Chapter VII

Optimization and Approximation Topics

1 Dirichlet's Principle

When we considered elliptic boundary value problems in Chapter III we found it useful to pose them in a weak form. For example, the Dirichlet problem

$$\left.\begin{aligned} -\Delta_n u(x) &= F(x)\ , \qquad x \in G\ , \\ u(s) &= 0\ , \qquad s \in \partial G \end{aligned}\right\} \tag{1.1}$$

on a bounded open set G in $\mathbb{R}^n$ was posed (and solved) in the form

$$u \in H_0^1(G)\ ; \quad \int_G \nabla u \cdot \nabla v\, dx = \int_G F(x)v(x)\, dx\ , \qquad v \in H_0^1(G)\ . \tag{1.2}$$

In the process of formulating certain problems of mathematical physics as boundary value problems of the type (1.1), integrals of the form appearing in (1.2) arise naturally. Specifically, in describing the displacement $u(x)$ at a point $x \in G$ of a stretched string ($n = 1$) or membrane ($n = 2$) resulting from a unit tension and distributed external force $F(x)$, we find the *potential energy* is given by

$$E(u) = \left(\tfrac{1}{2}\right) \int_G |\nabla u(x)|^2\, dx - \int_G F(x)u(x)\, dx\ . \tag{1.3}$$

Dirichlet's principle is the statement that the solution u of (1.2) is that function in $H_0^1(G)$ at which the functional $E(\cdot)$ attains its minimum. That

is, u is the solution of

$$u \in H_0^1(G) : E(u) \leq E(v) \ , \qquad v \in H_0^1(G) \ . \tag{1.4}$$

To prove that (1.4) characterizes u, we need only to note that for each $v \in H_0^1(G)$

$$E(u+v) - E(u) = \int_G (\nabla u \cdot \nabla v - Fv)\, dx + \left(\tfrac{1}{2}\right) \int_G |\nabla v|^2\, dx$$

and the first term vanishes because of (1.2). Thus $E(u+v) \geq E(u)$ and equality holds only if $v \equiv 0$.

The preceding remarks suggest an alternate proof of the existence of a solution of (1.2), hence, of (1.1). Namely, we seek the element u of $H_0^1(G)$ at which the energy function $E(\cdot)$ attains its minimum, then show that u is the solution of (1.2). This program is carried out in Section 2 where we minimize functions more general than (1.3) over closed convex subsets of Hilbert space. These more general functions permit us to solve some nonlinear elliptic boundary value problems.

By considering convex sets instead of subspaces we obtain some elementary results on unilateral boundary value problems. These arise in applications where the solution is subjected to a one-sided constraint, e.g., $u(x) \geq 0$, and their solutions are characterized by variational inequalities. These topics are presented in Section 3, and in Section 4 we give a brief discussion of some optimal control problems for elliptic boundary value problems.

Finally, Dirichlet's principle provides a means of numerically approximating the solution of (1.2). We pick a convenient finite-dimensional subspace of $H_0^1(G)$ and minimize $E(\cdot)$ over this subspace. This is the Rayleigh-Ritz method and leads to an approximate algebraic problem for (1.2). This method is described in Section 5, and in Section 6 we shall obtain related approximation procedures for evolution equations of first or second order.

2 Minimization of Convex Functions

Suppose F is a real-valued function defined on a closed interval K (possibly infinite). If F is continuous and if either K is bounded or $F(x) \to +\infty$ as $|x| \to +\infty$, then F attains its minimum value at some point of K. This result will be extended to certain real-valued functions on Hilbert space and the notions developed will be extremely useful in the remainder of this

chapter. An essential point is to characterize the minimum by the derivative of F. Throughout this section V is a real separable Hilbert space, K is a non-empty subset of V and $F : K \to \mathbb{R}$ is a function.

2.1

We recall from Section I.6 that the space V is weakly (sequentially) compact. It is worthwhile to consider subsets of V which inherit this property. Thus, K is called *weakly* (*sequentially*) *closed* if the limit of every weakly convergent sequence from K is contained in K. Since convergence (in norm) implies weak convergence, a weakly closed set is necessarily closed.

Lemma 2.1 *If K is closed and convex (cf. Section I.4.2), then it is weakly closed.*

Proof: Let x be a vector not in K. From Theorem I.4.3 there is an $x_0 \in K$ which is closest to x. By translation, if necessary, we may suppose $(x_0 + x)/2 = \theta$, i.e., $x = -x_0$. Clearly $(x, x_0) < 0$ so we need to show that $(z, x_0) \geq 0$ for all $z \in K$; from this the desired result follows easily. Since K is convex, the function $\varphi : [0, 1] \to \mathbb{R}$ given by

$$\varphi(t) = \|(1-t)x_0 + tz - x\|_V^2 \ , \qquad 0 \leq t \leq 1 \ ,$$

has its minimum at $t = 0$. Hence, the right-derivative $\varphi^+(0)$ is non-negative, i.e.,

$$(x_0 - x, z - x_0) \geq 0 \ .$$

Since $x = -x_0$, this gives $(x_0, z) \geq \|x_0\|_V^2 > 0$.

The preceding result and Theorem I.6.2 show that each closed, convex and bounded subset of V is weakly sequentially compact. We shall need to consider situations in which K is not bounded (e.g., $K = V$); the following is then appropriate.

Definition. The function F has the *growth property* at $x \in K$ if, for some $R > 0$, $y \in K$ and $\|y - x\| \geq R$ implies $F(y) > F(x)$.

The continuity requirement that is adequate for our purposes is the following.

Definition. The function $F : K \to \mathbb{R}$ is *weakly lower-semi-continuous* at $x \in K$ if for every sequence $\{x_n\}$ in K which weakly converges to $x \in K$

we have $F(x) \leq \liminf F(x_n)$. [Recall that for any sequence $\{a_n\}$ in $\mathbb{R}$, $\liminf(a_n) \equiv \sup_{k\geq 0}(\inf_{n\geq k}(a_n))$.]

Theorem 2.2 *Let K be closed and convex and $F : K \to \mathbb{R}$ be weakly lower-semi-continuous at every point of K. If* (a) *K is bounded or if* (b) *F has the growth property at some point in K, then there exists an $x_0 \in K$ such that $F(x_0) \leq F(x)$ for all $x \in K$. That is, F attains its minimum on K.*

Proof: Let $m = \inf\{F(x) : x \in K\}$ and $\{x_n\}$ a sequence in K for which $m = \lim F(x_n)$. If (a) holds, then by weak sequential compactness there is a subsequence of $\{x_n\}$ denoted by $\{x_{n'}\}$ which converges weakly to $x_0 \in V$; Lemma 2.1 shows $x_0 \in K$. The weak lower-semi-continuity of F shows $F(x_0) \leq \liminf F(x_{n'}) = m$, hence, $F(x_0) = m$ and the result follows. For the case of (b), let F have the growth property at $z \in K$ and let $R > 0$ be such that $F(x) > F(z)$ whenever $\|z - x\| \geq R$ and $x \in K$. Then set $B \equiv \{x \in V : \|x - z\| \leq R\}$ and apply (a) to the closed, convex and bounded set $B \cap K$. The result follows from the observation $\inf\{F(x) : x \in K\} = \inf\{F(x) : x \in B \cap K\}$.

We note that if K is bounded then F has the growth property at every point of K; thus the case (b) of Theorem 2.2 includes (a) as a special case. Nevertheless, we prefer to leave Theorem 2.2 in its (possibly) more instructive form as given.

2.2

The condition that a function be weakly lower-semi-continuous is in general difficult to verify. However for those functions which are convex (see below), the lower-semi-continuity is the same for the weak and strong notions; this can be proved directly from Lemma 2.1. We shall consider a class of functions for which convexity and lower semicontinuity are easy to check and, furthermore, this class contains all examples of interest to us here.

Definition. The function $F : K \to \mathbb{R}$ is *convex* if its domain K is convex and for all $x, y \in K$ and $t \in [0, 1]$ we have

$$F(tx + (1 - t)y) \leq tF(x) + (1 - t)F(y) \ . \tag{2.1}$$

Definition. The function $F : K \to \mathbb{R}$ is *G-differentiable* at $x \in K$ if K is convex and if there is a $F'(x) \in V'$ such that

$$\lim_{t \to 0^+} \frac{1}{t}\Big[F(x + t(y - x)) - F(x)\Big] = F'(x)(y - x)$$

for all $y \in K$. $F'(x)$ is called the *G-differential* of F at x. If F is G-differentiable at every point in K, then $F' : K \to V'$ is the *gradient* of F on K and F is the *potential* of the function F'.

The G-differential $F'(x)$ is precisely the directional derivative of F at the point x in the direction toward y. The following shows how it characterizes convexity of F.

Theorem 2.3 *Let $F : K \to \mathbb{R}$ be G-differentiable on the convex set K. The following are equivalent:* (a) *F is convex,* (b) *For each pair $x, y \in K$ we have*

$$F'(x)(y - x) \le F(y) - F(x) \ . \tag{2.2}$$

(c) *For each pair $x, y \in K$ we have*

$$(F'(x) - F'(y))(x - y) \ge 0 \ . \tag{2.3}$$

Proof: If F is convex, then $F(x + t(y - x)) \le F(x) + t(F(y) - F(x))$ for $x, y \in K$ and $t \in [0, 1]$, so (2.2) follows. Thus (a) implies (b). If (b) holds, we obtain $F'(y)(x - y) \le F(x) - F(y)$ and $F(x) - F(y) \le F'(x)(x - y)$, so (c) follows.

Finally, we show (c) implies (a). Let $x, y \in K$ and define $\varphi : [0, 1] \to \mathbb{R}$ by

$$\varphi(t) = F(tx + (1 - t)y) = F(y + t(x - y)) \ , \qquad t \in [0, 1] \ .$$

Then $\varphi'(t) = F'(y + t(x - y))(x - y)$ and we have for $0 \le s < t \le 1$ the estimate

$$(\varphi'(t) - \varphi'(s))(t - s) = (F'(y + t(x - y)) - F'(y + s(x - y)))((t - s)(x - y)) \ge 0$$

from (c), so φ' is non-decreasing. The Mean-Value Theorem implies that

$$\frac{\varphi(1) - \varphi(t)}{1 - t} \ge \frac{\varphi(t) - \varphi(0)}{t - 0} \ , \qquad 0 < t < 1 \ .$$

Hence, $\varphi(t) \le t\varphi(1) + (1 - t)\varphi(0)$, and this is just (2.1).

Corollary 2.4 *Let F be G-differentiable and convex. Then F is weakly lower-semi-continuous on K.*

Proof: Let the sequence $\{x_n\} \subset K$ converge weakly to $x \in K$. Since $F'(x) \in V'$, we have $\lim F'(x)(x_n) = F'(x)(x)$, so from (2.2) we obtain

$$\liminf(F(x_n) - F(x)) \geq \liminf F'(x)(x_n - x) = 0 \ .$$

This shows F is weakly lower-semi-continuous at $x \in K$.

