for boundary conditions.

(4) Those boundary conditions that are necessarily satisfied by the minimizing function are natural.

This method was used in Eq. 10.4.1 to show that the conditions \( u'(0) = 0 \) and \( u'(1) + u(1) = 0 \) are natural for the operator \( Au = -d^2u/dx^2, 0 < x < 1 \).

**Example 10.4.2**

As a further example, consider the equation

\[
Au \equiv -\nabla^2 u = f
\]  

subject to boundary conditions

\[
\frac{\partial u}{\partial n} + \sigma u = 0
\]

where \( \sigma > 0 \) on the boundary \( \Gamma \) of \( \Omega \). For \( u \) in \( D_A \),

\[
[u, v]_A = (-\nabla^2 u, v) = -\int_{\Omega} v \nabla^2 u \, d\Omega
\]

Integration by parts reduces this to

\[
[u, v]_A = \int_{\Omega} \nabla u \nabla v \, d\Omega - \int_{\Gamma} v \frac{\partial u}{\partial n} \, dS
\]

Using the boundary condition, this reduces to the symmetric form

\[
[u, v]_A = \int_{\Omega} \nabla u \nabla v \, d\Omega + \int_{\Gamma} \sigma uv \, dS
\]

Thus,

\[
[u, u]_A = \int_{\Omega} (\nabla u)^2 \, d\Omega + \int_{\Gamma} \sigma u^2 \, dS
\]

and

\[
F(u) = [u, u]_A - 2(f, u)
\]

\[
= \int_{\Omega} [(\nabla u)^2 - 2fu] \, d\Omega + \int_{\Gamma} \sigma u^2 \, dS
\]
Taking the first variation of $F$,
\[ \delta F(u; \eta) = 2 \int_\Omega \left[ \nabla u^T \nabla \eta - 2f\eta \right] \, d\Omega + \int_\Gamma 2\sigma u \eta \, dS \]
and integrating by parts yields
\[ \delta F(u; \eta) = -2 \int_\Omega \nabla^2 u \eta \, d\Omega + 2 \int_\Gamma \frac{\partial u}{\partial n} \eta \, dS \]
\[ - \int_\Omega 2f \eta \, d\Omega + \int_\Gamma 2\sigma u \eta \, dS \]
\[ = -2 \int_\Omega (\nabla^2 u - f) \eta \, d\Omega + 2 \int_\Gamma \left( \frac{\partial u}{\partial n} + \sigma u \right) \eta \, dS \]

A necessary condition for $u_0$ to minimize $F(u)$ is that $\delta F(u_0; \eta) = 0$, with no constraints imposed on the boundary values of $u$. Thus, it follows that
\[ \frac{\partial u_0}{\partial n} + \sigma u_0 \bigg|_\Gamma = 0 \]
from which it can be concluded that this is a natural boundary condition.

**EXERCISES 10.4**

1. Determine, by the procedure presented in this section, which of the following boundary conditions for the operator
\[ Au = \frac{d^2}{dx^2} \left[ a(x) \frac{d^2u}{dx^2} \right] \]
are natural:
\[ a(0) u''(0) - u'(0) = 0 \]
\[ u(0) = 0 \]
\[ u''(1) = u(1) = 0 \]
where $a(x) > 0$, $0 \leq x \leq 1$.

**10.5 NONHOMOGENEOUS BOUNDARY CONDITIONS**

Up to now, in discussing the equation
\[ Au = f \quad (10.5.1) \]
it has been assumed that the operator $A$ is defined for functions that satisfy homogeneous boundary conditions. In practice, non-homogeneous boundary conditions often occur, so it is of interest to explain how the variational method can be extended for the case of non-
homogeneous boundary conditions. Equation 10.5.1 is to be solved, with the boundary conditions

\[ G_1 u \big|_{\Gamma} = g_1, \ G_2 u \big|_{\Gamma} = g_2, \ldots, \ G_r u \big|_{\Gamma} = g_r \]  
\( (10.5.2) \)

where \( G_1, G_2, \ldots, G_r \) are linear operators and \( g_1, g_2, \ldots, g_r \) are given functions. The number of boundary conditions is determined by the order \( k \) of the differential operator of Eq. 10.5.1 and by whether \( u(x) \) is a scalar or a vector function.

This problem is solved, under the following hypothesis [21]: There exists a function \( \psi(x) \) that, together with its derivatives of order up to \( k - 1 \), is continuous in \( \Omega \) and that has derivatives of order \( k \) that are piecewise continuous in \( \Omega \) and satisfy the boundary conditions of the problem; i.e.,

\[ G_1 \psi \big|_{\Gamma} = g_1, \ G_2 \psi \big|_{\Gamma} = g_2, \ldots, \ G_r \psi \big|_{\Gamma} = g_r \]  
\( (10.5.3) \)

Defining \( u - \psi = v \) and substituting \( u = \psi + v \) into Eq. 10.5.1, the new unknown function \( v \) satisfies the equation

\[ A v = f_1(x) \]  
\( (10.5.4) \)

where \( f_1(x) = f(x) - A \psi \). Similarly, from Eqs. 10.5.2, \( v \) satisfies the homogeneous boundary conditions

\[ G_1 v \big|_{\Gamma} = 0, \ G_2 v \big|_{\Gamma} = 0, \ldots, \ G_r v \big|_{\Gamma} = 0 \]  
\( (10.5.5) \)

Let the operator \( A \) be positive definite for the set of functions that satisfy Eq. 10.5.5. According to the Minimum Functional Theorem, the solution of Eq. 10.5.4, subject to boundary conditions of Eq. 10.5.5, is equivalent to finding the function \( v \) that satisfies Eq. 10.5.5 and minimizes the functional

\[ F(v) = (A v, v) - 2 (v, f_1) \]  
\( (10.5.6) \)

Replacing \( v \) by \( u - \psi \) and \( f_1 \) by \( f - A \psi \), the functional to be minimized by \( v \), equivalently by \( u \), is

\[ F(v) = (A u - A \psi) - 2 (u - \psi, f - A \psi) \]

\[ = (A u, u) - 2 (u, f) + (u, A \psi) - (A u, \psi) + 2 (\psi, f) - (A \psi, \psi) \]  
\( (10.5.7) \)

Since \( \psi \) does not satisfy homogeneous boundary conditions, \( (u, A \psi) \neq (A u, \psi) \), but it is often possible to establish a formula for the operator \( A \) such that

\[ (u, A \psi) - (A u, \psi) = \int_{\Omega} (u A \psi - \psi A u) d\Omega = \int_{\Gamma} r(u, \psi) dS \]  
\( (10.5.8) \)
where the expression \( r(u, \psi) \) depends on the form of the operator \( A \). Using boundary conditions of Eqs. 10.5.2 and 10.5.3, it is often possible to represent \( r(u, \psi) \) in the form \( r(u, \psi) = N(u) + M \), where \( N(u) \) depends only on \( u \) and the functions \( g_1, g_2, \ldots, g_r \) and \( M \) does not depend on \( u \), but may depend on \( \psi \). In such a case, \( F(v) \) of Eq. 10.5.7 is reduced to the form

\[
F(v) = (Au, u) - 2(u, f) + \int_{\Gamma} N(u) \, dS
+ \left[ 2(\psi, f) - (A\psi, \psi) + \int_{\Gamma} M \, dS \right]
\]

Since the terms in square brackets do not depend on \( u \) or \( v \), minimization of the functional \( F(v) \) is thus equivalent to finding \( u \) to minimize

\[
\Phi(u) = (Au, u) - 2(u, f) + \int_{\Gamma} N(u) \, dS
\]  

over all functions that satisfy boundary conditions of Eq. 10.5.2. Note that, in Eq. 10.5.9, \((Au, u)\) is not in a symmetric form; i.e., an integration by parts has not been carried out.

The functional of Eq. 10.5.9 can be constructed without knowing the function \( \psi \). However, in order that the minimum of this functional has meaning, such a function must exist. Otherwise the variational problem may have no solution. In many classes of applications, it has been proved that a function \( \psi \) that satisfies Eqs. 10.5.2 exists if the surface \( \Gamma \) is sufficiently smooth and the functions \( g_1, g_2, \ldots, g_r \) that enter into Eqs. 10.5.2 are differentiable a sufficient number of times in any direction tangential to the surface \( \Gamma \) [22].

It may happen that some of Eqs. 10.5.5 will be natural for the functional \( F(v) \). The corresponding non-homogeneous equations in Eq. 10.5.2 will then be natural for the functional \( \Phi(u) \), the minimum of which can be sought over the class of functions that satisfy Eqs. 10.5.2, except perhaps the natural boundary conditions.

**Example 10.5.1**

Consider the problem of integrating the Laplace equation

\[
-\nabla^2 u = 0
\]  

subject to the boundary conditions

\[
u \mid_\Gamma = g(x)
\]

In this case, \( f = 0 \) and the second term in Eq. 10.5.9 vanishes. In order to find \( N(u) \), construct the expression

\[
\int_{\Omega} (uA\psi - \psiAu) \, d\Omega = \int_{\Omega} (\psi \nabla^2 u - u \nabla^2 \psi) \, d\Omega
\]
Integrating by parts,

\[ \int_{\Omega} (\psi \nabla^2 u - u \nabla^2 \psi) \, d\Omega = \int_{\Gamma} \left( \psi \frac{\partial u}{\partial n} - u \frac{\partial \psi}{\partial n} \right) \, dS \]

where \( n \) is the external normal to \( \Gamma \). Thus,

\[ r(u, \psi) = \psi \frac{\partial u}{\partial n} - u \frac{\partial \psi}{\partial n} \]

By virtue of Eq. 10.5.11, which must be satisfied by both \( u \) and \( \psi \),

\[ r(u, \psi) = g \left( \frac{\partial u}{\partial n} \right) - g \left( \frac{\partial \psi}{\partial n} \right) \]

Hence, in this case,

\[ N(u) = g \frac{\partial u}{\partial n} \]

\[ M = g \frac{\partial \psi}{\partial n} \]

This gives

\[ \Phi(u) = - \int_{\Omega} u \nabla^2 u \, d\Omega + \int_{\Gamma} g \frac{\partial u}{\partial n} \, dS \quad (10.5.12) \]

The first integral on the right can be simplified to

\[ - \int_{\Omega} u \nabla^2 u \, d\Omega = \int_{\Omega} (\nabla u)^2 \, d\Omega - \int_{\Gamma} u \frac{\partial u}{\partial n} \, dS \]

\[ = \int_{\Omega} (\nabla u)^2 \, d\Omega - \int_{\Gamma} g \frac{\partial u}{\partial n} \, dS \quad (10.5.13) \]

Substituting this into Eq. 10.5.12, the boundary integral vanishes. Solving Eqs. 10.5.10 and 10.5.11 thus reduces to finding the function \( u \) that minimizes

\[ \Phi(u) = \int_{\Omega} (\nabla u)^2 \, d\Omega \]

over functions in \( C^1(\Omega) \) that satisfy the principal boundary condition of Eq. 10.5.11.

