L.P. Lebedev I.I. Vorovich

Functional Analysis in Mechanics

Springer

L.P. Lebedev
Departamento de Matemáticas
Universidad Nacional de Colombia
Bogotá
Colombia
lebedev@uolpremium.net.co

I.I. Vorovich
(*deceased*)

Mathematics Subject Classification (2000): 46-01, 35-01, 74-01, 76-01, 74Kxx, 35Dxx, 35Qxx

Library of Congress Cataloging-in-Publication Data
Lebedev, L.P
Functional analysis in mechanics / L.P. Lebedev, I.I. Vorovich.
p. cm. — (Springer monographs in mathematics)
Includes bibliographical references and index.
ISBN 0-387-95519-4 (hc : alk. paper)
1. Functional analysis. I. Lebedev, Leonid Petrovich, 1946– II. Title. III. Series.
QA320 .L3483 2002
515′.7—dc21 2002075732

ISBN 0-387-95519-4 Printed on acid-free paper.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1 SPIN 10881937

Typesetting: Pages created by the authors using a Springer LaTeX 2e macro package.

www.springer-ny.com

Springer-Verlag New York Berlin Heidelberg
A member of BertelsmannSpringer Science+Business Media GmbH

Preface to the English Edition

This book started about 30 years ago as a course of lectures on functional analysis given by a youthful Prof. I.I. Vorovich to his students in the Department of Mathematics and Mechanics (division of Mechanics) at Rostov State University. That course was subsequently extended through the offering, to those same students, of another course called Applications of Functional Analysis. Later, the courses were given to pure mathematicians, and even to engineers, by both coauthors. Although experts in mechanics are quick to accept results concerning uniqueness or non-uniqueness of solutions, many of these same practitioners seem to hold a rather negative view concerning theorems of existence. Our goal was to overcome this attitude of reluctance toward existence theorems, and to show that functional analysis does contain general ideas that are useful in applications. This book was written on the basis of our lectures, and was then extended by the inclusion of some original results which, although not very new, are still not too well known.

We mentioned that our lectures were given to students of the Division of Mechanics. It seems that only in Russia are such divisions located within departments of mathematics. The students of these divisions study mathematics on the level of mathematicians, but they are also exposed to much material that is normally given at engineering departments in the West. So we expect that the book will be useful for western engineering departments as well.

This book is a revised and extended translation of the Russian edition of the book, and is published by permission of editor house Vuzovskskaya Kniga, Moscow. We would like to thank Prof. Michael Cloud of Lawrence

Technological University for assisting with the English translation, for producing the LaTeX files, and for contributing the problem hints that appear in the Appendix.

Department of Mechanics and Mathematics L.P. Lebedev
Rostov State University, Russia

&

Department of Mathematics
National University of Colombia, Colombia

Department of Mechanics and Mathematics I.I. Vorovich
Rostov State University, Russia

Preface to the Russian Edition

This is an extended version of a course of lectures we have given to third and fourth year students of mathematics and mechanics at Rostov State University. Our lecture audience typically includes students of applied mechanics and engineering. These latter students wish to master methods of contemporary mathematics in order to read the scientific literature, justify the numerical and analytical methods they use, and so on; they lack enthusiasm for courses in which applications appear only after long uninterrupted stretches of theory. Finally, the audience includes mathematicians. These listeners, already knowing more functional analysis than the course has to offer, are interested only in applications. In order to please such a diverse audience, we have had to arrange the course carefully and introduce sensible applications from the beginning. The brevity of the course — and the boundless extent of functional analysis — force us to present only those topics essential to the chosen applications. We do, however, try to make the course self-contained and to cover the foundations of functional analysis.

We assume that the reader knows the elements of mathematics at the beginning graduate or advanced undergraduate level. Those subjects assumed are typical of most engineering curricula: calculus, differential equations, mathematical physics, and linear algebra. A knowledge of mechanics, although helpful, is not necessary; we wish to attract all types of readers interested in the applications and foundations of functional analysis. We hope that not only students of engineering and applied mechanics will ben-

efit, but that some mathematicians or physicists will discover tools useful for their research as well.