Corollary 2.5 *In the situation of Corollary 2.4, for each pair $x, y \in K$ the function*

$$t \longmapsto F'(x + t(y - x))(y - x) \ , \qquad t \in [0, 1]$$

is continuous.

Proof: We need only observe that in the proof of Theorem 2.3 the function φ' is a monotone derivative and thereby must be continuous.

2.3

Our goal is to consider the special case of Theorem 2.2 that results when F is a convex potential function. It will be convenient in the applications to have the hypothesis on F stated in terms of its gradient F'.

Lemma 2.6 *Let F be G-differentiable and convex. Suppose also we have*

$$\lim_{\|x\| \to +\infty} \frac{F'(x)(x)}{\|x\|} = +\infty \ , \qquad x \in K \ .$$

Then $\lim_{\|x\| \to \infty} F(x) = +\infty$, so F has the growth property at every point in K.

Proof: We may assume $\theta \in K$. For each $x \in K$ we obtain from Corollary 2.5

$$\begin{aligned} F(x) - F(\theta) &= \int_0^1 F'(tx)(x)\, dt \\ &= \int_0^1 (F'(tx) - F'(\theta))(x)\, dt + F'(\theta)(x) \ . \end{aligned}$$

With (2.3) this implies

$$F(x) - F(\theta) \geq \int_{1/2}^{1} (F'(tx) - F'(\theta))(x)\, dt + F'(\theta)(x) \ . \tag{2.4}$$

From the Mean-Value Theorem it follows that for some $s = s(x) \in [\frac{1}{2}, 1]$

$$\begin{aligned} F(x) - F(\theta) &\geq \left(\tfrac{1}{2}\right) (F'(sx)(x) + F'(\theta)(x)) \\ &\geq \left(\tfrac{1}{2}\right) \|x\| \left\{ \frac{F'(sx)(sx)}{\|sx\|} - \|F'(\theta)\|_{V'} \right\} \ . \end{aligned}$$

Since $\|sx\| \geq (\frac{1}{2})\|x\|$ for all $x \in K$, the result follows.

Definitions. Let D be a non-empty subset of V and $T : D \to V'$ be a function. Then T is *monotone* if

$$(T(x) - T(y))(x - y) \geq 0 \ , \qquad x, y \in D \ ,$$

and *strictly monotone* if equality holds only when $x = y$. We call T *coercive* if

$$\lim_{\|x\| \to \infty} \left(\frac{T(x)(x)}{\|x\|} \right) = +\infty \ .$$

After the preceding remarks on potential functions, we have the following fundamental results.

Theorem 2.7 *Let K be a non-empty closed, convex subset of the real separable Hilbert space V, and let the function $F : K \to \mathbb{R}$ be G-differentiable on K. Assume the gradient F' is monotone and either* (a) *K is bounded or* (b) *F' is coercive. Then the set $M \equiv \{x \in K : F(x) \leq F(y)$ for all $y \in K\}$ is non-empty, closed and convex, and $x \in M$ if and only if*

$$x \in K : \quad F'(x)(y - x) \geq 0 \ , \qquad y \in K \ . \tag{2.5}$$

Proof: That M is non-empty follows from Theorems 2.2 and 2.3, Corollary 2.4 and Lemma 2.6. Each of the sets $M_y \equiv \{x \in K : F(x) \leq F(y)\}$ is closed and convex so their intersection, M, is closed and convex. If $x \in M$ then (2.5) follows from the definition of $F'(x)$; conversely, (2.2) shows that (2.5) implies $x \in M$.

2.4

We close with a sufficient condition for uniqueness of the minimum point.

Definition. The function $F : K \to \mathbb{R}$ is *strictly convex* if its domain is convex and for $x, y \in K$, $x \neq y$, and $t \in (0,1)$ we have

$$F(tx + (1-t)y) < tF(x) + (1-t)F(y) \ .$$

Theorem 2.8 *A strictly convex function $F : K \to \mathbb{R}$ has at most one point at which the minimum is attained.*

Proof: Suppose $x_1, x_2 \in K$ with $F(x_1) = F(x_2) = \inf\{F(y) : y \in K\}$ and $x_1 \neq x_2$. Since $\frac{1}{2}(x_1 + x_2) \in K$, the strict convexity of F gives

$$F\left((\tfrac{1}{2})(x_1 + x_2)\right) < \left(\tfrac{1}{2}\right)(F(x_1) + F(x_2)) = \inf\{F(y) : y \in K\} \ ,$$

and this is a contradiction.

The third part of the proof of Theorem 2.3 gives the following.

Theorem 2.9 *Let F be G-differentiable on K. If the gradient F' is strictly monotone, then F is strictly convex.*

3 Variational Inequalities

The characterization (2.5) of the minimum point u of F on K is an example of a *variational inequality*. It expresses the fact that from the minimum point the function does not decrease in any direction into the set K. Moreover, if the minimum point is an interior point of K, then we obtain the "variational equality" $F'(u) = 0$, a functional equation for the (gradient) operator F'.

3.1

We shall write out the special form of the preceding results which occur when F is a quadratic function. Thus, V is a real Hilbert space, $f \in V'$, and $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ is continuous, bilinear and symmetric. Define $F : V \to \mathbb{R}$ by

$$F(v) = \left(\tfrac{1}{2}\right) a(v,v) - f(v) \ , \qquad v \in V \ . \tag{3.1}$$

From the symmetry of $a(\cdot,\cdot)$ we find the G-differential of F is given by

$$F'(u)(v-u) = a(u, v-u) - f(v-u) \ , \qquad u, v \in V \ .$$

If $\mathcal{A} : V \to V'$ is the operator characterizing the form $a(\cdot,\cdot)$, cf. Section I.5.4, then we obtain

$$F'(u) = \mathcal{A}u - f \ , \qquad u \in V \ . \tag{3.2}$$

To check the convexity of F by the monotonicity of its gradient, we compute

$$(F'u - F'v)(u-v) = a(u-v, u-v) = \mathcal{A}(u-v)(u-v) \ .$$

Thus, F' is monotone (strictly monotone) exactly when $a(\cdot,\cdot)$ is non-negative (respectively, positive), and this is equivalent to $\mathcal{A}$ being monotone (respectively, positive) (cf. Section V.1). The growth of F is implied by the statement

$$\lim_{\|v\|\to\infty} \left(\frac{a(v,v)}{\|v\|} \right) = +\infty \ . \tag{3.3}$$

Since $F(v) \geq (\frac{1}{2})a(v,v) - \|f\| \cdot \|v\|$, from the identity (3.2) we find that (3.3) is equivalent to F' being coercive.

The preceding remarks show that Theorems 2.7 and 2.8 give the following.

Theorem 3.1 *Let $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ be continuous, bilinear, symmetric and non-negative. Suppose $f \in V'$ and K is a closed convex subset of V. Assume either* (a) *K is bounded or* (b) *$a(\cdot,\cdot)$ is V-coercive. Then there exists a solution of*

$$u \in K : \quad a(u, v-u) \geq f(v-u) \ , \qquad v \in K \ . \tag{3.4}$$

There is exactly one such u in the case of (b)*; there is exactly one in case* (a) *if we further assume $a(\cdot,\cdot)$ is positive.*

Finally we note that when K is the whole space V, then (3.4) is equivalent to

$$u \in V : \quad a(u,v) = f(v) \ , \qquad v \in V \ , \tag{3.5}$$

the generalized boundary value problem studied in Chapter III. For this reason, when (3.5) is equivalent to a boundary value problem, (3.5) is called the *variational form* of that problem and such problems are called *variational boundary value problems*.

We shall illustrate some very simple variational inequalities by examples in which we characterize the solution by other means.

3.2 Projection

Given the Hilbert space V, the closed convex subset K, and the point $u_0 \in V$, we define

$$a(u,v) = (u,v)_V \ , \quad f(v) = (u_0,v)_V \ , \qquad u,v \in V \ .$$

Then (3.1) gives the function

$$F(v) = \left(\tfrac{1}{2}\right)\left\{\|u_0 - v\|^2 - \|u_0\|^2\right\} \ , \qquad v \in V \ ,$$

so $u \in K$ is the minimum of F on K if and only if

$$\|u_0 - u\| \le \|u_0 - v\| \ , \qquad v \in K \ .$$

That is, u is that (unique) point of K which is closest to u_0. The existence and uniqueness follows from Theorem 3.1; in this case we have the equivalent of Theorem I.4.3. The computation

$$F'(u)(v-u) = (u - v_0, v - u)_V$$

shows that u is characterized by the variational inequality

$$u \in K: \quad (u - u_0, v - u)_V \ge 0 \ , \qquad v \in K \ ,$$

and the geometric meaning of this inequality is that the angle between $u - u_0$ and $v - u$ is between $-\pi/2$ and $\pi/2$ for each $v \in K$. If K is a subspace of V, this is equivalent to (3.5) which says $u - u_0$ is orthogonal to K. That is, u is the projection of u_0 on the space K, cf. Section I.4.3.

3.3 Dirichlet's Principle

Let G be a bounded open set in $\mathbb{R}^n$ and $V = H_0^1(G)$. Let $F \in L^2(\Omega)$ and define

$$a(u,v) = \int_G \nabla u \cdot \nabla v \, dx \ , \quad f(v) = \int_G F(x)v(x)\, dx \ , \qquad u,v \in V \ .$$

Thus, the function to be minimized is

$$E(v) = \left(\tfrac{1}{2}\right)\int_G \sum_{j=1}^n |\partial_j v|^2 \, dx - \int_G Fv \, dx \ , \qquad v \in V \ .$$

In the applications this is a measure of the "energy" in the system. Take K to be the whole space: $K = V$. The point u at which E attains its minimum is characterized by (3.5). Thus, the solution is characterized by the Dirichlet problem (1.1), cf. Chapter III.

3.4 Free Boundary Problem

We take the space V, the form $a(\cdot,\cdot)$ and functional f as above. Let $g \in H_0^1(\Omega)$ and define

$$K = \{v \in H_0^1(G) : v(x) \geq g(x) \quad \text{a.e.} \quad x \in G\} \ .$$

Since $a(\cdot,\cdot)$ is V-coercive, there exists a unique solution u of (3.4). This solution is characterized by the following:

$$\left.\begin{aligned} &u \geq g \text{ in } G \ , \quad u = 0 \text{ on } \partial G \ , \\ &-\Delta_n u - F \geq 0 \text{ in } G \ , \quad \text{and} \\ &(u-g)(-\Delta_n u - F) = 0 \text{ in } G \ . \end{aligned}\right\} \tag{3.6}$$

The first follows from $u \in K$ and the second is obtained from (3.4) by setting $v = u + \varphi$ for any $\varphi \in C_0^\infty(G)$ with $\varphi \geq 0$. Given the first two lines of (3.6), the third line follows by setting $v = g$ in (3.4). One can show, conversely, that any $u \in H^1(G)$ satisfying (3.6) is the solution of (3.4). Note that the region G is partitioned into two parts

$$G_0 = \{x : u(x) = g(x)\} \quad , \quad G_+ = \{x : u(x) > g(x)\}$$

and $-\Delta_n u = F$ in G_+. That is, in G_0 (G_+) the first (respectively, second) inequality in (3.6) is replaced by the corresponding equation. There is a *free boundary* at the interface between G_0 and G_+; locating this free boundary is equivalent to reducing (3.6) to a Dirichlet problem.