**EXERCISES 10.5**

1. Convert the following boundary-value problems with nonhomogeneous boundary conditions to problems with homogeneous boundary conditions:
(a) \[-\frac{d}{dx}\left[(1 + x^2)\frac{du}{dx}\right] + u = x, \quad 0 < x < 1\]
\[u(0) = 2\]
\[\frac{du}{dx}(1) = 1\]

(b) \[\frac{d^2}{dx^2}\left[x^2 \frac{d^2u}{dx^2}\right] + (1 + \sin x)u = e^x, \quad 0 < x < 1\]
\[u(0) = u(1) = 1\]
\[\frac{du}{dx}(0) + \frac{d^2u}{dx^2}(0) = 2\]
\[\frac{du}{dx}(1) = 1\]

(c) \[\nabla^2( f(x) \nabla^2 u ) = 0\]
\[u = x_1 + x_2\]
\[\nabla^2 u = 0\] on \(\Gamma\)

2. Find functionals whose minimization is equivalent to solving the boundary-value problems of Exercise 1.

3. Derive a functional that is minimized by the solution of
\[- u'' + u = f(x), \quad 0 < x < 1\]
\[u(0) = 0\]
\[u(1) = 1\]

4. Derive a functional that is minimized by the solution of
\[u'''' + u = f(x), \quad 0 < x < 1\]
\[u(0) = u'(0) = 0\]
\[u(1) = 1\]
\[u''(1) = 0\]
10.6 EIGENVALUE PROBLEMS

Completeness of Eigenfunctions in Energy

Properties of eigenvalues of positive bounded below operators and properties of the corresponding eigenfunctions were discussed in Section 9.2. Many of these properties carry over to the space of functions that have finite energy for the operator.

**Theorem 10.6.1.** If the eigenfunctions \( \{ \phi_i \} \) of a positive bounded below operator \( A \) are complete in the \( L_2 \) sense, then they are also complete in energy.

If \( f \) has finite energy and if \( \langle f, \phi_i \rangle_A = 0, i = 1, 2, \ldots \), then it follows that \( \langle A\phi_i, f \rangle = \lambda_i \langle \phi_i, f \rangle = 0, i = 1, 2, \ldots \). Since \( \lambda_i > 0, \langle \phi_i, f \rangle = 0, i = 1, 2, \ldots \), and since the \( \phi_i \) are complete in the \( L_2 \) sense, \( f = 0 \).

Theorem 10.6.1 permits construction of the solution of the operator equation \( Au = f \), where \( A \) is positive bounded below and has \( L_2 \) normalized eigenfunctions \( \{ \phi_i \} \) that are complete in the \( L_2 \) sense. This follows by constructing an energy complete orthonormal sequence \( \{ \psi_n \} \), as follows. First note that

\[
\| \phi_n \|_A^2 = \langle A\phi_n, \phi_n \rangle = \lambda_n \langle \phi_n, \phi_n \rangle = \lambda_n
\]

Normalizing the sequence \( \{ \phi_i \} \) in the energy norm yields the energy complete orthonormal sequence

\[
\psi_n = \frac{\phi_n}{\sqrt{\lambda_n}}
\]

These observations lead to the following results.

**Theorem 10.6.2.** The solution \( u_0 \) of \( Au = f \) can be represented by its energy Fourier series

\[
u_0 = \sum_{n=1}^{\infty} \langle u_0, \psi_n \rangle \psi_n = \sum_{n=1}^{\infty} \langle f, \psi_n \rangle \psi_n = \sum_{n=1}^{\infty} \frac{\langle f, \phi_n \rangle}{\lambda_n} \phi_n
\]

This series converges in energy and in the \( L_2 \) sense. Furthermore, for

\[
u_n = \sum_{i=1}^{n} \frac{\langle f, \phi_i \rangle}{\lambda_i} \phi_i
\]

\[
Au_n = \sum_{i=1}^{n} \frac{1}{\lambda_i} \langle f, \phi_i \rangle A\phi_i = \sum_{i=1}^{n} \langle f, \phi_i \rangle \phi_i
\]

so \( \lim_{n \to \infty} \| Au_n - f \| = 0 \).
If $A$ is a symmetric **bounded below operator**; i.e., $(Au, u) \geq k (u, u)$, $k$ possibly negative or zero, the result of Theorem 10.6.2 can still be shown to be valid [18, 22]. In particular, this shows that if $\phi_0$ is an eigenfunction of $A$ corresponding to a zero eigenvalue $\lambda_0 = 0$, then $Au = f$ has a solution only if $(f, \phi_0) = 0$. Another way of saying the same thing, which is a special case of the **Theorem of the Alternative**, is that $f$ must be orthogonal to all solutions of $Au = 0$.

**Minimization Principles for Eigenvalue Problems**

If an operator $A$ is positive bounded below, then there exists a positive constant $\gamma$ such that $(Au, u) \geq \gamma \|u\|^2$, for all $u$ in $D_A$, from which it follows that

$$\frac{(Au, u)}{(u, u)} \geq \gamma \quad (10.6.1)$$

for all $u \neq 0$ in $D_A$. This observation leads to the following theorem.

**Theorem 10.6.3.** Let $A$ be a symmetric operator with domain $D_A$ that is a dense subspace of $L_2$, such that there exists a greatest lower bound $d > -\infty$ of the functional

$$R(u) \equiv \frac{(Au, u)}{(u, u)} \quad (10.6.2)$$

which is called the **Rayleigh Quotient**. If there exists a function $u_0 \neq 0$ from $D_A$ such that

$$R(u_0) = \frac{(Au_0, u_0)}{(u_0, u_0)} = d \quad (10.6.3)$$

then $d$ is the smallest eigenvalue of the operator $A$ and $u_0$ is the corresponding eigenfunction.

To prove this theorem, let $\eta$ be an arbitrary function from $D_A$ and let $t$ be an arbitrary real number. Since for any $t$, $u_0 + t\eta$ is in $D_A$ and the function

$$\Phi(t) = \frac{(A(u_0 + t\eta), u_0 + t\eta)}{(u_0 + t\eta, u_0 + t\eta)}$$

$$= \frac{t^2 (A\eta, \eta) + 2t (Au_0, \eta) + (Au_0, u_0)}{t^2 (\eta, \eta) + 2t (u_0, \eta) + (u_0, u_0)}$$

has a minimum at $t = 0$, $\Phi'(0) = 0$. Evaluating $\Phi'(0)$ yields

$$(u_0, u_0) (Au_0, \eta) - (Au_0, u_0) (u_0, \eta) = 0$$
or, using Eq. 10.6.3,

\[
(Au_0 - d u_0, \eta) = 0
\]

(10.6.4)

for all \( \eta \) in \( D_A \). Since \( D_A \) is dense in \( L_2 \), it follows that

\[
Au_0 - d u_0 = 0
\]

i.e., \( d \) and \( u_0 \) are an eigenvalue and an associated eigenfunction of the operator \( A \).

It follows that \( d \) is the smallest eigenvalue, since if \( \lambda_1 \) and \( u_1 \neq 0 \) are another eigenvalue and its corresponding eigenfunction of the operator \( A \), then

\[
\lambda_1 = \frac{(A u_1, u_1)}{(u_1, u_1)} \geq \min_{u \in D_A} \frac{(A u, u)}{(u, u)} = d
\]

Theorem 10.6.3 reduces the problem of finding the smallest eigenvalue of an operator \( A \) that is only a bounded below operator; i.e., \( (Au, u) \geq \beta (u, u) \) for some \( \beta > -\infty \), but possibly \( \beta < 0 \), to the variational problem of finding the function that minimizes the Rayleigh Quotient

\[
R(u) = \frac{(Au, u)}{(u, u)}
\]

(10.6.5)

Consider next another formulation for this minimization problem, which is frequently more convenient. Letting \( \psi = u/\|u\|, \|\psi\| = 1 \), and

\[
\frac{(Au, u)}{(u, u)} = (A\psi, \psi)
\]

(10.6.6)

the variational problem can be formulated as follows: Find \( u \) in \( D_A \) to minimize the functional

\[
(Au, u)
\]

subject to the condition

\[
(u, u) = 1
\]

(10.6.8)

A method may now be stated for determining subsequent eigenvalues.

**Theorem 10.6.4.** Let \( \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n \) be the first \( n \) eigenvalues of an operator \( A \) that is bounded below and let \( u_1, u_2, \ldots, u_n \) be corresponding orthonormal eigenfunctions. Let there exist a function \( u = u_{n+1} \neq 0 \) that minimizes the functional of Eq. 10.6.5, under the supplementary conditions
Then, \( u_{n+1} \) is an eigenfunction of the operator \( A \) that corresponds to the eigenvalue

\[
\lambda_{n+1} = \frac{(Au_{n+1}, u_{n+1})}{(u_{n+1}, u_{n+1})} \tag{10.6.10}
\]

This is the next larger eigenvalue than \( \lambda_n \).