Department of Mechanics and Mathematics L.P. Lebedev
Rostov State University, Russia

Department of Mechanics and Mathematics I.I. Vorovich
Rostov State University, Russia

Contents

This page intentionally left blank

Introduction

Long ago it was traditional to apply mathematics only to mechanics and physics. Now it is almost impossible to find an area of knowledge in which mathematics is not used as a tool to create new models and to simulate them. This is due mainly to the fantastic ability of computers to process models having thousands of parameters.

In view of the fact that mathematics has become such a central tool, it is fortunate that mathematics itself tends to produce methods of great generality. Functional analysis, in particular, allows us to approach different mathematical facts and methods from a unified point of view. Let us consider some examples.

Example 1. A system of linear algebraic equations

$$x_i = \sum_{j=1}^{n} a_{ij} x_j + c_i, \qquad i = 1, \dots, n, \tag{1}$$

can be solved by the method of successive approximations in the form

$$\begin{aligned} x_i^{(0)} &= c_i, \\ x_i^{(k+1)} &= \sum_{j=1}^{n} a_{ij} x_j^{(k)} + c_i, \qquad i = 1, \dots, n, \quad k = 1, 2, \dots. \end{aligned}$$

To establish convergence of the scheme, let us consider the difference

$$x_i^{(k+1)} - x_i^{(k)} = \sum_{j=1}^{n} a_{ij} [x_j^{(k)} - x_j^{(k-1)}].$$

We have

$$\max_{1\le i\le n}|x_i^{(k+1)}-x_i^{(k)}| \le \max_{1\le i\le n}\sum_{j=1}^{n}|a_{ij}||x_j^{(k)}-x_j^{(k-1)}|$$
$$\le q\cdot\max_{1\le j\le n}|x_j^{(k)}-x_j^{(k-1)}|$$

where

$$q=\max_{1\le i\le n}\sum_{j=1}^{n}|a_{ij}|.$$

Convergence of the method is ensured if $q<1$. Then $(z_1,\dots,z_n)$, where

$$z_i=\lim_{k\to\infty}x_i^{(k)}$$

for $i=1,\dots,n$, is a solution to (1).

Now let us apply the successive approximation scheme to a system of integral equations

$$x_i(t)=\sum_{j=1}^{n}\int_0^1 a_{ij}(t,s)x_j(s)\,ds+c_i(t),\qquad i=1,\dots,n,\tag{2}$$

where $c_i(t)$ and $a_{ij}(t,s)$ are given continuous functions on $[0,1]$ and $[0,1]\times[0,1]$, respectively. The scheme is

$$x_i^{(0)}(t)=c_i(t),$$
$$x_i^{(k+1)}(t)=\sum_{j=1}^{n}\int_0^1 a_{ij}(t,s)x_j^{(k)}(s)\,ds+c_i(t),\qquad i=1,\dots,n.$$

For the difference of two successive approximations, we have

$$x_i^{(k+1)}(t)-x_i^{(k)}(t)=\sum_{j=1}^{n}\int_0^1 a_{ij}(t,s)[x_j^{(k)}(s)-x_j^{(k-1)}(s)]\,ds$$

so

$$|x_i^{(k+1)}(t)-x_i^{(k)}(t)|\le\sum_{j=1}^{n}\int_0^1|a_{ij}(t,s)|\,|x_j^{(k)}(s)-x_j^{(k-1)}(s)|\,ds.$$

Thus

$$\max_{\substack{1\le i\le n\\0\le t\le 1}}|x_i^{(k+1)}(t)-x_i^{(k)}(t)|\le\max_{\substack{1\le i\le n\\0\le t\le 1}}\sum_{j=1}^{n}\int_0^1|a_{ij}(t,s)|\,ds\cdot$$
$$\cdot\max_{\substack{1\le j\le n\\0\le s\le 1}}|x_j^{(k)}(s)-x_j^{(k-1)}(s)|.$$

It follows that for

$$q = \max_{\substack{1\le i\le n\\ 0\le t\le 1}} \sum_{j=1}^{n} \int_0^1 |a_{ij}(t,s)|\, ds < 1$$

the sequence $\{x_i^{(k)}(t)\}$ $(i = 1, \dots, n)$ is uniformly convergent on $[0, 1]$; hence there exists a limit $z_i(t) = \lim_{k\to\infty} x_i^{(k)}(t)$, and $(z_1(t), \dots, z_n(t))$ is a solution to (2).