3.5 Unilateral Boundary Condition

Choose $V = H^1(G)$ and $K = \{v \in V : v \geq g_1 \text{ on } \partial G\}$, where $g_1 \in H^1(G)$ is given. Let $F(\cdot) \in L^2(G)$, $g_2 \in L^2(\partial G)$ and define $f \in V'$ by

$$f(v) = \int_G Fv \, dx + \int_{\partial G} g_2 v \, ds \ , \qquad v \in V$$

where we suppress the trace operator in the above and hereafter. Set $a(u,v) = (u,v)_{H^1(G)}$. Theorem 3.1 shows there exists a unique solution

$u \in K$ of (3.4). This solution is characterized by the following:

$$\left.\begin{aligned} &-\Delta_n u + u = F \text{ in } G \ , \\ &u \geq g_1 \text{ on } \partial G \ , \\ &\frac{\partial u}{\partial \nu} \geq g_2 \text{ on } \partial G \ , \quad \text{and} \\ &\left(\frac{\partial u}{\partial \nu} - g_2\right)(u - g_1) = 0 \text{ on } \partial G \ . \end{aligned}\right\} \tag{3.7}$$

We shall show that the solution of (3.4) satisfies (3.7); the converse is left to an exercise. The first inequality in (3.7) follows from $u \in K$. If $\varphi \in C_0^\infty(G)$, then setting $v = u + \varphi$, then $v = u - \varphi$ in (3.4) we obtain the partial differential equation in (3.7). Inserting this equation in (3.4) and using the abstract Green's formula (Theorem III.2.3), we obtain

$$\int_{\partial G} \frac{\partial u}{\partial \nu}(v - u)\, ds \geq \int_{\partial G} g_2(v - u) \ , \qquad v \in K \ . \tag{3.8}$$

If $w \in H^1(G)$ satisfies $w \geq 0$ on ∂G, we may set $v = u + w$ in (3.8); this gives the second inequality in (3.7). Setting $v = g_1$ in (3.8) yields the last equation in (3.7). Note that there is a region Γ_0 in ∂G on which $u = g_1$, and $\partial u/\partial \nu = g_2$ on $\partial G \sim \Gamma_0$. Thus, finding u is equivalent to finding Γ_0, so we may think of (3.7) as another free boundary problem.

4 Optimal Control of Boundary Value Problems

4.1

Various optimal control problems are naturally formulated as minimization problems like those of Section 2. We illustrate the situation with a model problem which we discuss in this section.

Example. Let G be a bounded open set in $\mathbb{R}^n$ whose boundary ∂G is a C^1-manifold with G on one side. Let $F \in L^2(G)$ and $g \in L^2(\partial G)$ be given. Then for each *control* $v \in L^2(\partial G)$ there is a corresponding *state* $y \in H^1(G)$ obtained as the unique solution of the *system*

$$\left.\begin{aligned} &-\Delta_n y + y = F && \text{in } G \\ &\frac{\partial y}{\partial \nu} = g + v && \text{on } \partial G \end{aligned}\right\} \tag{4.1}$$

and we denote the dependence of y on v by $y = y(v)$. Assume that we may observe the state y only on ∂G and that our objective is to choose v so as to place the *observation* $y(v)|_{\partial G}$ closest to a given desired observation $w \in L^2(\partial G)$. Each control v is exerted at some *cost*, so the optimal control problem is to minimize the "error plus cost"

$$J(v) = \int_{\partial G} |y(v) - w|^2 \, dx + c \int_{\partial G} |v|^2 \, dx \tag{4.2}$$

over some given set of *admissible controls* in $L^2(\partial G)$. An admissible control u at which J attains its minimum is called an *optimal control*. We shall briefly consider problems of existence or uniqueness of optimal controls and alternate characterizations of them, and then apply these general results to our model problem.

We shall formulate the model problem (4.1), (4.2) in an abstract setting suggested by Chapter III. Thus, let V and H be real Hilbert spaces with V dense and continuously imbedded in H; identify the pivot space H with its dual and thereby obtain the inclusions $V \hookrightarrow H \hookrightarrow V'$. Let $a(\cdot, \cdot)$ be a continuous, bilinear and coercive form on V for which the corresponding operator $\mathcal{A} : V \to V'$ given by

$$a(u, v) = \mathcal{A}u(v) \ , \qquad u, v \in V$$

is necessarily a continuous bijection with continuous inverse. Finally, let $f \in V'$ be given. (The system (4.1) with $v \equiv 0$ can be obtained as the operator equation $\mathcal{A}y = f$ for appropriate choices of the preceding data; cf. Section III.4.2 and below.)

To obtain a control problem we specify in addition to the state space V and data space V' a Hilbert space U of controls and an operator $\mathcal{B} \in \mathcal{L}(U, V')$. Then for each control $v \in U$, the corresponding state $y = y(v)$ is the solution of the system (cf. (4.1))

$$\mathcal{A}y = f + \mathcal{B}v \ , \qquad y = y(v) \ . \tag{4.3}$$

We are given a Hilbert space W of observations and an operator $\mathcal{C} \in \mathcal{L}(V, W)$. For each state $y \in V$ there is a corresponding observation $\mathcal{C}y \in W$ which we want to force close to a given desired observation $w \in W$. The cost of applying the control $v \in U$ is given by $Nv(v)$ where $N \in \mathcal{L}(U, U')$ is symmetric and monotone. Thus, to each control $v \in U$ there is the "error plus cost" given by

$$J(v) \equiv \|\mathcal{C}y(v) - w\|_W^2 + Nv(v) \ . \tag{4.4}$$

The *optimal control problem* is to minimize (4.4) over a given non-empty closed convex subset $U_{\rm ad}$ of *admissible controls* in U. An *optimal control* is a solution of

$$u \in U_{\rm ad} : J(u) \le J(v) \quad \text{for all} \quad v \in U_{\rm ad} \ . \tag{4.5}$$

4.2

Our objectives are to give sufficient conditions for the existence (and possible uniqueness) of optimal controls and to characterize them in a form which gives more information.

We shall use Theorem 2.7 to attain these objectives. In order to compute the G-differential of J we first obtain from (4.3) the identity

$$\mathcal{C}y(v) - w = \mathcal{C}\mathcal{A}^{-1}\mathcal{B}v + \mathcal{C}\mathcal{A}^{-1}f - w$$

which we use to write (4.4) in the form

$$J(v) = \|\mathcal{C}\mathcal{A}^{-1}\mathcal{B}v\|_W^2 + Nv(v) + 2(\mathcal{C}\mathcal{A}^{-1}\mathcal{B}v, \mathcal{C}\mathcal{A}^{-1}f - w)_W + \|\mathcal{C}\mathcal{A}^{-1}f - w\|_W^2 \ .$$

Having expressed J as the sum of quadratic, linear and constant terms, we easily obtain the G-differential

$$\begin{aligned} J'(v)(\varphi) \ &= \ 2\Big\{(\mathcal{C}\mathcal{A}^{-1}\mathcal{B}v, \mathcal{C}\mathcal{A}^{-1}\mathcal{B}\varphi)_W \\ &\qquad\qquad + Nv(\varphi) + (\mathcal{C}\mathcal{A}^{-1}\mathcal{B}\varphi, \mathcal{C}\mathcal{A}^{-1}f - w)_W\Big\} \\ &= \ 2\Big\{(\mathcal{C}y(v) - w, \mathcal{C}\mathcal{A}^{-1}\mathcal{B}\varphi)_W + Nv(\varphi)\Big\} \ . \end{aligned} \tag{4.6}$$

Thus, we find that the gradient J' is monotone and

$$\left(\tfrac{1}{2}\right) J'(v)(v) \ge Nv(v) - (\text{const.})\|v\|_U \ ,$$

so J' is coercive if N is coercive, i.e., if

$$Nv(v) \ge c\|v\|_U^2 \ , \qquad v \in U_{\rm ad} \ , \tag{4.7}$$

for some $c > 0$. Thus, we obtain from Theorem 2.7 the following.

Theorem 4.1 *Let the optimal control problem be given as in Section 4.1. That is, we are to minimize* (4.4) *subject to* (4.3) *over the non-empty closed convex set* $U_{\rm ad}$. *Then if either* (a) $U_{\rm ad}$ *is bounded or* (b) N *is coercive over* $U_{\rm ad}$, *then the set of optimal controls is non-empty, closed and convex.*

Corollary 4.2 *In case* (b) *there is a unique optimal control.*

Proof: This follows from Theorem 2.9 since (4.7) implies J' is strictly monotone.

4.3

We shall characterize the optimal controls by variational inequalities. Thus, u is an optimal control if and only if

$$u \in U_{\rm ad} : J'(u)(v-u) \geq 0 \ , \qquad v \in U_{\rm ad} \ ; \tag{4.8}$$

this is just (2.5). This variational inequality is given by (4.6), of course, but the resulting form is difficult to interpret. The difficulty is that it compares elements of the observation space W with those of the control space U; we shall obtain an equivalent characterization which contains a variational inequality only in the control space U. In order to convert the first term on the right side of (4.6) into a more convenient form, we shall use the Riesz map R_W of W onto W' given by (cf. Section I.4.3)

$$R_W(x)(y) = (x,y)_W \ , \qquad x,y \in W$$

and the dual $\mathcal{C}' \in \mathcal{L}(W',V')$ of $\mathcal{C}$ given by (cf. Section I.5.1)

$$\mathcal{C}'(f)(x) = f(\mathcal{C}(x)) \ , \qquad f \in W' \ , \ x \in V \ .$$

Then from (4.6) we obtain

$$\begin{aligned} \left(\tfrac{1}{2}\right) J'(u)(v) &= (\mathcal{C}y(u) - w, \mathcal{C}\mathcal{A}^{-1}\mathcal{B}v)_W + Nu(v) \\ &= R_W(\mathcal{C}y(u) - w)(\mathcal{C}\mathcal{A}^{-1}\mathcal{B}v) + Nu(v) \\ &= \mathcal{C}' R_W(\mathcal{C}y(u) - w)(\mathcal{A}^{-1}\mathcal{B}v) + Nu(v) \ , \qquad u,v \in U \ . \end{aligned}$$

To continue we shall need the dual operator $\mathcal{A}' \in \mathcal{L}(V,V')$ given by

$$\mathcal{A}'x(y) = \mathcal{A}y(x) \ , \qquad x,y \in V \ ,$$

where V'' is naturally identified with V. Since $\mathcal{A}'$ arises from the bilinear form adjoint to $a(\cdot,\cdot)$, $\mathcal{A}'$ is an isomorphism. Thus, for each control $v \in U$ we

can define the corresponding *adjoint state* $p = p(v)$ as the unique solution of the system

$$\mathcal{A}'p = \mathcal{C}'R_W(\mathcal{C}y(v) - w) \ , \qquad p = p(v) \ . \tag{4.9}$$

From above we then have

$$\begin{aligned} \left(\tfrac{1}{2}\right) J'(u)(v) &= \mathcal{A}'p(u)(\mathcal{A}^{-1}\mathcal{B}v) + Nu(v) \\ &= \mathcal{B}v(p(u)) + Nu(v) \\ &= \mathcal{B}'p(u)(v) + Nu(v) \end{aligned}$$

where $\mathcal{B}' \in \mathcal{L}(V, U')$ is the indicated dual operator. These computations lead to a formulation of (4.8) which we summarize as follows.