To prove this theorem, let \( \zeta \) be an arbitrary function from \( D_A \) and define

\[
\eta = \zeta - \sum_{k=1}^{n} (\zeta, u_k) u_k
\]

Then, \( \eta \) satisfies Eqs. 10.6.9. In fact,

\[
(\eta, u_m) = (\zeta, u_m) - \sum_{k=1}^{n} (\zeta, u_k)(u_k, u_m), \quad m = 1, \ldots, n \tag{10.6.11}
\]

Since the \( u_k \) are orthonormal,

\[
(\eta, u_m) = (\zeta, u_m) - (\zeta, u_m) = 0, \quad m = 1, \ldots, n
\]

Just as \( \eta \), the product \( t\eta \) also satisfies Eqs. 10.6.9, where \( t \) is any real number. Thus, the sum \( u_{n+1} + t\eta \) satisfies Eqs. 10.6.9.

Since \( u_{n+1} \) minimizes the functional of Eq. 10.6.5, for all such \( \eta \),

\[
\frac{(A(u_{n+1} + t\eta), u_{n+1} + t\eta)}{(u_{n+1} + t\eta, u_{n+1} + t\eta)}
\]

has a minimum at \( t = 0 \). By repeating the same argument as in the proof of Theorem 10.6.3,

\[
(Au_{n+1} - \lambda_{n+1}u_{n+1}, \eta) = 0 \tag{10.6.12}
\]

Note at this point, however, that \( \eta \) is not arbitrary.

Consider the quantity

\[
(Au_{n+1} - \lambda_{n+1}u_{n+1}, \zeta)
\]

From the definition of \( \eta \), Eqs. 10.6.11 and 10.6.12 lead to
Furthermore,


\[
(Au_{n+1} - \lambda_{n+1}u_{n+1}, \zeta) = (Au_{n+1} - \lambda_{n+1}u_{n+1}, \eta)
\]

\[
+ \sum_{k=1}^{n} (u_k, \zeta) (Au_{n+1} - \lambda_{n+1}u_{n+1}, u_k)
\]

\[
= \sum_{k=1}^{n} (u_k, \zeta) (Au_{n+1} - \lambda_{n+1}u_{n+1}, u_k)
\]

The second term on the right of this equation vanishes, by virtue of Eqs. 10.6.9. Since \(A\) is symmetric, the first term is

\[
(Au_{n+1}, u_k) = (u_{n+1}, Au_k)
\]

Since \(u_k\) is an eigenfunction of \(A\), \(Au_k = \lambda_k u_k\), so \((Au_{n+1}, u_k) = \lambda_k (u_{n+1}, u_k) = 0\). It follows that \((Au_{n+1} - \lambda_{n+1}u_{n+1}, \zeta) = 0\) and, since \(\zeta\) is arbitrary,

\[
Au_{n+1} - \lambda_{n+1}u_{n+1} = 0
\]

Thus, \(\lambda_{n+1}\) is an eigenvalue of the operator and \(u_{n+1}\) is the corresponding eigenfunction.

It remains to prove that \(\lambda_{n+1}\) is the smallest eigenvalue that is greater than or equal to \(\lambda_n\). Let \(\bar{\lambda}\) be any eigenvalue of the operator \(A\) that is greater than \(\lambda_n\), with \(\bar{u}\) the corresponding eigenfunction. Since \(A\) is symmetric, \(\bar{u}\) satisfies Eqs. 10.6.9 and

\[
\bar{\lambda} = \frac{(Au, \bar{u})}{(\bar{u}, \bar{u})} \geq \lambda_{n+1}
\]

since \(\lambda_{n+1}\) is the minimum of the functional of Eq. 10.6.5, subject to Eq. 10.6.9.

As in the case of the smallest eigenvalue, the determination of \(\lambda_{n+1}\) can be reduced to the variational problem of minimizing the functional of Eq. 10.6.7, subject to the conditions of Eqs. 10.6.8 and 10.6.9.

**EXERCISES 10.6**

1. Verify that Theorems 10.6.1 and 10.6.2 remain valid if the operator is merely bounded below and does not have a zero eigenvalue; i.e., there is a number \(\beta\) (perhaps negative) such that \((Au, u) \geq \beta \|u\|^2\).

2. Note that the numerator of the Rayleigh Quotient is the square of the energy norm of \(u\), which contains only \(m\) derivatives of \(u\) when the order of \(A\) is \(2m\). Use the results
of Exercise 1 of Section 10.1.1 to construct Rayleigh Quotients for the following operators:

(a) \( A_1 u = -\frac{d}{dx} \left( a(x) \frac{du}{dx} \right) \)

\[ \mathcal{D}_{A_1} = \{ u \in C^2(0, \ell): u(0) = u'(0) = u(\ell) = 0 \} \]

(b) \( A_2 u = \frac{d^2}{dx^2} \left[ EI(x) \frac{d^2u}{dx^2} \right] \)

\[ \mathcal{D}_{A_2} = \{ u \in C^4(0, \ell): u(0) = u'(0) = u(\ell) = u'(\ell) = 0 \} \]

(c) \( A_3 u = \nabla^2( b(x) \nabla^2 u ) \)

\[ \mathcal{D}_{A_3} = \{ u \in C^4(\Omega): u = \frac{\partial u}{\partial n} = 0 \text{ on } \Gamma \} \]

3. Show that if \( \bar{a}(x) \geq a(x), 0 \leq x \leq \ell \); \( \bar{E}I(x) \geq EI(x), 0 \leq x \leq \ell \); and \( \bar{b}(x) \geq b(x), x \in \Omega \), then the eigenvalues \( \lambda_i \) of operators \( A_1, A_2, \) and \( A_3 \) are greater than the eigenvalues \( \lambda_i \) of operators \( A_1, A_2, \) and \( A_3 \) of Exercise 2.

4. Let the natural frequency \( \omega \) of a body be determined by the positive definite operator equation \( Au = \omega^2 u \), with \( u \) in \( \mathcal{D}_A \). Show that if the system is more severely constrained; i.e., has a smaller domain \( \mathcal{D}_A \subset \mathcal{D}_A \), then the natural frequencies will increase (or at least not decrease).

5. Let the symmetric operators \( A \) and \( B \) be defined on the same domain \( \mathcal{D}_{AB} \), which is a dense subspace of \( L^2 \), such that there exists a greatest lower bound \( d > -\infty \) of the functional (generalized Rayleigh Quotient)

\[ R(u) \equiv \frac{(Au, u)}{(Bu, u)} = \frac{\|u\|^2_A}{\|u\|^2_B} \]

Show that if there exists a function \( u_0 \neq 0 \) from \( \mathcal{D}_{AB} \) such that

\[ R(u_0) = \frac{(Au_0, u_0)}{(Bu_0, u_0)} = d \]

then \( d \) is the smallest eigenvalue of the eigenvalue problem \( Au = \lambda Bu \) and \( u_0 \) is the corresponding eigenfunction.
10.7 MINIMIZING SEQUENCES AND THE RITZ METHOD

Minimizing Sequences

**Definition 10.7.1.** Let $\Phi(u)$ be a functional that is bounded below on a domain $D_{\Phi}$, so there is a greatest lower bound $d$; i.e., $d = \inf_{u \in D_{\Phi}} \Phi(u)$. A sequence $\{u_n\}$ of functions in $D_{\Phi}$ is called a minimizing sequence if

$$\lim_{n \to \infty} \Phi(u_n) = d$$

Let $A$ be a positive definite operator with domain in $H_A$. If the equation

$$Au = f$$

has a solution $u_0$ in $D_A$, then, as in the proof of Theorem 10.2.1, the functional

$$F(u) = (Au, u) - 2(u, f)$$

reduces to the form

$$F(u) = \|u - u_0\|_A^2 - \|u_0\|_A^2$$

(10.7.2)

It is clear that $\inf_{u \in D_A} F(u) = \min_{u \in D_A} F(u) = -\|u_0\|_A^2$. A minimizing sequence $\{u_n\}$ for the functional $F(u)$ is characterized by the equation

$$\lim_{n \to \infty} F(u_n) = -\|u_0\|_A^2$$

(10.7.3)

The following theorem is of great importance for the variational method.

**Theorem 10.7.1.** If Eq. 10.7.1 has a solution in $D_A$, then any minimizing sequence for the functional of Eq. 10.7.2 converges in energy to this solution. If $A$ is positive bounded below, the sequence converges in the mean.

To prove this theorem, let $\{u_n\}$ be a minimizing sequence for the functional of Eq. 10.7.2. From Eqs. 10.7.2 and 10.7.3, it follows that

$$\lim_{n \to \infty} F(u_n) = \lim_{n \to \infty} \{\|u_n - u_0\|_A^2 - \|u_0\|_A^2\} = -\|u_0\|_A^2$$

Hence, $\lim_{n \to \infty} \|u_n - u_0\|_A^2 = 0$; i.e., $u_n$ converges to $u$ in energy. If $A$ is a positive bounded below operator, $u_n$ converges to $u$ in the $L_2$ norm.

Theorem 10.7.1 underlies many numerical methods. Thus, to solve Eq. 10.7.1, assuming a solution exists, it suffices to construct a minimizing sequence for the functional of Eq. 10.7.2.
A minimizing sequence can also be defined for $F(u) = \langle u, u \rangle_A - 2\langle u, f \rangle$, in the energy space $H_A$ associated with the symmetric and positive definite operator $A$. Let the sequence $\{ u_n \}$ in $H_A$ be such that $\lim_{n \to \infty} F(u_n) = -\| u_0 \|_A^2$, where $u_0$ is in $H_A$ and minimizes $F(u)$ on $H_A$; i.e., it is the generalized solution of $Au = f$. The sequence $\{ u_n \}$ is thus a minimizing sequence in $H_A$ and Theorem 10.7.1 may be generalized, as follows.