The obvious similarity between the treatments of (1) and (2) suggests that some general approach might cover these and many other cases of interest.

Example 2. In what follows, we shall deal mainly with spaces of infinite dimension. For example, the wave equation

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} \tag{3}$$

describes the vibrations $u = u(x, t)$ of a stretched string. Let the string ends be fixed:

$$u(0, t) = u(1, t) = 0.$$

It is natural to seek a solution with finite potential and kinetic energies, i.e., with

$$\int_0^1 \left(\frac{\partial u}{\partial x}\right)^2 dx < \infty, \qquad \int_0^1 \left(\frac{\partial u}{\partial t}\right)^2 dx < \infty.$$

We could seek a solution in the form of a Fourier series

$$u(x, t) = \sum_{k,m} A_{km} \sin \pi k x \sin \pi m t.$$

This solution is evidently described by a denumerable set of numbers A_{km}, which can be regarded as the components of a vector having an infinite number of components. The set of such "vectors" clearly constitutes a space that is infinite dimensional.

One of the difficulties in dealing with an infinite dimensional space is that the Bolzano–Weierstrass principle (that any bounded infinite sequence contains a convergent subsequence) breaks down. For example, we cannot select a convergent subsequence from the bounded sequence of functions $y_k = \sin kx$, $k = 1, 2, \dots$.

Example 3. In contemporary mathematical physics, generalized solutions are typical. Without going into too much detail, we may briefly consider the problem of a bar with clamped ends bending under a load $q(x)$. A corresponding boundary value problem is

$$(B(x)y''(x))'' - q(x) = 0, \qquad y(0) = y'(0) = y(l) = y'(l) = 0, \tag{4}$$

where $B(x)$ and l are the stiffness and length, respectively, of the bar. This formulation supposes $y = y(x)$ to possess derivatives up to fourth order.

The same boundary value problem can be posed differently through the use of variational principles. It can be shown that the functional I defined by

$$I(y) = \frac{1}{2}\int_0^l [B(y'')^2 - 2q(x)y]\,dx$$

takes on a minimum value at an equilibrium state of the bar (here all $y(x)$ under consideration must satisfy the boundary conditions stated in (4)). The variation

$$\delta I = \int_0^l [B(x)y''(x)\varphi''(x) - q(x)\varphi(x)]\,dx$$

vanishes for any $\varphi(x)$ satisfying the boundary conditions in (4) if $y(x)$ satisfies (4). A function $y(x)$ is said to be a generalized solution to the problem (4) if the equation

$$\int_0^l [B(x)y''(x)\varphi''(x) - q(x)\varphi(x)]\,dx = 0 \quad (y(0) = y'(0) = y(l) = y'(l) = 0)$$

holds for any sufficiently smooth function $\varphi(x)$ such that

$$\varphi(0) = \varphi'(0) = \varphi(l) = \varphi'(l) = 0.$$

So a generalized solution satisfies the equilibrium equation in a Lagrange principle sense. For a moving system, we can introduce generalized solutions using Hamilton's variational principle.

Since the restrictions on smoothness for generalized solutions are milder than those for classical solutions, the above approach extends the circle of problems we may investigate. In particular, problems with non-smooth loads often occur in industrial applications. The approach also arises naturally when we study convergence of the finite element method — one of the most powerful tools of mathematical physics.

At this point we hope the reader has begun to picture functional analysis as a powerful tool in applications. We are therefore ready to begin a more systematic study of its fundamentals. Let us close this introduction by presenting two theorems of classical analysis. Both theorems bear Weierstrass's name, and will be used frequently in what follows.