Theorem 4.3 *Let the optimal control problem be given as in* (4.1)*. Then a necessary and sufficient condition for u to be an optimal control is that it satisfy the following system:*

$$\left.\begin{aligned} &u \in U_{\text{ad}} \ , \quad \mathcal{A}y(u) = f + \mathcal{B}u \ , \\ &\mathcal{A}'p(u) = \mathcal{C}'R_W(\mathcal{C}y(u) - w) \ , \\ &(\mathcal{B}'p(u) + Nu)(v - u) \geq 0 \ , \quad \textit{all} \ \ v \in U_{\text{ad}} \ . \end{aligned}\right\} \tag{4.10}$$

The system (4.10) is called the *optimality system* for the control problem. We leave it as an exercise to show that a solution of the optimality system satisfies (4.8).

4.4

We shall recover the Example of Section 4.1 from the abstract situation above. Thus, we choose $V = H^1(G)$, $a(u, v) = (u, v)_{H^1(G)}$, $U = L^2(\partial G)$ and define

$$\begin{aligned} f(v) &= \int_G F(x)v(x)\,dx + \int_{\partial G} g(s)v(s)\,ds \ , \qquad && v \in V \ , \\ \mathcal{B}u(v) &= \int_{\partial G} u(s)v(s)\,ds \ , && u \in U \ , \ v \in V \ . \end{aligned}$$

The state $y(u)$ of the system determined by the control u is given by (4.3), i.e.,

$$\left.\begin{aligned} &-\Delta_n y + y = F \ \ \text{in} \ \ G \ , \\ &\frac{\partial y}{\partial \nu} = g + u \ \ \text{on} \ \ \partial G \ . \end{aligned}\right\} \tag{4.11}$$

Choose $W = L^2(\partial G)$, $w \in W$, and define

$$Nu(v) = c\int_{\partial G} u(s)v(s)\,ds\ , \qquad u, v \in W\ ,\ (c \geq 0)$$
$$\mathcal{C}u(v) \equiv \int_{\partial G} u(s)v(s)\,ds\ , \qquad u \in V\ ,\ v \in W\ .$$

The adjoint state equation (4.9) becomes

$$\begin{aligned} -\Delta_n p + p &= 0 \ \text{ in } \ G \\ \frac{\partial p}{\partial \nu} &= y - w \ \text{ on } \ \partial G \end{aligned} \tag{4.12}$$

and the variational inequality is given by

$$u \in U_{\text{ad}} : \int_{\partial G} (p + cu)(v - u)\,ds \geq 0\ , \qquad v \in U_{\text{ad}}\ . \tag{4.13}$$

From Theorem 4.1 we obtain the existence of an optimal control if U_{ad} is bounded or if $c > 0$. Note that

$$J(v) = \int_{\partial G} |y(v) - w|^2\,ds + c\int_{\partial G} |v|^2\,ds \tag{4.14}$$

is strictly convex in either case, so uniqueness follows in both situations. Theorem 4.3 shows the unique optimal control u is characterized by the optimality system (4.11), (4.12), (4.13). We illustrate the use of this system in two cases.

4.5 $U_{\text{ad}} = L^2(\partial G)$

This is the case of *no constraints* on the control. Existence of an optimal control follows if $c > 0$. Then (4.13) is equivalent to $p + cu = 0$. The optimality system is equivalent to

$$\begin{aligned} -\Delta_n y + y &= F\ , & -\Delta_n p + p &= 0 \ \text{ in } G \\ \frac{\partial y}{\partial \nu} &= g - \left(\frac{1}{c}\right) p\ , & \frac{\partial p}{\partial \nu} &= y - w \ \text{ on } \ \partial G \end{aligned}$$

and the optimal control is given by $u = -(1/c)p$.

Consider the preceding case with $c = 0$. We show that an optimal control might not exist. First show $\inf\{J(v) : v \in U\} = 0$. Pick a sequence $\{w_m\}$ of

very smooth functions on ∂G such that $w_m \to w$ in $L^2(\partial G)$. Define y_m by

$$\begin{aligned} -\Delta_n y_m + y_m &= F \ \text{ in } \ G \\ y_m &= w_m \ \text{ on } \ \partial G \end{aligned}$$

and set $v_m = (\partial y_m/\partial \nu) - g$, $m \geq 1$. Then $v_m \in L^2(\partial G)$ and $J(v_m) = \|w_m - w\|^2_{L^2(\partial G)} \to 0$. Second, note that if u is an optimal control, then $J(u) = 0$ and the corresponding state y satisfies

$$\begin{aligned} -\Delta_n y + y &= F \ \text{ in } \ G \\ y &= w \ \text{ on } \ \partial G \ . \end{aligned}$$

Then we have (formally) $u = (\partial y/\partial \nu) - g$. However, if $w \in L^2(\partial G)$ one does not in general have $(\partial y/\partial \nu) \in L^2(\partial G)$. Thus u might not be in $L^2(\partial G)$ in which case there is no optimal control (in $L^2(\partial G)$).

4.6

$U_{\text{ad}} = \{v \in L^2(\partial G) : 0 \leq v(s) \leq M \text{ a.e.}\}$. Since the set of admissible controls is bounded, there exists a unique optimal control u characterized by the optimality system (4.10). Thus, u is characterized by (4.11), (4.12) and

$$\begin{aligned} &\text{if } \ 0 < u < M \ , \ \text{ then } \ p + cu = 0 \\ &\text{if } \ u = 0 \ , \ \text{ then } \ p \geq 0 \ , \text{ and} \\ &\text{if } \ u = M \ , \ \text{ then } \ p + cu \leq 0 \ . \end{aligned} \tag{4.15}$$

We need only to check that (4.13) and (4.15) are equivalent. The boundary is partitioned into the three regions determined by the three respective cases in (4.15). This is analogous to the free boundary problems encountered in Sections 3.3 and 3.4.

We specialize the above to the case of "free control," i.e., $c = 0$. One may search for an optimal control in the following manner. Motivated by (4.11) and (4.14), we consider the solution Y of the Dirichlet problem

$$\begin{aligned} -\Delta_n Y + Y &= F \ \text{ in } \ G \ , \\ Y &= w \ \text{ on } \ \partial G \ . \end{aligned}$$

If it happens that

$$0 \leq \frac{\partial Y}{\partial \nu} - g \leq M \quad \text{on} \quad \partial G \ , \tag{4.16}$$

then the optimal control is given by (4.11) as

$$u = \frac{\partial Y}{\partial \nu} - g \ .$$

Note that $u \in U_{\text{ad}}$ and $J(u) = 0$.

We consider the contrary situation in which (4.16) does not hold. Specifically we shall show that (when all aspects of the problem are regular) the set $\Gamma \equiv \{s \in \partial G : 0 < u(s) < M,\ p(s) = 0\}$ is empty. This implies that the control takes on only its extreme values $0, M$; this is a result of "bang-bang" type.

Partition Γ into the three parts $\Gamma_0 = \{s \in \Gamma : y(s) = w(s)\}$, $\Gamma_+ = \{s \in \Gamma : y(s) > w(s)\}$ and $\Gamma_- = \{s \in \Gamma : y(s) < w(s)\}$. On any interval in Γ_0 we have $p = 0$ (by definition of Γ) and $\frac{\partial p}{\partial \nu} = 0$ from (4.12). From the uniqueness of the Cauchy problem for the elliptic equation in (4.12), we obtain $p = 0$ in G, hence, $y = w$ on ∂G. But this implies $y = Y$, hence (4.16) holds. This contradiction shows Γ_0 is empty. On any interval in Γ_+ we have $p = 0$ and $\frac{\partial p}{\partial \nu} > 0$. Thus, $p < 0$ in some neighborhood (in $\bar{G}$) of that interval. But $\Delta p < 0$ in the neighborhood follows from (4.12), so a maximum principle implies $\frac{\partial p}{\partial \nu} \leq 0$ on that interval. This contradiction shows Γ_+ is empty. A similar argument holds for Γ_- and the desired result follows.

5 Approximation of Elliptic Problems

We shall discuss the *Rayleigh-Ritz-Galerkin* procedure for approximating the solution of an elliptic boundary value problem. This procedure can be motivated by the situation of Section 3.1 where the abstract boundary value problem (3.5) is known to be equivalent to minimizing a quadratic function (3.1) over the Hilbert space V. The procedure is to pick a closed subspace S of V and minimize the quadratic function over S. This is the Rayleigh-Ritz method. The resulting solution is close to the original solution if S closely approximates V. The approximate solution is characterized by the abstract boundary vlaue problem obtained by replacing V with S; this gives the (equivalent) Galerkin method of obtaining an approximate solution. The very important *finite-element method* consists of the above

procedure applied with a space S of piecewise polynomial functions which approximates the whole space V. The resulting finite-dimensional problem can be solved efficiently by computers. Our objectives are to describe the Rayleigh-Ritz-Galerkin procedure, obtain estimates on the error that results from the approximation, and then to give some typical convergence rates that result from standard finite-element or *spline* approximations of the space. We shall also construct some of these approximating subspaces and prove the error estimates as an application of the minimization theory of Section 2.

5.1

Suppose we are given an abstract boundary value problem: V is a Hilbert space, $a(\cdot,\cdot) : V \times V \to \mathbb{K}$ is continuous and sesquilinear, and $f \in V'$. The problem is to find u satisfying

$$u \in V : a(u,v) = f(v) \ , \qquad v \in V \ . \tag{5.1}$$

Let S be a subspace of V. Then we may consider the related problem of determining u_s satisfying

$$u_s \in S : a(u_s,v) = f(v) \ , \qquad v \in S \ . \tag{5.2}$$

We shall show that the error $u_s - u$ is small if S approximates V sufficiently well.