**Theorem 10.7.2.** If a symmetric operator $A$ is positive definite, then any minimizing sequence of $F(u)$ in $H_A$ converges in energy to the generalized solution $u_0$ of Eq. 10.7.1 in $H_A$, which is guaranteed to exist. If $A$ is positive bounded below, then convergence is also in $L_2$.

The proof requires no change from that of Theorem 10.7.1, except to name the limit $u_0$.

Similarly, but less complete, the following theorem provides a basis for numerical methods of solving the eigenvalue problem

$$Au = \lambda u$$

**Theorem 10.7.3.** Let $A$ be a positive definite operator and let $\{ \phi_n \}$, with $\| \phi_n \| = 1$, be a minimizing sequence in $H_A$ for the functional $\| u \|_A^2$. If a subsequence $\{ \phi_{n_k} \}$ converges in the mean, then $\lambda_1 = \lim_{k \to \infty} \| \phi_{n_k} \|_A^2$ is the smallest eigenvalue of Eq. 10.7.4 and $u = \lim_{k \to \infty} \phi_{n_k}$ is the corresponding eigenfunction.

For proof of this theorem, see Ref. 18.

This result is not as powerful as Theorem 10.7.2, which guarantees convergence of the minimizing sequence. Sharper results for the eigenvalue problem require a knowledge of the method that is employed to construct the minimizing sequence.

**The Ritz Method of Generating Minimizing Sequences**

One of the basic difficulties with applications of minimizing sequences, often called direct methods for solving boundary-value problems, is the construction of a complete sequence of functions, called coordinate functions, that are used to generate minimizing sequences. In view of the importance of this question, key theoretical results that are of practical value in applications are first summarized. The first such result is a practical test for completeness of a sequence of coordinate functions.

**Theorem 10.7.4.** Let $A$ be a positive bounded below operator and let $\{ \phi_n \}$ be a sequence of functions from its domain $D_A$. If the sequence of functions $\{ A\phi_n \}$ is complete in $L_2$, then $\{ \phi_n \}$ is complete in $D_A$, hence also in $H_A$, in the sense of energy convergence.
To prove this theorem, let \( v(x) \) be an arbitrary function from \( D_A \). Since the sequence \( \{ A\phi_n \} \) is complete in \( L_2 \), for any \( \varepsilon > 0 \) it is possible to find a positive integer \( n \) and constants \( \alpha_1, \alpha_2, \ldots, \alpha_n \) such that

\[
\left\| A v - \sum_{k=1}^{n} \alpha_k A \phi_k \right\| = \left\| A \left[ v - \sum_{k=1}^{n} \alpha_k \phi_k \right] \right\| < \frac{1}{2} \gamma \varepsilon \tag{10.7.5}
\]

where \( \gamma \) is the constant in \( (Au, u) \geq \gamma^2 \| u \|^2 \). To obtain an estimate for the quantity \( \| v - \sum_{k=1}^{n} \alpha_k \phi_k \|_A \), let \( v_n = \sum_{k=1}^{n} \alpha_k \phi_k \). Thus,

\[
\| v - v_n \|_A^2 = (A(v - v_n), v - v_n)
\]

Using the Schwartz inequality and Eq. 10.7.5,

\[
\| v - v_n \|_A^2 \leq \| A(v - v_n) \| \| v - v_n \| < \frac{1}{2} \varepsilon \gamma \| v - v_n \| \tag{10.7.6}
\]

Substituting \( \| v - v_n \| \leq (1/\gamma) \| v - v_n \|_A \) into Eq. 10.7.6 and dividing by \( \| v - v_n \|_A \),

\[
\| v - v_n \|_A < \frac{1}{2} \varepsilon
\tag{10.7.7}
\]

so \( \{ \phi_n \} \) is complete in \( D_A \).

To prove \( \{ \phi_n \} \) is complete in \( H_A \), let \( u(x) \) be an arbitrary function with finite energy; i.e., \( u \) is in \( H_A \). Since \( D_A \) is dense in \( H_A \), there exists a function \( \tilde{v} \) in \( D_A \) such that

\[
\| u - \tilde{v} \|_A \leq \varepsilon/2.
\]

Since \( \tilde{v} \) is in \( D_A \), there is a sequence \( \tilde{v}_m = \sum_{i=1}^{m} \tilde{\alpha}_i \phi_i \) such that for \( m \) large enough, Eq. 10.7.7 is satisfied; i.e., \( \| \tilde{v} - \tilde{v}_m \|_A \leq \varepsilon/2 \). By the triangle inequality,

\[
\left\| u - \sum_{k=1}^{m} \tilde{\alpha}_k \phi_k \right\|_A = \| u - \tilde{v}_m \|_A \leq \| u - \tilde{v} \|_A + \| \tilde{v} - \tilde{v}_m \|_A < \varepsilon
\]

which was to be proved.

\[\boxed{}\]

**Example 10.7.1**

Let \( \Omega \) be the rectangle \( 0 \leq x \leq a, 0 \leq y \leq b \) and let \( Au = -\nabla^2 u \), with \( u = 0 \) on the boundary \( \Gamma \). Choose as coordinate functions
\[ \phi_{m,n}(x, y) = \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b}, \quad m, n = 1, 2, \ldots \]  

Then,

\[ A\phi_{m,n} = -\nabla^2 \phi_{m,n} = \pi^2 \left( \frac{m^2}{a^2} + \frac{n^2}{b^2} \right) \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b} \]

From the theory of double Fourier series, it is known that the sequence of functions \{ \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b} \} is complete in the sense of convergence in the mean. Since multiplication of each function by a non-zero constant does not destroy completeness of the sequence, the sequence \{ A\phi_{m,n} \} is complete in \( L_2(\Omega) \). It follows from Theorem 10.7.4 that the functions in Eq. 10.7.8 are complete in \( H_A \).

This implies the following. Let \( u(x, y) \) be any function that is continuous and continuously differentiable in the rectangle \( 0 \leq x \leq a, 0 \leq y \leq b \) and vanishes on its boundary. For any \( \varepsilon > 0 \), it is possible to find positive integers \( m \) and \( n \) and constants \( \alpha_{k,\ell}, \ k = 1, 2, \ldots, m, \ell = 1, 2, \ldots, n \), such that

\[
\int_0^a \int_0^b \left\{ \left( \frac{\partial u}{\partial x} - \pi \sum_{k=1}^m \sum_{\ell=1}^n \alpha_{k,\ell} \cos \frac{k\pi x}{a} \sin \frac{\ell \pi y}{b} \right)^2 + \left( \frac{\partial u}{\partial y} - \pi \sum_{k=1}^m \sum_{\ell=1}^n \ell \alpha_{k,\ell} \sin \frac{k\pi x}{a} \cos \frac{\ell \pi y}{b} \right)^2 \right\} \, dx \, dy < \varepsilon^2
\]

**EXERCISES 10.7**

1. Prove that the sequence

\[ (x - a)^m (b - x)^m x^k, \quad k = 0, 1, 2, \ldots \]

is complete in the energy norm of the operator

\[ (-1)^m \frac{d^{2m}u}{dx^{2m}} \]

on the interval \([a, b]\), under the boundary conditions

\[ u^{(k)}(a) = u^{(k)}(b) = 0, \quad k = 0, 1, \ldots, m - 1 \]

given that the polynomials \( \phi_i(x) = x^i \) are complete in \( L_2(a, b) \).
10.8 RITZ METHOD FOR EQUILIBRIUM PROBLEMS

The solution of the equation
\[ A u = f \]  
(10.8.1)
where \( A \) is a positive definite operator, reduces to finding a function in the domain \( D_A \) of \( A \) that minimizes the functional
\[ F(u) = (A u, u) - 2 \langle u, f \rangle \]  
(10.8.2)

Selection of Coordinate Functions

An approximate solution to this problem may be represented as a linear combination of a sequence of functions
\[ \phi_1(x), \phi_2(x), \ldots, \phi_n(x), \ldots \]  
(10.8.3)
called coordinate functions, which belong to the domain \( D_A \) of the operator. In order to yield good results, this sequence should be subject to two conditions:

1. The sequence is in \( D_A \) and is complete in energy.
2. For any \( n \), the functions \( \phi_1(x), \phi_2(x), \ldots, \phi_n(x) \) are linearly independent.

A linear combination of the first \( n \) coordinate functions
\[ u_n(x) = \sum_{j=1}^{n} a_j \phi_j(x) \]  
(10.8.4)

with numerical coefficients \( a_j \) is sought to approximate the solution of Eq. 10.8.1. Since the \( \phi_j(x) \) are known, substituting \( u_n(x) \) for \( u(x) \) in the energy functional of Eq. 10.8.2 makes \( F(u_n) \) a function of the variables \( a_1, a_2, \ldots, a_n \); i.e.,
\[ F(u_n) = \left( \sum_{j=1}^{n} a_j A \phi_j, \sum_{k=1}^{n} a_k \phi_k \right) - 2 \sum_{j=1}^{n} a_j \phi_j, f \right) \]
\[ = \sum_{j,k=1}^{n} (A \phi_j, \phi_k) a_j a_k - 2 \sum_{j=1}^{n} (\phi_j, f) a_j \]  
(10.8.5)

Ritz Approximate Solutions

To create an approximation of the function \( u \) that minimizes \( F(u) \), the coefficients \( a_j \) are selected so that the function of Eq. 10.8.5 is a minimum. This requires that the \( a_j \) satisfy the conditions
\[ \frac{\partial F(u_n)}{\partial a_i} = 0, \quad i = 1, 2, \ldots, n \]  
(10.8.6)
Since the operator \( A \) is positive definite and the \( \phi_i \) are linearly independent, if
\[
a = [a_1, \ldots, a_n]^T \neq 0, \sum_{i=1}^{n} a_i \phi_i \neq 0 \text{ and}
\]
\[
\sum_{j,k=1}^{n} (A \phi_j, \phi_k) a_j a_k = \left( A \sum_{i=1}^{n} a_i \phi_i, \sum_{i=1}^{n} a_i \phi_i \right) > 0
\]

Thus, the quadratic form of Eq. 10.8.5 is positive definite, and the coefficients
\( a_1, a_2, \ldots, a_n \) that satisfy Eqs. 10.8.6 yield a minimum of \( F(u_n) \).