Theorem 1. Let a sequence $\{f_n(\mathbf{x})\}$ of functions continuous on a compact set $\Omega \subset \mathbb{R}^k$ converge uniformly; that is, for any $\varepsilon > 0$ there is an integer $N = N(\varepsilon)$ such that

$$|f_{n+m}(\mathbf{x}) - f_n(\mathbf{x})| < \varepsilon$$

for any $n > N$, $m > 0$, and $\mathbf{x} \in \Omega$. Then there exists a limit function

$$f(\mathbf{x}) = \lim_{n\to\infty} f_n(\mathbf{x})$$

that is continuous on Ω.

Theorem 2. Let $f(\mathbf{x})$ be a function continuous on a compact set $\Omega \subset \mathbb{R}^k$. For any $\varepsilon > 0$ there is a polynomial $P_n(\mathbf{x})$ of the nth degree such that

$$|f(\mathbf{x}) - P_n(\mathbf{x})| < \varepsilon$$

for any $\mathbf{x} \in \Omega$.

We recall that in $\mathbb{R}^k$ the term "compact set" refers to a closed and bounded set.

This page intentionally left blank

1
Metric Spaces

1.1 Preliminaries

Consider a set of particles P_i, $i = 1, \ldots, n$. To locate these particles in the space $\mathbb{R}^3$, we need a reference system. Let the Cartesian coordinates of P_i be (ξ_i, η_i, ζ_i) for each i. Identifying (ξ_1, η_1, ζ_1) with (x_1, x_2, x_3), (ξ_2, η_2, ζ_2) with (x_4, x_5, x_6), and so on, we obtain a vector $\mathbf{x}$ of the Euclidean space $\mathbb{R}^{3n}$ with coordinates $(x_1, x_2, \ldots, x_{3n})$. This vector determines the positions of all particles in the set.

To distinguish different configurations $\mathbf{x}$ and $\mathbf{y}$ of the system, we can introduce a distance from $\mathbf{x}$ to $\mathbf{y}$:

$$d_E(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{3n} (x_i - y_i)^2 \right)^{1/2}.$$

This is the *Euclidean distance* (or *metric*) of $\mathbb{R}^{3n}$. Alternatively, we could characterize the distance from $\mathbf{x}$ to $\mathbf{y}$ using the function

$$d_S(\mathbf{x}, \mathbf{y}) = \max\{|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_{3n} - y_{3n}|\}.$$

It is easily seen that each of the metrics d_E and d_S satisfy the following properties, known as the *metric axioms*:

D1. $d(\mathbf{x}, \mathbf{y}) \geq 0$;

D2. $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$;

D3. $d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x})$;

D4. $d(\mathbf{x},\mathbf{y}) \le d(\mathbf{x},\mathbf{z}) + d(\mathbf{z},\mathbf{y})$.

Any real valued function $d(\mathbf{x},\mathbf{y})$ defined for all $\mathbf{x},\mathbf{y} \in \mathbb{R}^{3n}$ is called a *metric* on $\mathbb{R}^{3n}$ if it satisfies properties D1–D4. Property D1 is called the *axiom of positiveness*, property D3 is called the *axiom of symmetry*, and property D4 is called the *triangle inequality*.

Problem 1.1.1. Let a real valued function $d(\mathbf{x},\mathbf{y})$ be defined for all $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$. Show that if d satisfies D2, D3, and D4, then it also satisfies D1. Confirm that this does not depend on the nature of the elements $\mathbf{x}$ and $\mathbf{y}$.

Remark 1.1.1. It follows from Problem 1.1.1 that the set of axioms for the metric can be restricted to just three of them: D2, D3, and D4.

With regard to sequence convergence in $\mathbb{R}^{3n}$, the metrics d_E and d_S are *equivalent* since there exist positive constants m_1, m_2 independent of $\mathbf{x}$ and $\mathbf{y}$ such that

$$0 < m_1 \le \frac{d_E(\mathbf{x},\mathbf{y})}{d_S(\mathbf{x},\mathbf{y})} \le m_2 < \infty \tag{1.1.1}$$

whenever $\mathbf{x},\mathbf{y} \in \mathbb{R}^{3n}$ and $\mathbf{x} \ne \mathbf{y}$. So

$$\lim_{k\to\infty} d_E(\mathbf{x}_k,\mathbf{x}) = 0 \implies \lim_{k\to\infty} d_S(\mathbf{x}_k,\mathbf{x}) = 0$$

and vice versa.