Theorem 5.1 *Let $a(\cdot,\cdot)$ be a V-coercive continuous sesquilinear form and $f \in V'$. Let S be a closed subspace of V. Then* (5.1) *has a unique solution u and* (5.2) *has a unique solution u_s. Furthermore we have the estimate*

$$\|u_s - u\| \le (K/c) \inf\{\|u - v\| : v \in S\} \ , \tag{5.3}$$

where K is the bound on $a(\cdot,\cdot)$ (cf. the inequality I.(5.2)*) and c is the coercivity constant (cf. the inequality* III.(2.3)*).*

Proof: The existence and uniqueness of the solutions u and u_s of (5.1) and (5.2) follow immediately from Theorem III.2.1 or Theorem 3.1, so we need only to verify the estimate (5.3). By subtracting (5.1) from (5.2) we obtain

$$a(u_s - u, v) = 0 \ , \qquad v \in S \ . \tag{5.4}$$

Thus for any $w \in S$ we have

$$a(u_s - u, u_s - u) = a(u_s - u, w - u) + a(u_s - u, u_s - w) \ .$$

Since $u_s - w \equiv v \in S$ it follows that the last term is zero, so we obtain

$$c\|u_s - u\|^2 \leq K\|u_s - u\| \, \|w - u\| \ , \qquad w \in S \ .$$

This gives the desired result.

Consider for the moment the case of V being separable. Thus, there is a sequence $\{v_1, v_2, v_3, \ldots\}$ in V which is a basis for V. For each integer $m \geq 1$, the set $\{v_1, v_2, \ldots, v_m\}$ is linearly independent and its linear span will be denoted by V_m. If P_m is the projection of V into V_m, then $\lim_{m\to\infty} P_m v = v$ for all $v \in V$. The problem (5.2) with $S = V_m$ is equivalent to

$$u_m \in V_m : a(u_m, v_j) = f(v_j) \ , \qquad 1 \leq j \leq m \ .$$

There is exactly one such u_m for each integer $m \geq 1$ and we have the estimates $\|u_m - u\| \leq (K/c)\|u - P_m u\|$. Hence, $\lim_{m\to\infty} u_m = u$ in V and the rate of convergence is determined by that of $\{P_m u\}$ to the solution u of (5.1). Thus we are led to consider an approximating finite-dimensional problem. Specifically u_m is determined by the point $x = (x_1, x_2, \ldots, x_m) \in \mathbb{K}^m$ through the identity $u_m = \sum_{i=1}^m x_i v_i$, and (5.2) is equivalent to the $m \times m$ system of linear equations

$$\sum_{i=1}^{m} a(v_i, v_j) x_i = f(v_j) \ , \qquad 1 \leq j \leq m \ . \tag{5.5}$$

Since $a(\cdot, \cdot)$ is V-coercive, the $m \times m$ coefficient matrix $(a(v_i, v_j))$ is invertible and the linear system (5.5) can be solved for x. The dimension of the system is frequently of the order $m = 10^2$ or 10^3, so the actual computation of the solution may be a non-trivial consideration. It is helpful to choose the basis functions so that most of the coefficients are zero. Thus, the matrix is *sparse* and various special techniques are available for efficiently solving the large linear system. This sparseness of the coefficient matrix is one of the computational advantages of using finite-element spaces. A very special example will be given in Section 5.4 below.

5.2

The fundamental estimate (5.3) is a bound on the error in the norm of the Hilbert space V. In applications to elliptic boundary value problems this corresponds to an *energy estimate*. We shall estimate the error in the norm of a pivot space H. Since this norm is weaker we expect an improvement on the rate of convergence with respect to the approximation of V by S.

Theorem 5.2 *Let $a(\cdot,\cdot)$ be a continuous, sesquilinear and coercive form on the Hilbert space V, and let H be a Hilbert space identified with its dual and in which V is dense, and continuously imbedded. Thus, $V \hookrightarrow H \hookrightarrow V'$. Let $A^* : D^* \to H$ be the operator on H which is determined by the adjoint sesquilinear form, i.e.,*

$$\overline{a(v,w)} = (A^*w, v)_H \ , \qquad w \in D^* \ , \ v \in V$$

(cf. Section III.7.5). Let S be a closed subspace of V and $e^*(S)$ a corresponding constant for which we have

$$\inf\{\|w - v\| : v \in S\} \leq e^*(S)|A^*w|_H \ , \qquad w \in D^* \ . \tag{5.6}$$

Then the solutions u of (5.1) and u_s of (5.2) satisfy the estimate

$$|u - u_s|_H \leq (K^2/c)\inf\{\|u - v\| : v \in S\}e^*(S) \ . \tag{5.7}$$

Proof: We may assume $u \neq u_s$; define $g = (u - u_s)/|u - u_s|_H$ and choose $w \in D^*$ so that $A^*w = g$. That is,

$$a(v,w) = (v,g)_H \ , \qquad v \in V \ ,$$

and this implies that

$$a(u - u_s, w) = (u - u_s, g)_H = |u - u_s|_H \ .$$

From this identity and (5.4) we obtain for any $v \in S$

$$|u-u_s|_H = a(u-u_s, w-v) \leq K\|u-u_s\| \, \|w-v\| \leq K\|u-u_s\|e^*(S)|A^*w|_H \ .$$

Since $|A^*w|_H = |g|_H = 1$, the estimate (5.7) follows from (5.3).

Corollary 5.3 *Let $A : D \to H$ be the operator on H determined by $a(\cdot,\cdot)$, V, H, i.e.,*

$$a(w,v) = (Aw,v)_H \ , \qquad w \in D \ , \ v \in V \ .$$

Let $e(S)$ be a constant for which

$$\inf\{\|w - v\| : v \in S\} \le e(S)|Aw|_H \ , \qquad w \in D \ .$$

Then the solutions of (5.1) *and* (5.2) *satisfy the estimate*

$$|u - u_s|_H \le (K^2/c)e(S)e^*(S)|Au|_H \ . \tag{5.8}$$

The estimate (5.7) will provide the rate of convergence of the error that is faster than that of (5.3). The added factor $e^*(S)$ arising in (5.6) will depend on how well S approximates the subspace D^* of "smoother" or "more regular" elements of V.

5.3

We shall combine the estimates (5.3) and (5.7) with approximation results that are typical of finite-element or spline function subspaces of $H^1(G)$. This will result in rate of convergence estimates in terms of a parameter $h > 0$ related to mesh size in the approximation scheme. The *approximation assumption* that we make is as follows: Suppose $\mathcal{H}$ is a set of positive numbers, M and $k \ge 0$ are integers, and $\mathcal{S} \equiv \{S_h : h \in \mathcal{H}\}$ is a collection of closed subspaces of $V \subset H^1(G)$ such that

$$\inf\{\|w - v\|_{H^1(G)} : v \in S_h\} \le Mh^{j-1}\|w\|_{H^j(G)} \tag{5.9}$$

for all $h \in \mathcal{H}$, $1 \le j \le k+2$, and $w \in H^j(G) \cap V$. The integer $k+1$ is called the *degree* of $\mathcal{S}$.

Theorem 5.4 *Let V be a closed subspace of $H^1(G)$ with $H_0^1(G) \subset V$ and let $a(\cdot,\cdot) : V \times V \to \mathbb{K}$ be continuous, sesquilinear and V-coercive. Let $\mathcal{S}$ be a collection of closed subspaces of V satisfying* (5.9) *for some $k \ge 0$, and assume $a(\cdot,\cdot)$ is k-regular on V. Let $F \in H^k(G)$ and define $f \in V'$ by $f(v) = (F,v)_H$, $v \in V$, where $H \equiv L^2(G)$. Let u be the solution of* (5.1) *and, for each $h \in \mathcal{H}$, u_h be the solution of* (5.2) *with $S = S_h$. Then for some constant c_1 we have*

$$\|u - u_h\|_{H^1(G)} \le c_1 h^{k+1} \ , \qquad h \in \mathcal{H} \ . \tag{5.10}$$

If in addition the sesquilinear form adjoint to $a(\cdot,\cdot)$ is 0-regular, then for some constant c_2 we have

$$\|u - u_h\|_{L^2(G)} \le c_2 h^{k+2} , \qquad h \in \mathcal{H} . \tag{5.11}$$

Proof: Since $F \in H^k(G)$ and $a(\cdot,\cdot)$ is k-regular it follows that $u \in H^{k+2}(G)$. Hence we combine (5.3) with (5.9) to obtain (5.10). If the adjoint form is 0-regular, then in Theorem 5.2 we have $D^* \subset H^2(G)$ and $\|w\|_{H^2(G)} \le$ (const.)$\|A^*w\|_{L^2(G)}$. Hence (5.9) with $j = 2$ gives (5.6) with $e^*(S_h) =$ (const.)h. Thus (5.11) follows from (5.7).

Sufficient conditions for $a(\cdot,\cdot)$ to be k-regular were given in Section III.6. Note that this permits all the hypotheses in Theorem 5.4 to be placed on the *data* in the problem (5.1) being solved. For problems for which appropriate regularity results are not available, one may of course assume the appropriate smoothness of the solution.

5.4

Let G be the interval $(0,1)$ and V a closed subspace of $H^1(G)$. Any function $f \in V$ can be approximated by a piecewise-linear $f_0 \in V$; we need only to choose f_0 so that it agrees with f at the endpoints of the intervals on which f_0 is affine. This is a simple *Lagrange interpolation* of f by the linear finite-element function f_0, and it leads to a family of approximating subspaces of degree 1. We shall describe the spaces and prove the estimates (5.9) for this example.

Let $P = \{0 = x_0 < x_1 < \cdots < x_N < x_{N+1} = 1\}$ be a partition of G and denote by $\mu(P)$ the mesh of P : $\mu(P) = \max\{x_{j+1} - x_j : 0 \le j \le N\}$. The closed convex set $K = \{v \in V : v(x_j) = 0,\ 0 \le j \le N+1\}$ is basic to our construction. Let $f \in V$ be given and consider the function $F(v) = (\frac{1}{2})|\partial(v-f)|_H^2$ on V, where $H = L^2(G)$. The G-differential is given by

$$F'(u)(v) = (\partial(u-f), \partial v)_H , \qquad u, v \in V .$$

We easily check that F' is strictly monotone on K; this follows from Theorem II.2.4. Similarly the estimate

$$F'(v)(v) = |\partial v|_H^2 - (\partial f, \partial v)_H \ge |\partial v|_H^2 - |\partial f|_H |\partial v|_H , \qquad v \in V ,$$

shows F' is coercive on K. It follows from Theorems 2.7 and 2.9 that there is a unique $u_f \in K$ at which F takes its minimal value on K, and it is

characterized in (2.5) by

$$u_f \in K: \quad (\partial(u_f - f), \partial v)_H = 0 \ , \qquad v \in K \ .$$

This shows that for each $f \in V$, there exists exactly one $f_0 \in V$ which satisfies

$$f_0 - f \in K \ , \quad (\partial f_0, \partial v)_H = 0 \ , \qquad v \in K \ . \tag{5.12}$$

(They are clearly related by $f_0 = f - u_f$.) The second part of (5.12) states that $-\partial^2 f_0 = 0$ in each subinterval of the partition so f_0 is affine on each subinterval. The first part of (5.12) determines the value of f_0 at each of the points of the partition, so it follows that f_0 is that function in V which is affine in the intervals of P and equals f at the points of P. This f_0 is the linear finite-element interpolant of f.