In order to evaluate the derivatives on the left side of Eq. 10.8.6, the function \( F(u_n) \)
can be rewritten from Eq. 10.8.5 as
\[
F(u_n) = (A \phi_1, \phi_1) a_1^2 + \sum_{k \neq i} (A \phi_i, \phi_k) a_j a_k
\]
\[
+ \sum_{j \neq i} (A \phi_j, \phi_i) a_j a_i - 2 (f, \phi_i) a_i + C
\]

where \( i \) is a typical index, \( 1 \leq i \leq n \), and \( C \) denotes a group of terms that do not contain \( a_i \).

Using symmetry of \( A \) and combining terms, Eq. 10.8.7 may be rewritten as
\[
F(u_n) = (A \phi_1, \phi_1) a_1^2 + 2 \sum_{k \neq i} (A \phi_i, \phi_k) a_j a_k - 2 (f, \phi_i) a_i + C
\]

Now, Eq. 10.8.6 is just
\[
\frac{\partial F(u_n)}{\partial a_i} = \sum_{k=1}^{n} 2 (A \phi_i, \phi_k) a_k - 2 (f, \phi_i) = 0
\]
or,
\[
\sum_{k=1}^{n} \left[ \phi_i, \phi_k \right]_A a_k = (f, \phi_i), \quad i = 1, 2, \ldots, n
\]

Writing this system in detail yields the **Ritz equations**,\n\[
\begin{align*}
[\phi_1, \phi_1]_A a_1 + [\phi_1, \phi_2]_A a_2 + \ldots + [\phi_1, \phi_n]_A a_n &= (f, \phi_1) \\
[\phi_2, \phi_1]_A a_1 + [\phi_2, \phi_2]_A a_2 + \ldots + [\phi_2, \phi_n]_A a_n &= (f, \phi_2) \\
&\vdots \\
[\phi_n, \phi_1]_A a_1 + [\phi_n, \phi_2]_A a_2 + \ldots + [\phi_n, \phi_n]_A a_n &= (f, \phi_n)
\end{align*}
\]
The determinant of the coefficient matrix of this system of equations is

$$\begin{vmatrix}
[\phi_1, \phi_1]_A & [\phi_1, \phi_2]_A & \cdots & [\phi_1, \phi_n]_A \\
[\phi_2, \phi_1]_A & [\phi_2, \phi_2]_A & \cdots & [\phi_2, \phi_n]_A \\
\vdots & \vdots & \ddots & \vdots \\
[\phi_n, \phi_1]_A & [\phi_n, \phi_2]_A & \cdots & [\phi_n, \phi_n]_A
\end{vmatrix} \neq 0 \quad (10.8.10)$$

which is the **Gram determinant** of the linearly independent functions $\phi_1(x), \phi_2(x), \ldots, \phi_n(x)$ with respect to the $A$-scalar product, so by Theorem 2.4.3 it is different from zero. It follows that the system of Ritz equations of Eq. 10.8.8 or 10.8.9 has a unique solution for the $a_i$, provided the operator $A$ is positive definite. By finding the coefficients $a_1, a_2, \ldots, a_n$ and substituting them in Eq. 10.8.4, the function $u_n(x)$ is obtained. It is called the **Ritz approximate solution** of Eq. 10.8.1.

**Theorem 10.8.1.** If $A$ is positive bounded below and if the coordinate functions used are complete in energy in $D_A$, the Ritz approximate solution $u_n$ of the operator equation of Eq. 10.8.1 constitutes a minimizing sequence for the functional of Eq. 10.8.2.

For proof, denote the least value of the functional of Eq. 10.8.2 by $d$; i.e.,

$$d = \min_{u \in H_A} F(u) = -\| u_0 \|_A^2$$

where $u_0$ is the generalized solution of Eq. 10.8.1 in $H_A$. Note that $F(u_n) \geq d$; i.e., the Ritz method gives an upper bound for the value of the functional $F$. This follows from the fact that $u_n(x)$ is in $D_A$, since it is a linear combination of the functions $\phi_1(x), \phi_2(x), \ldots, \phi_n(x)$ that are in $D_A$, and the quantity $d$ is the greatest lower bound of the functional $F(u)$. For any $\varepsilon > 0$, by definition of the greatest lower bound and of $H_A$, there exists a function $v(x)$ in $D_A$ such that $d < F(v) < d + \varepsilon/2$.

Since $F(v) = [v, v]_A - (f, v)$ and $f = Au_0$,

$$F(v) = [v, v]_A - [v, u_0]_A$$

$$= [v - u_0, v - u_0]_A - [u_0, u_0]_A$$

Thus, for any $v_n = \alpha_1\phi_1 + \alpha_2\phi_2 + \ldots + \alpha_n\phi_n$,

$$F(v_n) - F(v) = \| v_n - u_0 \|_A^2 - \| v - u_0 \|_A^2$$

$$= [\| v_n - u_0 \|_A - \| v - u_0 \|_A]$$

$$\times [\| v_n - u_0 \|_A - \| v - u_0 \|_A]$$

By the triangle inequality,
so,
\[ F(v_n) - F(v) \leq \left[ \| v_n - u_0 \|_A + \| v - u_0 \|_A \right] \| v_n - v \|_A \]

Since the sequence \( \{ \phi_n \} \) is complete in energy, for any \( \varepsilon > 0 \), it is possible to choose an integer \( n \) and constants \( \alpha_1, \alpha_2, \ldots, \alpha_n \) such that
\[ \| v_n - v \|_A < \frac{\varepsilon}{k} \]
The value of \( k \) will be chosen later. Since \( \| v_n \|_A - \| v \|_A \leq \| v_n - v \|_A \),
\[ \| v_n \|_A < \| v \|_A + \frac{\varepsilon}{k} \]
and
\[ F(v_n) - F(v) < \left[ 2 \| v \|_A + 2 \| u_0 \|_A + \frac{\varepsilon}{k} \right] \frac{\varepsilon}{k} \]
By selecting \( k \) such that
\[ \frac{1}{k} \left[ 2 \| v \|_A + 2 \| u_0 \|_A + \frac{\varepsilon}{k} \right] < \frac{1}{2} \]
the inequality \( F(v_n) - F(v) < \varepsilon/2 \) is obtained. Hence, it follows that
\[ d \leq F(v_n) \leq F(v) + \frac{\varepsilon}{2} < d + \varepsilon \]
Let \( u_n \) be the function constructed by the Ritz method. Then,
\[ d \leq F(u_n) \leq F(v_n) \]
or,
\[ d \leq F(u_n) \leq d + \varepsilon \]
Letting \( \varepsilon \to 0 \), \( F(u_n) \to d \); i.e., the sequence \( \{ u_n \} \) is a minimizing sequence. \( \blacksquare \)

From the theorem just proved and from Theorem 10.7.1, it follows that Ritz approximate solutions converge to the solution of Eq. 10.8.1. To secure high accuracy, it may be necessary to take a large number of coordinate functions. This leads to the necessity of solving Eqs. 10.8.8 with a large number of equations and unknowns.

Consider next the question: to what extent are the stipulations of completeness and linear independence that have been imposed on the coordinate functions necessary? The
first stipulation is not entirely necessary. In fact, it is not necessary that every function from \( D_A \) be approximated in the energy sense by a linear combination of the coordinate functions selected. It is sufficient that the solution permits such an approximation. In the proof of Theorem 10.8.1, it is possible to take for \( v \) an appropriate linear combination of coordinate functions and the whole proof holds true. If it is known beforehand that the solution sought belongs to a class of functions that is more restricted than \( D_A \), then it suffices that the sequence of coordinate functions be complete in this class. For example, if it is known that the function sought is even in a certain independent variable, then the coordinate functions may be selected as even functions of that variable. In this case, it suffices that the system of coordinate functions selected be complete with respect to even functions from \( D_A \).

Consider next what would happen if the completeness condition were satisfied, but the coordinate functions were linearly dependent. Suppose, for simplicity, that there is only one function in Eq. 10.8.3 that is linearly dependent on a finite number of other functions of this system, and that this function is different from zero. In this case it is possible, by substituting linear combinations of other coordinate functions for a certain coordinate function, to eliminate a dependent function, say \( \phi_1(x) \), and that the functions \( \phi_2(x), \phi_3(x), \ldots \) are linearly independent. Equation 10.8.4 can then be written in the form

\[
    u_n(x) = \sum_{i=2}^{n} a_j \phi_j(x)
\]

and the remainder of the analysis follows without alteration. In particular, the coefficients \( a_2, a_3, \ldots, a_n \) are uniquely defined. Hence, it follows that the coefficients \( a_1, a_2, \ldots, a_n \) are also defined, but not uniquely, since the coefficient \( a_1 \) can be specified arbitrarily. This means that Eqs. 10.8.8 can be solved, although not uniquely, and any solution leads to the same function \( u_n(x) \). Thus, the presence of one or several coordinate functions that are linearly dependent on the others does not alter the rule for constructing a Ritz approximate solution, although in this case the solution of the Ritz equations of Eq. 10.8.9 is more complex, because the coefficient matrix is singular.

In practical calculations, cases are encountered in which the coordinate functions are "quasi-linearly dependent"; i.e., the determinant of Eq. 10.8.10 is nearly zero, while the coordinate functions are technically linearly independent. In such a case, the relative error in the evaluation of the numerical values of the coefficients \( a_1, a_2, \ldots, a_n \) fluctuates markedly with the degree of accuracy of the coefficients and of the independent terms in Eqs. 10.8.8, but these fluctuations have little effect on the function \( u_n(x) \). Thus, the presence of linearly dependent or quasilinearly dependent functions does not preclude the use of the Ritz method. It should, however, be borne in mind that it complicates the nature and solution of the Ritz equations.