Remark 1.1.2. In what follows, we shall use the notation "m_i" for those constants whose exact values are not important.

Equation (1.1.1) shows that, in a certain way, $d_E(\mathbf{x},\mathbf{y})$ and $d_S(\mathbf{x},\mathbf{y})$ have the same standing as metrics on $\mathbb{R}^{3n}$. We can introduce other functions on $\mathbb{R}^{3n}$ satisfying axioms D1–D4: for example,

$$d_p(\mathbf{x},\mathbf{y}) = \left(\sum_{i=1}^{3n} |x_i - y_i|^p\right)^{1/p}, \qquad p = \text{ constant } \ge 1,$$

and

$$d_k(\mathbf{x},\mathbf{y}) = \left(\sum_{i=1}^{3n} k_i |x_i - y_i|^2\right)^{1/2}, \qquad k_i > 0.$$

Problem 1.1.2. Show that any two of the metrics introduced above are equivalent on $\mathbb{R}^n$. Note that two metrics $d_1(\mathbf{x},\mathbf{y})$ and $d_2(\mathbf{x},\mathbf{y})$ on $\mathbb{R}^n$ are equivalent if there exist m_1, m_2 such that

$$0 < m_1 \le \frac{d_1(\mathbf{x},\mathbf{y})}{d_2(\mathbf{x},\mathbf{y})} \le m_2 < \infty$$

for any $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$ such that $\mathbf{x} \ne \mathbf{y}$.

The notion of metric generalizes the notion of distance in $\mathbb{R}^3$. It can be applied not only to particle locations but also to particle velocities, accelerations, and masses, in order to distinguish between different states of a given system or between different systems of particles. The same can be done for any system described by a finite number of parameters (forces, temperatures, etc.).

Now let us consider how to extend the idea of distance to continuum problems. Take a string, with fixed ends, of length π. For a loaded string, we can use the Fourier expansion

$$u(s) = \sum_{k=1}^{\infty} x_k \sin ks \tag{1.1.2}$$

to describe the string displacement $u(s)$. Any state of the string can be identified with a vector $\mathbf{x}$ having infinitely many coordinates x_i, $i = 1, 2, \ldots$. The dimension of the space S of all such vectors is obviously not finite.

We can modify the metric of $\mathbb{R}^n$ to determine the distance from $\mathbf{x}$ to $\mathbf{y}$ in S. The necessary changes are evident; we can use

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{\infty} |x_i - y_i|^p \right)^{1/p}, \qquad p \geq 1,$$

or

$$d(\mathbf{x}, \mathbf{y}) = \sup_i |x_i - y_i|.$$

The distances so defined satisfy D1–D4, hence are metrics. So we have an analogy between $\mathbb{R}^n$ and S — but there are some differences. Consider, for instance, the distance from $\mathbf{0} = (0, 0, \ldots)$ to $\mathbf{x}_0 = (1, 1/2, 1/3, \ldots)$ using the metrics

$$d_1(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{\infty} |x_i - y_i|, \qquad d_2(\mathbf{x}, \mathbf{y}) = \sup_i |x_i - y_i|.$$

Since

$$d_1(\mathbf{x}_0, \mathbf{0}) = \sum_{i=1}^{\infty} 1/i, \qquad d_2(\mathbf{x}_0, \mathbf{0}) = 1,$$

and the series diverges, we do not have

$$d_1(\mathbf{x}_0, \mathbf{0}) / d_2(\mathbf{x}_0, \mathbf{0}) \leq m_2 < \infty.$$

Hence these metrics are not equivalent (moreover, they are defined on different subsets of S). So on an infinite dimensional space, different metrics can determine different properties of sequence convergence.

Definition 1.1.1. A set X is called a *metric space* if for each pair of points $x, y \in X$ there is defined a metric (a real valued function) which satisfies axioms D1–D4.

Roughly speaking, a metric space consists of a set X along with an appropriate metric d; it can therefore be regarded as an ordered pair (X, d).