To compute the error in this interpolation procedure, we first note that

$$|\partial f_0|_H^2 + |\partial(f_0 - f)|_H^2 = |\partial f|_H^2$$

follows from setting $v = f_0 - f$ in (5.12). Thus we obtain the estimate

$$|\partial(f_0 - f)|_H \le |\partial f|_H \ .$$

If $g = f_0 - f$, then from Theorem II.2.4 we have

$$\int_{x_j}^{x_{j+1}} |g|^2 \, dx \le 4\mu(P)^2 \int_{x_j}^{x_{j+1}} |\partial g|^2 \, dx \ , \qquad 0 \le j \le N \ ,$$

and summing these up gives

$$|f - f_0|_H \le 2\mu(P)|\partial(f_0 - f)|_H \ . \tag{5.13}$$

This proves the first two estimates in the following.

Theorem 5.5 *For each $f \in V$ and partition P as above, the linear finite-element interpolant f_0 of f with respect to P is characterized by* (5.12) *and it satisfies*

$$|\partial(f_0 - f)|_H \le |\partial f|_H \ , \tag{5.14}$$

and

$$|f_0 - f|_H \le 2\mu(P)|\partial f|_H \ . \tag{5.15}$$

If also $f \in H^2(G)$, then we have

$$|\partial(f_0 - f)|_H \le 2\mu(P)|\partial^2 f|_H \tag{5.16}$$

$$|f_0 - f|_H \le 4\mu(P)^2|\partial^2 f|_H \ . \tag{5.17}$$

Proof: We need only to verify (5.16) and (5.17). Since $(f - f_0)(x_j) = 0$ for $0 \le j \le N+1$, we obtain for each $f \in H^2(G) \cap V$

$$|\partial(f_0 - f)|_H^2 = \sum_{j=0}^{N} \int_{x_j}^{x_{j+1}} (-\partial^2(f_0 - f))(f_0 - f)\, dx = (\partial^2 f, f_0 - f)_H \ ,$$

and thereby the estimate

$$|\partial(f_0 - f)|_H^2 \le |f_0 - f|_H |\partial^2 f|_H \ .$$

With (5.13) this gives (5.16) after dividing the factor $|\partial(f_0 - f)|_H$. Finally, (5.17) follows from (5.13) and (5.16).

Corollary 5.6 *For each h with $0 < h < 1$ let P_h be a partition of G with mesh $\mu(P_h) < h$, and define L_h to be the space of all linear finite-element function in $H^1(G)$ corresponding to the partition P_h. Then $\mathcal{L} \equiv \{L_h : 0 < h < 1\}$ satisfies the approximation assumption* (5.9) *with $k = 0$. The degree of $\mathcal{L}$ is* 1.

Finally we briefly consider the computations that are involved in implementing the Galerkin procedure (5.2) for one of the spaces L_h above. Let $P_h = \{x_0, x_1, \ldots, x_{N+1}\}$ be the corresponding partition and define ℓ_j to be the unique function in L_h which satisfies

$$\ell_j(x_i) = \begin{cases} 1 & \text{if } i = j \ , \\ 0 & \text{if } i \ne j \ , \end{cases} \qquad 0 \le i, j \le N+1 \ . \tag{5.18}$$

For each $f \in H^1(G)$, the interpolant f_0 is given by

$$f_0 = \sum_{j=0}^{N+1} f(x_j)\ell_j \ .$$

We use the basis (5.18) to write the problem in the form (5.5), and we must then invert the matrix $(a(\ell_i, \ell_j))$. Since $a(\cdot, \cdot)$ consists of integrals over G of products of ℓ_i and ℓ_j and their derivatives, and since any such product is identically zero when $|i - j| \ge 2$, it follows that the coefficient matrix is tridiagonal. It is also symmetric and positive-definite. There are efficient methods for inverting such matrices.

6 Approximation of Evolution Equations

We present here the Faedo-Galerkin procedure for approximating the solution of evolution equations of the types considered in Chapters IV, V and VI. As in the preceding section, the idea is to project a weak form of the problem onto a finite-dimensional subspace. We obtain thereby a system of ordinary differential equations whose solution approximates the solution of the original problem. In the applications to initial-boundary-value problems, this corresponds to a discretization of the space variable by a finite-element or spline approximation. We shall describe these semi-discrete approximation procedures, obtain estimates on the error that results from the approximation, and give the convergence rates that result from standard finite-element or spline approximations in the space variable. This program is carried out for first-order evolution equations and also for second-order evolution equations.

6.1

We first consider some first-order equations of the implicit type discussed in Section V.2. Let $\mathcal{M}$ be the Riesz map of the Hilbert space V_m with scalar-product $(\cdot,\cdot)_m$. Let V be a Hilbert space dense and continuously imbedded in V_m and let $\mathcal{L} \in \mathcal{L}(V, V')$. For a given $f \in C((0,\infty), V_m')$ and $u_0 \in V_m$, we consider the problem of approximating a solution $u \in C([0,\infty), V_m) \cap C^1((0,\infty), V_m)$ of

$$\mathcal{M}u'(t) + \mathcal{L}u(t) = f(t) \ , \qquad t > 0 \ , \tag{6.1}$$

with $u(0) = u_0$. Since $\mathcal{M}$ is symmetric, such a solution satisfies

$$D_t(u(t), u(t))_m + 2\ell(u(t), u(t)) = 2f(t)(u(t)) \ , \qquad t > 0 \ , \tag{6.2}$$

where $\ell(\cdot,\cdot) : V \times V \to \mathbb{R}$ is the bilinear form associated with $\mathcal{L}$. This gives the identity

$$\|u(t)\|_m^2 + 2\int_0^t \ell(u(s), u(s))\, ds = \|u_0\|_m^2 + 2\int_0^t f(s)(u(s))\, ds \ , \qquad t > 0 \ , \tag{6.3}$$

involving the V_m norm $\|\cdot\|_m$ of the solution. Since the right side of (6.2) is bounded by $T\|f\|_{V_m'}^2 + T^{-1}\|u\|_m^2$ for any given $T > 0$, we obtain from (6.2)

$$D_t(e^{-t/T}\|u(t)\|_m^2) + e^{-t/T} 2\ell(u(t), u(t)) \le Te^{-t/T}\|f(t)\|_{V_m'}^2$$

and from this follows the a-priori estimate

$$\|u(t)\|_m^2 + 2\int_0^t \ell(u(s),u(s))\,ds \le e\|u_0\|^2 + Te\int_0^t \|f(s)\|_{V_m'}^2\,ds\,, \qquad 0 \le t \le T\,. \tag{6.4}$$

In the situations we consider below, $\mathcal{L}$ is monotone, hence, (6.4) gives an upper bound on the V_m norm of the solution.

In order to motivate the Faedo-Galerkin approximation, we note that a solution u of (6.1) satisfies

$$(u'(t),v)_m + \ell(u(t),v) = f(t)(v)\,, \qquad v \in V\,,\ t > 0\,. \tag{6.5}$$

Since V is dense in V_m, (6.5) is actually equivalent to (6.1). Let S be a subspace of V. Then we consider the related problem of determining $u_s \in C([0,\infty),S) \cap C^1((0,\infty),S)$ which satisfies

$$(u_s'(t),v)_m + \ell(u_s(t),v) = f(t)(v)\,, \qquad v \in S\,,\ t > 0 \tag{6.6}$$

and an initial condition to be specified.

Consider the case of S being a finite-dimensional subspace of V; let $\{v_1, v_2, \ldots, v_m\}$ be a basis for S. Then the solution of (6.6) is of the form

$$u_s(t) = \sum_{i=1}^m x_i(t) v_i$$

where $x(t) \equiv (x_1(t), x_2(t), \ldots, x_m(t))$ is in $C([0,\infty),\mathbb{R}^m) \cap C^1((0,\infty),\mathbb{R}^m)$, and (6.6) is equivalent to the system of ordinary differential equations

$$\sum_{i=1}^m (v_i,v_j)_m x_i'(t) + \sum_{i=1}^m \ell(v_i,v_j)x_i(t) = f(t)(v_j)\,, \qquad 1 \le j \le m\,. \tag{6.7}$$

The linear system (6.7) has a unique solution $x(t)$ with the initial condition $x(0)$ determined by $u_s(0) = \sum_{i=1}^m x_i(0)v_i$. (Note that the matrix coefficient of $x'(t)$ in (6.7) is symmetric and positive-definite, hence, nonsingular.) As in the preceding section, it is helpful to choose the basis functions so that most of the coefficients in (6.7) are zero. Special efficient computational techniques are then available for the resulting sparse system.

6.2

We now develop estimates on the error, $u(t)-u_s(t)$, in the situation of Theorem V.2.2. This covers the case of parabolic and pseudoparabolic equations. It will be shown that the error in the Faedo-Galerkin procedure for (6.1) is bounded by the error in the corresponding Rayleigh-Ritz-Galerkin procedure for the elliptic problem determined by the operator $\mathcal{L}$. Thus, we consider for each $t > 0$ the *$\mathcal{L}$-projection* of $u(t)$ defined by

$$u_\ell(t) \in S: \quad \ell(u_\ell(t), v) = \ell(u(t), v) \ , \qquad v \in S \ . \tag{6.8}$$

Theorem 6.1 *Let the real Hilbert spaces V and V_m, operators $\mathcal{M}$ and $\mathcal{L}$, and data u_0 and f be given as in Theorem V.2.2, and let S be a closed subspace of V. Then there exists a unique solution u of* (6.1) *with $u(0) = u_0$ and there exists a unique solution u_s of* (6.6) *for any prescribed initial value $u_s(0) \in S$. Assume $u \in C([0,\infty), V)$ and choose $u_s(0) = u_\ell(0)$, the $\mathcal{L}$-projection* (6.8) *of $u(0)$. Then we have the error estimate*

$$\|u(t) - u_s(t)\|_m \le \|u(t) - u_\ell(t)\|_m + \int_0^t \|u'(s) - u_\ell'(s)\|_m \, ds \ , \quad t \ge 0 \ . \tag{6.9}$$

Proof: The existence-uniqueness results are immediate from Theorem V.2.2, so we need only to verify (6.9). Note that $u(0) = u_0$ necessarily belongs to V, so (6.8) defines $u_\ell(0) = u_s(0)$. For any $v \in S$ we obtain from (6.5) and (6.6)

$$(u'(t) - u_s'(t), v)_m + \ell(u(t) - u_s(t), v) = 0 \ ,$$

so (6.8) gives the identity

$$(u'(t) - u_\ell'(t), v)_m = (u_s'(t) - u_\ell'(t), v)_m + \ell(u_s(t) - u_\ell(t), v) \ .$$

Setting $v = u_s(t) - u_\ell(t)$ and noting that $\mathcal{L}$ is monotone, we obtain

$$D_t \|u_s(t) - u_\ell(t)\|_m^2 \le 2\|u'(t) - u_\ell'(t)\|_m \|u_s(t) - u_\ell(t)\|_m \ .$$

The function $t \mapsto \|u_s(t) - u_\ell(t)\|_m$ is absolutely continuous, hence differentiable almost everywhere, and satisfies

$$D_t \|u_s(t) - u_\ell(t)\|_m^2 = 2\|u_s(t) - u_\ell(t)\|_m D_t \|u_s(t) - u_\ell(t)\|_m \ .$$

Let $Z = \{t > 0 : \|u_s(t) - u_\ell(t)\|_m = 0\}$. Clearly, for any $t \notin Z$ we have from above

$$D_t \|u_s(t) - u_\ell(t)\|_m \le \|u'(t) - u_\ell'(t)\|_m \ . \tag{6.10}$$

At an accumulation point of Z, the estimate (6.10) holds, since the left side is zero at such a point. Since Z has at most a countable number of isolated points, this shows that (6.10) holds at almost every $t > 0$. Integrating (6.10) gives the estimate

$$\|u_s(t) - u_\ell(t)\|_m \le \int_0^t \|u'(s) - u'_\ell(s)\|_m \, ds \ , \qquad t \ge 0 \ ,$$

from which (6.9) follows by the triangle inequality.