Finally, it is important to note that for the Ritz method, it is only required that the coordinate functions \( \{ \phi_i \} \) be in \( H_A \). Hence they need not satisfy natural boundary conditions.
Theorem 10.8.2. If the coordinate functions used are complete in $H_A$ and the operator $A$ is positive bounded below, the Ritz approximate solution of the operator equation $Au = f$, on $H_A$, constitutes a minimizing sequence for $F(u) = [u, u]_A - 2 (u, f)$.

The proof given in the foregoing remains unchanged.

EXERCISES 10.8

1. Show that if an operator $A$ is positive definite and $\phi_i, i = 1, \ldots, n$, are linearly independent in $D_A$, then the matrix $[ (A \phi_i, \phi_j) ] = \{ [\phi_i, \phi_j]_A \}$ is positive definite.

2. Given the result of Exercise 1, show that Eq. 10.8.8 has a unique solution $\bar{a}_1, \ldots, \bar{a}_k$ and that $F(\bar{u}_n) \leq F(u_n)$, where $u_n$ is given by Eq. 10.8.4 and $\bar{u}_n = \sum_{i=1}^{n} \bar{a}_i \phi_i(x)$.

3. Let $\{ \phi_i \}$ be orthonormal eigenfunctions of a positive definite operator $A$. Simplify and solve the Ritz equations for the operator equation $Au = f$.

10.9 RITZ METHOD FOR EIGENVALUE PROBLEMS

Let $A$ be a positive bounded below operator. Clearly,

$$d = \inf_{u \neq 0} \frac{(Au, u)}{(u, u)}$$  \hspace{1cm} (10.9.1)

is the lowest eigenvalue of operator $A$, if there exists a function $u_0 \neq 0$ in $D_A$ such that

$$d = \frac{(Au_0, u_0)}{(u_0, u_0)}$$

Assuming that such a function exists, the determination of the lowest eigenvalue of operator $A$ reduces to finding the greatest lower bound of the functional of Eq. 10.9.1. Equivalently, this is the determination of the greatest lower bound

$$\inf_{u \in D_A} (Au, u)$$  \hspace{1cm} (10.9.2)

under the supplementary condition

$$(u, u) = 1$$  \hspace{1cm} (10.9.3)

This problem can also be solved by the Ritz method.
Sec. 10.9  Ritz Method for Eigenvalue Problems

Selection of Coordinate Functions

Consider a sequence of coordinate functions \( \{ \phi_i(x) \} \) in \( D_A \), subject to the requirements:

1. The sequence of coordinate functions is complete in energy.
2. For any \( n \), the functions \( \phi_1(x), \phi_2(x), \ldots, \phi_n(x) \), are linearly independent.

The goal is to approximate an eigenfunction as

\[
   u_n(x) = \sum_{k=1}^{n} a_k \phi_k(x)
\]

where \( a_k \) are coefficients to be selected. Motivated by Eqs. 10.9.1 and 10.9.2, these coefficients are to be chosen so that \( u_n \) satisfies Eq. 10.9.3 and such that the quantity \( (A u_n, u_n) \) is a minimum. This requires the minimization of

\[
   (A u_n, u_n) = \sum_{k,i=1}^{n} (A \phi_k, \phi_i) a_k a_i
\]

subject to the condition

\[
   (u_n, u_n) = \sum_{k,i=1}^{n} (\phi_k, \phi_i) a_k a_i = 1
\]

Ritz Approximate Solutions

In order to solve this constrained minimization problem, Lagrange multipliers are introduced to construct the function \( \Phi = (A u_n, u_n) - \lambda [(u_n, u_n) - 1] \), where \( \lambda \) is a multiplier. As a necessary condition [8] for minimization of the function of Eq. 10.9.4, subject to the constraint of Eq. 10.9.5, the partial derivatives of \( \Phi \) with respect to the \( a_i \) must be zero. This leads to the system of equations

\[
   \sum_{k=1}^{n} a_k [ (A \phi_k, \phi_i) - \lambda (\phi_k, \phi_i) ] = 0, \quad i = 1, 2, \ldots, n
\]

Equations 10.9.6 are linear and homogeneous in the unknowns \( a_i \), which can not all be zero. It follows that the determinant of the coefficient matrix of Eq. 10.9.6 must be zero. This gives the following characteristic equation for \( \lambda \):

\[
   \begin{vmatrix}
   (A \phi_1, \phi_1) - \lambda (\phi_1, \phi_1) & (A \phi_2, \phi_1) - \lambda (\phi_2, \phi_1) & \cdots & (A \phi_n, \phi_1) - \lambda (\phi_n, \phi_1) \\
   (A \phi_1, \phi_2) - \lambda (\phi_1, \phi_2) & (A \phi_2, \phi_2) - \lambda (\phi_2, \phi_2) & \cdots & (A \phi_n, \phi_2) - \lambda (\phi_n, \phi_2) \\
   \vdots & \vdots & \ddots & \vdots \\
   (A \phi_1, \phi_n) - \lambda (\phi_1, \phi_n) & (A \phi_2, \phi_n) - \lambda (\phi_2, \phi_n) & \cdots & (A \phi_n, \phi_n) - \lambda (\phi_n, \phi_n)
   \end{vmatrix}
   = 0
\]

(10.9.7)
If the sequence \( \{ \phi_n \} \) is orthonormal; i.e., \( (\phi_i, \phi_j) = \delta_{ij} \), then Eq. 10.9.7 simplifies to

\[
\begin{bmatrix}
( A\phi_1, \phi_1 ) - \lambda & ( A\phi_2, \phi_1 ) & \ldots & ( A\phi_n, \phi_1 ) \\
( A\phi_1, \phi_2 ) & ( A\phi_2, \phi_2 ) - \lambda & \ldots & ( A\phi_n, \phi_2 ) \\
\vdots & \vdots & \ddots & \vdots \\
( A\phi_1, \phi_n ) & ( A\phi_2, \phi_n ) & \ldots & ( A\phi_n, \phi_n ) - \lambda
\end{bmatrix} = 0 \quad (10.9.8)
\]

As noted previously, the coordinate functions \( \phi_1, \phi_2, \ldots, \phi_n \) are to be selected as linearly independent. Equation 10.9.7 will then be of \( n \)th degree, since the coefficient of \((-1)^n \lambda^n\) is the Gram determinant of the functions \( \phi_1, \phi_2, \ldots, \phi_n \) relative to the A-scalar product. Hence, it follows that Eq. 10.9.7 has precisely \( n \) roots. Let \( \lambda_0 \) be any one of these roots. Substituting it in Eq. 10.9.6, the system has a non-trivial solution \( a_k^{(0)} \), \( k = 1, 2, \ldots, n \), that also satisfies Eq. 10.9.5.

Substituting \( \lambda = \lambda_0 \) and \( a_k = a_k^{(0)} \) in Eq. 10.9.6,

\[
\sum_{k=1}^{n} a_k^{(0)} ( A\phi_k, \phi_i ) = \lambda_0 \sum_{k=1}^{n} a_k^{(0)} ( \phi_k, \phi_i ), \quad i = 1, 2, \ldots, n \quad (10.9.9)
\]

Multiplying by \( a_i^{(0)} \) and summing over \( i \),

\[
\sum_{k,i=1}^{n} ( A\phi_k, \phi_i ) a_k^{(0)} a_i^{(0)} = \lambda_0 \sum_{k,i=1}^{n} ( \phi_k, \phi_i ) a_k^{(0)} a_i^{(0)}
\]

By virtue of Eq. 10.9.5, the right side of the above equation equals \( \lambda_0 \) and its left side equals \( ( A u_n^{(0)}, u_n^{(0)} ) \), where

\[
u_n^{(0)} = \sum_{k=1}^{n} a_k^{(0)} \phi_k
\]

Thus,

\[
\lambda_0 = ( A u_n^{(0)}, u_n^{(0)} ) \quad (10.9.10)
\]

Equation 10.9.6 shows that the roots of Eq. 10.9.7 are real, if \( A \) is symmetric. Further, one of the functions \( u_n^{(0)} \) minimizes the function of Eq. 10.9.4, subject to the constraint of Eq. 10.9.5. Equation 10.9.10 shows that this minimum equals the least of the roots of Eq. 10.9.7. The minimum thus constructed is denoted \( \lambda_n^{(0)} \). It does not increase with increasing \( n \) and it is bounded below by \( d \). Thus, \( \lambda_n^{(0)} \) tends to a limit as \( n \to \infty \), which is greater than or equal to \( d \).
Theorem 10.9.1. The sequence \( \lambda_n^{(0)} \) generated by the Ritz method for the positive bounded below operator equation \( Au = \lambda u \) converges monotonically to \( d \) of Eq. 10.9.1; i.e., \( \lim_{n \to \infty} \lambda_n^{(0)} = d, \lambda_n^{(0)} \geq \lambda_{n+k}^{(0)} \geq d \) for all \( n \) and \( k \geq 1 \).