Remark 1.1.3. In the following pages, we shall not distinguish between metric spaces based on the same set of elements if their metrics are equivalent. That is, if d_1 and d_2 are equivalent metrics, then we shall not distinguish between (X, d_1) and (X, d_2). Metric spaces with non-equivalent metrics, even those consisting of the same set of elements, are different for us. By the above example we are made to distinguish the metric spaces consisting of elements of S. Moreover, these spaces with non-equivalent metrics consist of different subsets of elements of S. So *S is not a metric space, but only a (linear) set of infinite dimensional vectors whose linear subsets (subspaces) of elements, together with their metrics, can be various metric spaces of vectors with infinite numbers of coordinates.*

In the definition of metric space the nature of the elements of the space is unimportant. The elements could be abstract objects, even ordinary objects such as chairs or tables — it is merely necessary that we can introduce for each pair of elements of the set a function satisfying the axioms of a metric. In mathematical physics, metric spaces of functions are usually employed. These are the spaces to which solutions of some equations and/or the parameters of a problem must belong. During the rigorous investigation of such problems, some restrictions are always imposed on the properties of the solutions sought. This is due not only to a desire for rigor and formalism; some mathematical problems have several solutions, some parts of which contradict our ideas about the nature of the process described by the problem. Additional restrictions based on the physical nature of the problem allow us to select physically reasonable solutions. One way to impose such restrictions is to require that the solution belong to a metric space. Thus the choice of space in which one seeks a solution can be crucial for the solution of realistic problems. Depending on this choice, solutions may exist or not, be unique or not, etc. Metric spaces in mathematical physics are usually linear and infinite dimensional.

Let us enumerate some metric spaces of infinite dimensional vectors $\mathbf{x} = (x_1, x_2, \ldots)$ (equivalently, of sequences $\mathbf{x} = \{x_i\}$).

1. *The metric space m.* The space m is the set of all bounded sequences; the metric is given by

$$d(\mathbf{x}, \mathbf{y}) = \sup_i |x_i - y_i|. \tag{1.1.3}$$

2. *The metric space ℓ^p.* The space ℓ^p $(p \geq 1)$ is the set of all sequences $\{x_i\}$ such that $\sum_{i=1}^{\infty} |x_i|^p < \infty$; the metric is

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{\infty} |x_i - y_i|^p \right)^{1/p}. \tag{1.1.4}$$

3. *The metric space c.* The space c is the linear subspace of m that consists of all convergent sequences; the metric is the metric of m.

4. *The metric space c_0.* The space c_0 is the subspace of c consisting of all sequences converging to 0; again, the metric is the metric of m.

The metrics of these spaces were introduced by analogy with metrics on $\mathbb{R}^n$. We now consider another class of metrics: the energy metrics.

5. *The energy space for a string.* The potential energy of a string is proportional to

$$\int_0^{2\pi} \left(\frac{\partial u}{\partial s} \right)^2 ds = \pi \sum_{k=1}^{\infty} k^2 x_k^2,$$

x_k being defined by (1.1.2). We can compare two states of the string by introducing the metric

$$d(u, v) \equiv d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{\infty} k^2 (x_k - y_k)^2 \right)^{1/2} \tag{1.1.5}$$

where

$$v(s) = \sum_{k=1}^{\infty} y_k \sin ks.$$

The energy space of the string is the set of all sequences of Fourier coefficients such that $\sum_{k=1}^{\infty} k^2 x_k^2 < \infty$; the metric is given by (1.1.5).

Problem 1.1.3. Show that (1.1.3)–(1.1.5) are indeed metrics on their respective sets.

Energy spaces are advantageous when applied to mechanics problems, as we shall see later.