The fundamental estimate (6.9) shows that the error in the approximation procedure is determined by the error in the $\mathcal{L}$-projection (6.8) which is just the Rayleigh-Ritz-Galerkin procedure of Section 5. Specifically, when $u \in C^1((0,\infty), V)$ we differentiate (6.8) with respect to t and deduce that $u'_\ell(t) \in S$ is the $\mathcal{L}$-projection of $u'(t)$. This regularity of the solution u holds in both parabolic and pseudoparabolic cases.

We shall illustrate the use of the estimate (6.9) by applying it to a second order parabolic equation which is approximated by using a set of finite-element subspaces of degree one. Thus, suppose $\mathcal{S} \equiv \{S_h : h \in \mathcal{H}\}$ is a collection of closed subspaces of the closed subspace V of $H^1(G)$ and $\mathcal{S}$ is of degree 1; cf. Section 5.3. Let the continuous bilinear form $a(\cdot,\cdot)$ be V-elliptic and 0-regular; cf. Section III.6.4. Set $H = L^2(G) = H'$, so $\mathcal{M}$ is the identity, let $f \equiv 0$, and let $\ell(\cdot,\cdot) = a(\cdot,\cdot)$. If u is the solution of (6.1) and u_h is the solution of (6.6) with $S = S_h$, then the differentiability in $t > 0$ of these functions is given by Corollary IV.6.4 and their convergence at $t = 0^+$ is given by Exercise IV.7.8. We assume the form adjoint to $a(\cdot,\cdot)$ is 0-regular and obtain from (5.11) the estimates

$$\left.\begin{aligned} \|u(t) - u_\ell(t)\|_{L^2(G)} &\le c_2 h^2 \|Au(t)\|_{L^2(G)} \ , \\ \|u'(t) - u'_\ell(t)\|_{L^2(G)} &\le c_2 h^2 \|A^2 u(t)\|_{L^2(G)} \ , \qquad t > 0 \ . \end{aligned}\right\} \tag{6.11}$$

The a-priori estimate obtained from (6.3) shows that $|u(t)|_H$ is non-increasing and it follows similarly that $|Au(t)|_H$ is non-increasing for $t > 0$. Thus, if $u_0 \in D(A^2)$ we obtain from (6.9), and (6.11) the error estimate

$$\|u(t) - u_h(t)\|_{L^2(G)} \le c_2 h^2 \{\|Au_0\|_{L^2(G)} + t\|A^2 u_0\|_{L^2(G)}\} \ . \tag{6.12}$$

Although (6.12) gives the correct rate of convergence, it is far from optimal in the hypotheses assumed. For example, one can use estimates from Theorem IV.6.2 to play off the factors t and $\|Au'(t)\|_H$ in the second term of (6.12) and

thereby relax the assumption $u_0 \in D(A^2)$. Also, corresponding estimates can be obtained for the non-homogeneous equation and faster convergence rates can be obtained if approximating subspaces of higher degree are used.

6.3

We turn now to consider the approximation of second-order evolution equations of the type discussed in Section VI.2. Thus, we let $\mathcal{A}$ and $\mathcal{C}$ be the respective Riesz maps of the Hilbert spaces V and W, where V is dense and continuously embedded in W, hence, W' is identified with a subspace of V'. Let $\mathcal{B} \in \mathcal{L}(V, V')$, $u_0 \in V$, $u_1 \in W$ and $f \in C((0,\infty), W')$. We shall approximate the solution $u \in C([0,\infty), V) \cap C^1((0,\infty), V) \cap C^1([0,\infty), W) \cap C^2((0,\infty), W)$ of

$$\mathcal{C}u''(t) + \mathcal{B}u'(t) + \mathcal{A}u(t) = f(t) \ , \qquad t > 0 \ , \tag{6.13}$$

with the initial conditions $u(0) = u_0$, $u'(0) = u_1$. Equations of this form were solved in Section VI.2 by reduction to an equivalent first-order system of the form (6.1) on appropriate product spaces. We recall here the construction, since it will be used for the approximation procedure. Define $V_m \equiv V \times W$ with the scalar product

$$([x_1, x_2], [y_1, y_2]) = (x_1, y_1)_V + (x_2, y_2)_W \ , \qquad [x_1, x_1], [y_1, y_1] \in V \times W \ ,$$

so $V_m' = V' \times W'$; the Riesz map $\mathcal{M}$ of V_m onto V_m' is given by

$$\mathcal{M}([x_1, x_2]) = [\mathcal{A}x_1, \mathcal{C}x_2] \ , \qquad [x_1, x_2] \in V_m \ .$$

Define $V_\ell = V \times V$ and $\mathcal{L} \in \mathcal{L}(V_\ell, V_\ell')$ by

$$\mathcal{L}([x_1, x_2]) = [-\mathcal{A}x_2, \mathcal{A}x_1 + \mathcal{B}x_2] \ , \qquad [x_1, x_2] \in V_\ell \ .$$

Then Theorem VI.2.1 applies if $\mathcal{B}$ is monotone to give existence and uniqueness of a solution $w \in C^1([0,\infty), V_m)$ of

$$\mathcal{M}w'(t) + \mathcal{L}w(t) = [0, f(t)] \ , \qquad t > 0 \tag{6.14}$$

with $w(0) = [u_0, u_1]$ and $f \in C^1([0,\infty), W')$ given so that $u_0, u_1 \in V$ with $\mathcal{A}u_0 + \mathcal{B}u_1 \in W'$. The solution is given by $w(t) = [u(t), u'(t)]$, $t \geq 0$;

from the inclusion $[u, u'] \in C^1([0,\infty), V \times W)$ and (6.14) we obtain $[u, u'] \in C^1([0,\infty), V \times V)$. From (6.4) follows the a-priori estimate

$$\|u(t)\|_V^2 + \|u'(t)\|_W^2 + 2\int_0^t \mathcal{B}u'(s)(u'(s))\,ds$$
$$\leq e(\|u_0\|_V^2 + \|u_1\|_W^2) + Te\int_0^t \|f(s)\|_{W'}^2\,ds\ , \qquad 0 \leq t \leq T\ ,$$

on a solution $w(t) = [u(t), u'(t)]$ of (6.14).

The Faedo-Galerkin approximation procedure for the second-order equation is just the corresponding procedure for (6.14) as given in Section 6.1. Thus, if S is a finite-dimensional subspace of V, then we let w_s be the solution in $C^1([0,\infty), S \times S)$ of the equation

$$(w_s'(t), v)_m + \ell(w(t), v) = [0, f(t)](v)\ , \qquad v \in S \times S\ ,\ t > 0\ , \qquad (6.15)$$

with an initial value $w_s(0) \in S \times S$ to be prescribed below. If we look at the components of $w_s(t)$ we find from (6.15) that $w_s(t) = [u_s(t), u_s'(t)]$ for $t > 0$ where $u_s \in C^2([0,\infty), S)$ is the soluton of

$$(u_s''(t), v)_W + b(u_s'(t), v) + (u_s(t), v)_V = f(t)(v)\ , \quad v \in S\ ,\ t > 0\ . \quad (6.16)$$

Here $b(\cdot,\cdot)$ denotes the bilinear form on V corresponding to $\mathcal{B}$. As in Section 6.1, we can choose a basis for S and use it to write (6.16) as a system of m ordinary differential equations of second order. Of course this system is equivalent to a system of $2m$ equations of first order as given by (6.15), and this latter system may be the easier one in which to do the computation.

6.4

Error estimates for the approximation of (6.13) by the related (6.16) will be obtained in a special case by applying Theorem 6.1 directly to the situation described in Section 6.3. Note that in the derivation of (6.9) we needed only that $\mathcal{L}$ is monotone. Since $\mathcal{B}$ is monotone, the estimate (6.9) holds in the present situation. This gives an error bound in terms of the $\mathcal{L}$-projection $w_\ell(t) \in S \times S$ of the solution $w(t)$ of (6.14) as defined by

$$\ell(w_\ell(t), v) = \ell(w(t), v)\ , \qquad v \in S \times S\ . \qquad (6.17)$$

The bilinear form $\ell(\cdot,\cdot)$ is not coercive over V_ℓ so we might not expect $w_\ell(t) - w(t)$ to be small. However, in the special case of $\mathcal{B} = \varepsilon\mathcal{A}$ for some $\varepsilon \geq 0$ we

find that (6.17) is equivalent to a pair of similar identities in the component spaces. That is, if $e(t) \equiv w(t) - w_\ell(t)$ denotes the error in the $\mathcal{L}$-projection, and if $e(t) = [e_1(t), e_2(t)]$, then (6.17) is equivalent to

$$(e_j(t), v)_V = 0 \ , \qquad v \in S \ , \ j = 1, 2 \ . \tag{6.18}$$

Thus, if we write $w_\ell(t) = [u_\ell(t), v_\ell(t)]$, we see that $u_\ell(t)$ is the V-projection of $u(t)$ on S and $v_\ell(t) = u_\ell'(t)$ is the projection of $u'(t)$ on S. It follows from these remarks that we have

$$\|u(t) - u_\ell(t)\|_V \le \inf\{\|u(t) - v\|_V : v \in S\} \tag{6.19}$$

and corresponding estimates on $u'(t) - u_\ell'(t)$ and $u''(t) - u_\ell''(t)$. Our approximation results for (6.13) can be summarized as follows.