To prove this, note first that for any \( \varepsilon > 0 \) there exists, by the definition of greatest lower bound, a function \( \bar{u} \) in \( D_A \) such that \( (\bar{u}, \bar{u}) = 1 \) and

\[
d \leq (Au, \bar{u}) < d + \varepsilon
\]

Equivalently,

\[
\sqrt{d} \leq \| \bar{u} \|_A < \sqrt{d + \varepsilon}
\]  

(10.9.11)

Since the sequence \{ \( \phi_n(x) \) \} is complete in energy, it is possible to find a function \( \bar{u}_N \) of the form

\[
\bar{u}_N = \sum_{k=1}^{N} b_k \phi_k
\]

such that \( \| \bar{u} - \bar{u}_N \|_A < \sqrt{\varepsilon} \). Using the triangle inequality, \( \| \bar{u}_N \|_A - \| \bar{u} \|_A \leq \| \bar{u} - \bar{u}_N \|_A < \sqrt{\varepsilon} \). This and Eq. 10.9.11 yield

\[
\| \bar{u}_N \|_A \leq \| \bar{u} \|_A + \sqrt{\varepsilon} < \sqrt{d + \varepsilon} + \sqrt{\varepsilon}
\]

Squaring both sides of this inequality,

\[
\| \bar{u}_N \|_A^2 = (A\bar{u}_N, \bar{u}_N) < \left( \sqrt{d + \varepsilon} + \sqrt{\varepsilon} \right)^2
\]  

(10.9.12)

From Eq. 10.9.1, it follows that \( \| \bar{u}_N - \bar{u} \| \leq (1/\sqrt{d}) \| \bar{u}_N - \bar{u} \|_A < \sqrt{\varepsilon/d} \). Hence, \( \| \bar{u}_N \| \geq \| \bar{u} \| - \sqrt{\varepsilon/d} = 1 - \sqrt{\varepsilon/d} \). This result and Eqs. 10.9.1 and 10.9.12 yield

\[
d \leq \frac{(A\bar{u}_N, \bar{u}_N)}{(\bar{u}_N, \bar{u}_N)} = \frac{\| \bar{u}_N \|_A^2}{\| \bar{u}_N \|_A^2} < \frac{(\sqrt{d + \varepsilon} + \sqrt{\varepsilon})^2}{\left(1 - \sqrt{\frac{\varepsilon}{d}}\right)^2} = d + \eta
\]

where \( \eta \) tends to zero with \( \varepsilon \). Further, \( \lambda_{n_N}^{(0)} \) is the minimum of
\[
\frac{(Au_N, u_N)}{(u_N, u_N)}
\]
over all \(a_k\), where \(u_N = \sum_{k=1}^{N} a_k \phi_k\). Thus,

\[
d \leq \lambda_N^{(0)} \leq \frac{(Au_N, \bar{u}_N)}{(u_N, u_N)} < d + \eta
\]  
(10.9.13)

If \(n \geq N\), then \(\lambda_n^{(0)} \leq \lambda_N^{(0)}\) and \(d \leq \lambda_n^{(0)} < d + \eta\). Equation 10.9.13 shows that

\[
\lim_{n \to \infty} \lambda_n^{(0)} = d
\]  
(10.9.14)

To complete the proof, note that as \(N\) increases, the space spanned by the coordinate functions chosen to minimize the Rayleigh Quotient is enlarged. This yields the monotonicity property \(\lambda_n^{(0)} \geq \lambda_{n+k}^{(0)}\), for \(k \geq 1\).

Consider now the determination of the larger eigenvalues. To obtain an approximate value of the second eigenvalue, \(u_n\) is sought to minimize the scalar product of Eq. 10.9.4, subject to the constraints

\[
( u_n, u_n ) = 1
\]
(10.9.15)

\[
( u_n^{(0)}, u_n ) = \sum_{k,m=1}^{n} ( \phi_k, \phi_m ) a_k^{(0)} a_m = 0
\]

where \(u_n^{(0)} = \sum_{k=1}^{n} a_k^{(0)} \phi_k\) is the approximate value of the first normalized eigenfunction of operator \(A\).

Using Lagrange multipliers \(\lambda\) and \(\mu\) associated with the constraints of Eq. 10.9.15, the expression

\[
\Psi = (Au_n, u_n) - \lambda [ (u_n, u_n) - 1 ] - 2\mu (u_n, u_n^{(0)})
\]
is constructed and its partial derivatives with respect to the \(a_i\) are set to zero [8]. This leads to the system

\[
\sum_{k=1}^{n} \{ a_k [ (A\phi_k, \phi_i) - \lambda (\phi_k, \phi_i) ] - \mu (\phi_k, \phi_i) a_k^{(0)} \} = 0, \ i = 1, \ldots, n
\]  
(10.9.16)
Multiplying both sides of this equation by \( a_i^{(0)} \) and summing over \( i \) yields

\[
\sum_{k,i=1}^{n} a_k a_i^{(0)} \left[ (A\phi_k, \phi_i) - \lambda (\phi_k, \phi_i) \right] - \mu \sum_{k,i=1}^{n} a_k^{(0)} a_i^{(0)} (\phi_k, \phi_i) = 0
\]

(10.9.17)

The second summation is simply \( (u_n^{(0)}, u_n^{(0)}) = 1 \). Changing the order of the first summation,

\[
\sum_{i=1}^{n} a_i \sum_{k=1}^{n} a_k^{(0)} \left[ (A\phi_i, \phi_k) - \lambda (\phi_i, \phi_k) \right]
\]

(10.9.18)

The inner sum may be written as

\[
\sum_{k=1}^{n} a_k^{(0)} \left[ (\phi_k, A\phi_i) - \lambda (\phi_k, \phi_i) \right]
\]

or, since \( A \) is symmetric,

\[
\sum_{k=1}^{n} a_k^{(0)} \left[ (A\phi_k, \phi_i) - \lambda (\phi_k, \phi_i) \right]
\]

By virtue of Eq. 10.9.9, with \( a_k^{(0)} \) corresponding to \( \lambda = \lambda_n^{(0)} \), this is

\[
(\lambda_n^{(0)} - \lambda) \sum_{k=1}^{n} a_k^{(0)} (\phi_k, \phi_i) = (\lambda_n^{(0)} - \lambda) \left( \sum_{k=1}^{n} a_k^{(0)} \phi_k, \phi_i \right)
\]

\[
= (\lambda_n^{(0)} - \lambda) (u_n^{(0)}, \phi_i)
\]

\[
= (\lambda_n^{(0)} - \lambda) (\phi_i, u_n^{(0)})
\]

Equation 10.9.18 now becomes,

\[
(\lambda_n^{(0)} - \lambda) \sum_{i=1}^{n} a_i (\phi_i, u_n^{(0)}) = (\lambda_n^{(0)} - \lambda) (u_n, u_n^{(0)}) = 0
\]

where the last equality follows from Eq. 10.9.15. From Eq. 10.9.17, \( \mu = 0 \), so Eqs. 10.9.16 are identical to Eqs. 10.9.6. Hence, the minimum sought is a root of Eq. 10.9.7. Similarly, it may be shown that all roots of Eq. 10.9.7 are approximations for larger eigenvalues of \( Au = \lambda u \).
The Ritz method can also be developed for finding the minimum of the ratio $\frac{\|u\|^2_A}{\|u\|^2_2}$ over all $u \neq 0$ in $H_A$; i.e., ignoring natural boundary conditions. The coordinate functions are selected as before, with the requirement that they satisfy only principal boundary conditions. As before,

$$u_n(x) = \sum_{k=1}^{n} a_k \phi_k(x)$$

where it is required that

$$\|u_n\|^2_A = \sum_{k,i=1}^{n} a_k a_i [\phi_k, \phi_i]_A$$

should be a minimum, subject to the condition that

$$\|u_n\|^2 = \sum_{k,i=1}^{n} a_k a_i (\phi_k, \phi_i) = 1$$

The necessary conditions for a constrained minimum yield

$$\begin{vmatrix} [\phi_1, \phi_1]_A - \lambda (\phi_1, \phi_1) & [\phi_2, \phi_1]_A - \lambda (\phi_2, \phi_1) & \ldots & [\phi_n, \phi_1]_A - \lambda (\phi_n, \phi_1) \\
[\phi_1, \phi_2]_A - \lambda (\phi_1, \phi_2) & [\phi_2, \phi_2]_A - \lambda (\phi_2, \phi_2) & \ldots & [\phi_n, \phi_2]_A - \lambda (\phi_n, \phi_2) \\
\vdots & \vdots & & \vdots \\
[\phi_1, \phi_n]_A - \lambda (\phi_1, \phi_n) & [\phi_2, \phi_n]_A - \lambda (\phi_2, \phi_n) & \ldots & [\phi_n, \phi_n]_A - \lambda (\phi_n, \phi_n) 
\end{vmatrix} = 0$$

(10.9.19)

Note that Eq. 10.9.19 is equivalent to Eq. 10.9.7, if the coordinate functions are in $D_A$.

As before, the roots of Eq. 10.9.19 are approximate values of the eigenvalues of the operator and, as $n \to \infty$, the smallest root of Eq. 10.9.19 tends to the smallest eigenvalue of the operator.

Finally, consider briefly the **generalized eigenvalue problem**

$$Au = \lambda Bu$$

(10.9.20)

Assume that both operators $A$ and $B$ are positive bounded below and that $D_A$ is a subset of $D_B$. For any function in $D_A$ it is possible to define the energy in two ways, by associating it either with operator $A$ or with operator $B$. These two forms of energy are called the energy of operator $A$ or the energy of operator $B$.

The details of applying the Ritz method to Eq. 10.9.20 are left as an exercise. Let the sequence of coordinate functions $\{\phi_i\}$ be in $D_A$, be linearly independent, and be complete.
in the energy of operator $A$. Equations 10.9.6 and 10.9.7 are then replaced by the Ritz equations

$$\sum_{k=1}^{n} a_k \left[ (A\phi_k, \phi_i) - \lambda (B\phi_k, \phi_i) \right] = 0, \quad i = 1, 2, \ldots, n \quad (10.9.21)$$

and the characteristic equation

$$\begin{bmatrix}
A\phi_1, \phi_1 & \lambda (B\phi_1, \phi_1) \\
A\phi_2, \phi_1 & \lambda (B\phi_2, \phi_1) \\
\vdots & \vdots \\
A\phi_n, \phi_1 & \lambda (B\phi_n, \phi_1)
\end{bmatrix}
\begin{bmatrix}
A\phi_1, \phi_1 \\
A\phi_2, \phi_1 \\
\vdots \\
A\phi_n, \phi_1
\end{bmatrix} = 0$$

The condition that the $\phi_i$ be in $D_A$ can be relaxed by merely requiring that the coordinate functions be in $H_A$. It can be proved that they also possess finite energy for operator $B$. Instead of Eqs. 10.9.21 and 10.9.22, the Ritz equations for the eigenvalue problem $Au = \lambda Bu$ are

$$\sum_{k=1}^{n} a_k \left\{ [\phi_k, \phi_i]_A - \lambda [\phi_k, \phi_i]_B \right\} = 0, \quad i = 1, 2, \ldots, n \quad (10.9.23)$$

and the characteristic equation is

$$\begin{bmatrix}
[\phi_1, \phi_1]_A - \lambda [\phi_1, \phi_1]_B & [\phi_2, \phi_1]_A - \lambda [\phi_2, \phi_1]_B & \cdots & [\phi_n, \phi_1]_A - \lambda [\phi_n, \phi_1]_B \\
[\phi_1, \phi_2]_A - \lambda [\phi_1, \phi_2]_B & [\phi_2, \phi_2]_A - \lambda [\phi_2, \phi_2]_B & \cdots & [\phi_n, \phi_2]_A - \lambda [\phi_n, \phi_2]_B \\
\vdots & \vdots & \ddots & \vdots \\
[\phi_1, \phi_n]_A - \lambda [\phi_1, \phi_n]_B & [\phi_2, \phi_n]_A - \lambda [\phi_2, \phi_n]_B & \cdots & [\phi_n, \phi_n]_A - \lambda [\phi_n, \phi_n]_B
\end{bmatrix} = 0$$

**EXERCISES 10.9**

1. Use the result of Exercise 5 of Section 10.6 to derive the Ritz equations of Eqs. 10.9.21 and 10.9.23 for the generalized eigenvalue problem $Au = \lambda Bu$.