Theorem 6.2 *Let the Hilbert spaces V and W, operators $\mathcal{A}$ and $\mathcal{C}$, and data u_0, u_1 and f be given as in Theorem VI.2.1. Suppose furthermore that $\mathcal{B} = \varepsilon\mathcal{A}$ for some $\varepsilon \ge 0$ and that S is a finite-dimensional subspace of V. Then there exists a unique solution $u \in C^1([0,\infty), V) \cap C^2([0,\infty), W)$ of* (6.13) *with $u(0) = u_0$ and $u'(0) = u_1$; and there exists a unique solution $u_s \in C^2([0,\infty), S)$ of* (6.16) *with initial data determined by*

$$(u_s(0) - u_0, v)_V = (u_s'(0) - u_1, v)_V = 0 \ , \qquad v \in S \ .$$

We have the error estimate

$$\begin{aligned} &(\|u(t) - u_s(t)\|_V^2 + \|u'(t) - u_s'(t)\|_W^2)^{1/2} \\ &\quad \le (\|u(t) - u_\ell(t)\|_V^2 + \|u'(t) - u_\ell'(t)\|_W^2)^{1/2} \\ &\qquad + \int_0^t (\|u'(s) - u_\ell'(s)\|_V^2 + \|u''(s) - u_\ell''(s)\|_W^2)^{1/2}\, ds \ , \qquad t \ge 0 \end{aligned} \tag{6.20}$$

where $u_\ell(t) \in S$ is the V-projection of $u(t)$ defined by

$$(u_\ell(t), v)_V = (u(t), v)_V \ , \qquad v \in S \ .$$

Thus (6.19) *holds and provides a bound on* (6.20).

Finally we indicate how the estimate (6.20) is applied with finite-element or spline function spaces. Suppose $\mathcal{S} = \{S_h : h \in \mathcal{H}\}$ is a collection of finite-dimensional subspaces of the closed subspace V of $H^1(G)$. Let $k+1$ be the

degree of $\mathcal{S}$ which satisfies the approximation assumption (5.9). The scalar-product on V is equivalent to the $H^1(G)$ scalar-product and we assume it is k-regular on V. For each $h \in \mathcal{H}$ let u_h be the solution of (6.16) described above with $S = S_h$, and suppose that the solution u satisfies the regularity assumptions $u, u' \in L^\infty([0,T], H^{k+2}(G))$ and $u'' \in L^1([0,T], H^{k+2}(G))$. Then there is a constant c_0 such that

$$(\|u(t) - u_h(t)\|_V^2 + \|u'(t) - u_h'(t)\|_h^2)^{1/2}$$
$$\leq c_0 h^{k+1} , \qquad h \in \mathcal{H} , \ 0 \leq t \leq T . \tag{6.21}$$

The preceding results apply to wave equations (cf. Section VI.2.1), viscoelasticity equations such as VI.(2.9), and Sobolev equations (cf. Section VI.3).

Exercises

1.1. Show that a solution of the Neumann problem $-\Delta_n u = F$ in G, $\partial u/\partial v = 0$ on ∂G is a $u \in H^1(G)$ at which the functional (1.3) attains its minimum value.

2.1. Show that $F : K \to \mathbb{R}$ is weakly lower-semi-continuous at each $x \in K$ if and only if $\{x \in V : F(x) \leq a\}$ is weakly closed for every $a \in \mathbb{R}$.

2.2. In the proof of Theorem 2.3, show that $\varphi'(t) = F'(y + t(x-y))(x-y)$.

2.3. In the proof of Theorem 2.7, verify that M is closed and convex.

2.4. Prove Theorem 2.9.

2.5. Let F be G-differentiable on K. If F' is strictly monotone, prove directly that (2.5) has at most one solution.

2.6. Let G be bounded and open in $\mathbb{R}^n$ and let $F : G \times \mathbb{R} \to \mathbb{R}$ satisfy the following: $F(\cdot, u)$ is measurable for each $u \in \mathbb{R}$, $F(x, \cdot)$ is absolutely continuous for almost every $x \in G$, and the estimates

$$|F(x,u)| \leq a(x) + b|u|^2 \quad , \quad |\partial_u F(x,u)| \leq c(x) + b|u|$$

hold for all $u \in \mathbb{R}$ and a.e. $x \in G$, where $a(\cdot) \in L^1(G)$ and $c(\cdot) \in L^2(G)$.

(a) Define $E(u) = \int_G F(x, u(x))\,dx$, $u \in L^2(G)$, and show

$$E'(u)(v) = \int_G \partial_u F(x, u(x))v(x)\,dx\ , \qquad u, v \in L^2(G)\ .$$

(b) Show E' is monotone if $\partial_u F(x, \cdot)$ is non-decreasing for a.e. $x \in G$.

(c) Show E' is coercive if for some $k > 0$ and $c_0(\cdot) \in L^2(G)$ we have

$$\partial_u F(x, u) \cdot u \geq k|u|^2 - c_0(x)|u|\ ,$$

for $u \in \mathbb{R}$ and a.e. $x \in G$.

(d) State and prove some existence theorems and uniqueness theorems for boundary value problems containing the semi-linear equation

$$-\Delta_n u + f(x, u(x)) = 0\ .$$

2.7. Let G be bounded and open in $\mathbb{R}^n$. Suppose the function $F : G \times \mathbb{R}^{n+1} \to \mathbb{R}$ satisfies the following: $F(\cdot, \hat{u})$ is measurable for $\hat{u} \in \mathbb{R}^{n+1}$, $F(x, \cdot) : \mathbb{R}^{n+1} \to \mathbb{R}$ is (continuously) differentiable for a.e. $x \in G$, and the estimates

$$|F(x, \hat{u})| \leq a(x) + b\sum_{j=0}^{n} |u_j|^2 \quad , \quad |\partial_k F(x, \hat{u})| \leq c(x) + b\sum_{j=0}^{n} |u_j|$$

as above for every k, $0 \leq k \leq n$, where $\partial_k = \frac{\partial}{\partial u_k}$.

(a) Define $E(u) = \int_G F(x, u(x), \nabla u(x))\,dx$, $u \in H^1(G)$, and show

$$E'(u)(v) = \int_G \sum_{j=0}^{n} \partial_j F(x, u, \nabla u)\partial_j v(x)\,dx\ , \qquad u, v \in H^1(G)\ .$$

(b) Show E' is monotone if

$$\sum_{j=0}^{n} (\partial_j F(x, u_0, u_1, \ldots, u_n) - \partial_j F(x, v_0, v_1, \ldots, v_n))(u_j - v_j) \geq 0$$

for all $\hat{u}, \hat{v} \in \mathbb{R}^{n+1}$ and a.e. $x \in G$.

(c) Show E' is coercive if for some $k > 0$ and $c_0(\cdot) \in L^2(G)$

$$\sum_{j=0}^{n} \partial_j F(x, \hat{u})u_j \geq k\sum_{j=0}^{n} |u_j|^2 - c_0(x)\sum_{j=0}^{n} |u_j|$$

for $\hat{u} \in \mathbb{R}^{n+1}$ and a.e. $x \in \mathbb{R}^n$.

(d) State and prove an existence theorem and a uniqueness theorem for a boundary value problem containing the nonlinear equation

$$\sum_{j=0}^{n} \partial_j F_j(x, u, \nabla u) = f(x) \ .$$

3.1. Prove directly that (3.4) has at most one solution when $a(\cdot, \cdot)$ is (strictly) positive.

3.2. Give an example of a stretched membrane (or string) problem described in the form (3.6). Specifically, what does g represent in this application?

4.1. Show the following optimal control problem is described by the abstract setting of Section 4.1: find an admissible control $u \in U_{\text{ad}} \subset L^2(G)$ which minimizes the function

$$J(u) = \int_G |y(u) - w|^2 \, dx + c \int_G |u|^2 \, dx$$

subject to the state equations

$$\begin{cases} -\Delta_n y = F + u & \text{in } G \ , \\ y = 0 & \text{on } \partial G \ . \end{cases}$$

Specifically, identify all the spaces and operators in the abstract formulation.

4.2. Give sufficient conditions on the data above for existence of an optimal control. Write out the optimality system (4.10) for cases analogous to Sections 4.5 and 4.6.

5.1. Write out the special cases of Theorems 5.1 and 5.2 as they apply to the boundary value problem

$$\begin{cases} -\partial(p(x)\partial u(x)) + q(x)u(x) = f(x) \ , & 0 < x < 1 \ , \\ u(0) = u(1) = 0 \ . \end{cases}$$

Give the algebraic problem (5.5) and error estimates that occur when the piecewise-linear functions of Section 5.4 are used.

5.2. Repeat the above for the boundary value problem

$$\begin{cases} -\partial(p(x)\partial u(x)) + q(x)u(x) = f(x) \ , \\ u'(0) = u'(1) = 0 \ . \end{cases}$$

(Note that the set K and subspaces are not exactly as above.)

5.3. We describe an *Hermite interpolation* by piecewise-cubics. Let the interval G and partition P be given as in Section 5.4. Let $V \leq H^2(G)$ and define

$$K = \{v \in V : v(x_j) = v'(x_j) = 0 \ , \quad 0 \leq j \leq N+1\} \ .$$

(a) Let $f \in V$ and define $F(v) = (\frac{1}{2})|\partial^2(v-f)|_{L^2(G)}$. Show there is a unique $u_f \in K : (\partial^2(u_f - f), \partial^2 v)_{L^2(G)} = 0$, $v \in K$.

(b) Show there exists a unique $f_0 \in H^2(G)$ for which f_0 is a cubic polynomial on each $[x_j, x_{j+1}]$, $f_0(x_j) = f(x_j)$ and $f_0'(x_j) = f'(x_j)$ for $j = 0, 1, \ldots, N+1$.

(c) Construct a corresponding family of subspaces as in Theorem 5.4 and show it is of degree 3.

(d) Repeat exercise 5.1 using this family of approximating subspaces.

5.4. Repeat exercise 5.3 but with $V = H_0^2(G)$ and

$$K = \{v \in V : v(x_j) = 0 \ , \quad 0 \leq j \leq N+1\} \ .$$

Show that the corresponding *Spline interpolant* is a piecewise-cubic, $f_0(x_j) = f(x_j)$ for $0 \leq j \leq N+1$, and f_0 is in $C^2(G)$.

6.1. Describe the results of Sections 6.1 and 6.2 as they apply to the problem

$$\begin{cases} \partial_t u(x,t) - \partial_x(p(x)\partial_x u(x,t)) = F(x,t) \ , \\ u(0,t) = u(1,t) = 0 \ , \\ u(x,0) = u_0(x) \ . \end{cases}$$

Use the piecewise-linear approximating subspaces of Section 5.4.

6.2. Describe the results of Sections 6.3 and 6.4 as they apply to the problem

$$\begin{cases} \partial_t^2 u(x,t) - \partial_x(p(x)\partial_x u(x,t)) = F(x,t) \ , \\ u(0,t) = u(1,t) = 0 \ , \\ u(x,0) = u_0(x) \ , \quad \partial_t u(x,0) = u_1(x) \ . \end{cases}$$

Use the subspaces of Section 5.4.