2. Write integral formulas for the coefficients in the Ritz equations for the eigenvalue problems

$$A_1u = \lambda g_i(x) u, \quad g_i(x) \geq g_0 > 0, \quad 0 \leq x \leq \ell$$

for the operators $A_1, A_2$, and $A_3$ of Exercise 2 of Section 10.1.
10.10 OTHER VARIATIONAL METHODS

The variational methods discussed thus far in this chapter are based on an equivalence between minimization of an energy functional and solving the linear operator equations \( Au = f \) and \( Au = \lambda Bu \). There are a number of other variational methods for approximating solutions of operator equations, but which are less restrictive in the hypotheses required in order that they are applicable. In particular, if a residual, or error term \( R = Au - f \), associated with the operator equation \( Au = f \) is formed, then the operator equation is satisfied when \( R = 0 \). This idea, and others that tend to cause the equation to be satisfied, form the basis for many variational methods.

Two commonly employed variational methods that are based on these ideas are summarized in this section. The presentation here is limited to formulation and application of these methods and does not treat their convergence properties, which are developed in Refs. 18 and 21.

The Method of Least Squares

Consider first the linear operator equation

\[
Au = f
\]

(10.10.1)

for \( u \) in \( D_A \). The operator \( A \) is not necessarily symmetric or positive definite. It is clear that a function in \( D_A \) is the solution of Eq. 10.10.1 if and only if the norm of the residual or error function \( R = Au - f \) is zero. This is equivalent to stating that the solution of Eq. 10.1.1 minimizes

\[
\| R \|^2 = \| Au - f \|^2
\]

(10.10.2)

Selecting a sequence \( \{ \phi_k(x) \} \) in \( D_A \), an approximate solution of the form

\[
u_n = \sum_{k=1}^{n} a_k \phi_k(x)
\]

(10.10.3)

is constructed by choosing the \( a_k \) to minimize the function of Eq. 10.10.2. This is called the least square method of solution of Eq. 10.10.1. As for the Ritz method, the coordinate functions selected should be complete, in order that it is reasonable to expect that they can provide a good approximation of the solution. Substituting Eq. 10.10.3 into the functional of Eq. 10.10.2, the following function is obtained, which depends only on the undetermined coefficients:

\[
\| Au_n - f \|^2 = \left\| \sum_{k=1}^{n} a_k A \phi_k - f \right\|^2
\]

\[
= \left( \sum_{k=1}^{n} a_k A \phi_k - f, \sum_{j=1}^{n} a_j A \phi_j - f \right)
\]

(10.10.4)
In order for this function to be a minimum, it is necessary that its derivative with respect to each of the coefficients $a_i$ be zero. This yields

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

(10.10.5)

This set of linear equations can now be solved for the coefficients $a_j$ to construct an approximate solution of Eq. 10.10.3. This approximation is called the least square approximate solution of Eq. 10.10.1.

It is shown in Ref. 21 that

(a) The least square approximation converges more slowly than the Ritz approximation, when the Ritz method applies.
(b) $Au_n$ approaches $f$ in the $L_2$ sense, which is not necessarily the case in the Ritz method.

While the least square method is not as efficient as the Ritz method, when the Ritz method applies, the least square method can be used for a much broader class of problems. For example, consider the nonlinear operator equation

$$Bu = f$$

(10.10.6)

where $u$ is in the linear space $D_B$. In this case, the least square method can still be applied, using the approximation

$$u_n = \sum_{k=1}^{n} a_k \phi_k$$

(10.10.7)

but the equation

$$\frac{\partial}{\partial a_i} \left\| Bu_i - f \right\|^2 = \frac{\partial}{\partial a_i} \left\| B \left( \sum_{k=1}^{n} a_k \phi_k \right) - f \right\|^2 = 0, \quad i = 1, 2, \ldots, n$$

(10.10.8)

is nonlinear in the unknown coefficients $a_k$. This creates the rather difficult task of solving a system of nonlinear algebraic equations. Formidable as this problem is, it is still solvable by approximate methods of numerical analysis, such as the Newton method for solution of nonlinear algebraic equations [8].

**The Galerkin Method**

Consider again the linear operator equation

$$Au = f$$

(10.10.9)

for $u$ in $D_A$, where the operator $A$ is linear, but it may not be symmetric or positive definite. A function $u$ in $D_A$ is sought such that the residual
R = Au - f \quad (10.10.10)

is zero.

If a sequence \( \{ \phi_k(x) \} \) in \( D_A \) is complete in \( L_2 \), then \( R = 0 \) if and only if

\[
( \phi_i, R ) = 0, \quad i = 1, 2, \ldots
\]

(10.10.11)

To use this fact, consider an approximate solution of the form

\[
u_n = \sum_{k=1}^{n} a_k \phi_k \quad (10.10.12)\]

Substituting this approximation into Eqs. 10.10.10 and 10.10.11 yields the linear equations

\[
\left( \phi_i, \sum_{k=1}^{n} a_k A \phi_k - f \right) = 0, \quad i = 1, 2, \ldots, n
\]

in the \( a_k \), or

\[
\sum_{k=1}^{n} a_k ( A \phi_k, \phi_i ) = ( f, \phi_i ), \quad i = 1, 2, \ldots, n
\]

(10.10.13)

These equations serve to determine the coefficients \( a_k \) in the approximation. This process is called the \textit{Galerkin method} of finding an approximate solution.

Note that if the operator \( A \) is symmetric and positive definite, Eqs. 10.10.13 are precisely the Ritz equations of Eqs. 10.8.8. Thus, the Galerkin method reduces to the Ritz method, when the Ritz method applies. The Galerkin method, however, is applicable to a much broader class of problems than is the Ritz method.

Likewise, for the eigenvalue problem \( Au = \lambda Bu \), the residual

\[
\bar{R} = Au - \lambda Bu \quad (10.10.14)
\]

is to be zero. With the approximation of Eq. 10.10.12 and the conditions of Eqs. 10.10.11,

\[
\sum_{k=1}^{n} \left\{ ( A \phi_k, \phi_i ) - \lambda ( B \phi_k, \phi_i ) \right\} a_k = 0, \quad i = 1, 2, \ldots, n
\]

(10.10.15)

as conditions that determine the unknown coefficients \( a_k \) and the eigenvalue \( \lambda \). It is clear that the determinant of the coefficient matrix in Eq. 10.10.15 must be zero. This provides an approximate eigenvalue and the associated approximate eigenvector.

Note that Eqs. 10.10.15 are the Ritz equations of Eqs. 10.9.21 when the operators \( A \) and \( B \) are positive definite and symmetric. Thus, the Galerkin method for the eigenvalue problem also reduces to the Ritz method, when the Ritz method applies.
Note finally that if an operator B is nonlinear, with a linear domain \( D_B \), then the Galerkin method applies and results in the following nonlinear equations for the coefficients \( a_k \):

\[
\left( B \left( \sum_{k=1}^{n} a_k \phi_k \right) - f, \phi_i \right) = 0, \quad i = 1, 2, \ldots, n
\]  

as an approximate solution in \( D_B \) of the operator equation

\[
Bu = f
\]  

The formidable problem of solving these nonlinear algebraic equations for the unknown coefficients \( a_k \) remains.

**EXERCISES 10.10**

1. Consider the boundary-value problem

\[
- \frac{d}{dx} \left( x^2 \frac{du}{dx} \right) + u = 2, \quad 0 < x < 1
\]

\( u(0) = u(1) = 1 \)

(a) Formulate this problem as a linear operator equation.

(b) Select a set of coordinate functions that can be used in construction of an approximate solution.

(c) Using the first two coordinate functions selected in (b), find an approximate solution using the Galerkin method.

2. Consider the operator equation

\[
A u = 4 \frac{d^4 u}{dx^4} - \frac{d}{dx} \left( x^2 \frac{du}{dx} \right) + e^x u = \sin x
\]

\( D_A = \{ u \in C^4(0,1): u(0) = u'(0) = u(1) = u''(1) = 0 \} \)

(a) Determine rigorously which boundary conditions are principal.

(b) Describe the energy space \( H_A \) of the operator.

(c) Using the Galerkin method, find a two term approximate solution for the following coordinate functions in \( H_A \):

\[
\phi_1(x) = x^2 \ (x - 1), \quad \phi_2(x) = x^3 \ (x - 1)^2, \ldots, \quad \phi_n(x) = x^{n+1} \ (x - 1)^n
\]

3. Derive equations for the least square solution of

\[
- u'' + u' + u = f(x), \quad 0 < x < \ell
\]

\( u(0) = u(\ell) = 0 \)