6. *The metric space of straight lines.* The notion of metric space is abstract. A metric space can consist of elements that are not vectors. Consider, for example, the set M of all straight lines in the plane which do not pass through the coordinate origin. A straight line L is given by the equation $x \cos \alpha + y \sin \alpha - p = 0$. Let us show that

$$d(L_1, L_2) = \left[(p_1 - p_2)^2 + 4 \sin^2 \frac{\alpha_1 - \alpha_2}{2} \right]^{1/2}$$

is a metric on M. Axioms D1 and D3 are obviously satisfied. Consider D2. Certainly $d(L_1, L_2) = 0$ whenever $L_1 = L_2$. Conversely, $d(L_1, L_2) = 0$

implies both $p_1 = p_2$ and $\sin(\alpha_1 - \alpha_2)/2 = 0$; the latter condition gives $\alpha_1 - \alpha_2 = 2\pi n$ $(n = 0, \pm 1, \pm 2, \ldots)$ hence $L_1 = L_2$. Finally, consider D4. Since

$$4 \sin^2 \frac{\alpha_1 - \alpha_2}{2} = (\sin \alpha_1 - \sin \alpha_2)^2 + (\cos \alpha_1 - \cos \alpha_2)^2$$

we have

$$d(L_1, L_2) = \left[(p_1 - p_2)^2 + (\sin \alpha_1 - \sin \alpha_2)^2 + (\cos \alpha_1 - \cos \alpha_2)^2\right]^{1/2}.$$

Let $(p_i, \sin \alpha_i, \cos \alpha_i)$, for $i = 1, 2, 3$, be the coordinates of a point A_i in 3-dimensional Euclidean space. Noting that $d(L_i, L_j)$ equals the Euclidean distance from A_i to A_j in $\mathbb{R}^3$, we see that D4 is also satisfied.

1.2 Some Metric Spaces of Functions

To describe the behavior or change in state of a body in space, we use functions of one or more variables. Displacements, velocities, loads, and temperatures are all functions of position. So we must learn how to distinguish different states of a body; the appropriate tool for this is, of course, the notion of metric space. In mechanics of materials, we deal mostly with real-valued continuous or differentiable functions.

Let Ω be a closed and bounded domain in $\mathbb{R}^n$. A natural measure of the deviation between two continuous functions $f(\mathbf{x})$ and $g(\mathbf{x})$, $\mathbf{x} \in \Omega$, is

$$d(f, g) = \max_{\mathbf{x} \in \Omega} |f(\mathbf{x}) - g(\mathbf{x})|. \tag{1.2.1}$$

It is obvious that $d(f, g)$ satisfies axioms D1–D3. Let us verify D4. Since $|f(\mathbf{x}) - g(\mathbf{x})|$ is a continuous function on Ω, there exists a point $\mathbf{x}_0 \in \Omega$ such that

$$d(f, g) = \max_{\mathbf{x} \in \Omega} |f(\mathbf{x}) - g(\mathbf{x})| = |f(\mathbf{x}_0) - g(\mathbf{x}_0)|.$$

For any function $h(\mathbf{x})$ which is continuous on Ω, we get

$$\begin{aligned} d(f, g) &= |f(\mathbf{x}_0) - g(\mathbf{x}_0)| \\ &\le |f(\mathbf{x}_0) - h(\mathbf{x}_0)| + |h(\mathbf{x}_0) - g(\mathbf{x}_0)| \\ &\le d(f, h) + d(h, g). \end{aligned}$$

(Here we use the Weierstrass theorem that on a compact set a continuous function attains its maximum and minimum values.) Thus $d(f, g)$ in (1.2.1) is a metric.

Definition 1.2.1. Let Ω be a closed and bounded domain. $C(\Omega)$ is the metric space consisting of the set of all continuous functions on Ω supplied with the metric (1.2.1).

To take into account the derivatives of functions, we must use other metrics. One of these is

$$d(f,g) = \sum_{|\alpha|\le k} \max_{\mathbf{x}\in\Omega} |D^\alpha f(\mathbf{x}) - D^\alpha g(\mathbf{x})| \tag{1.2.2}$$

where

$$D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}, \qquad |\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_n.$$

Problem 1.2.1. Let $C^{(k)}(\Omega)$ denote the set of all continuous functions on a closed and bounded domain Ω whose derivatives up to order k are continuous on Ω. Show that this set is a metric space under the distance function (1.2.2).

On $C(\Omega)$, let us consider another metric:

$$d(f,g) = \left(\int_\Omega |f(\mathbf{x}) - g(\mathbf{x})|^p \, d\Omega \right)^{1/p} \qquad (p \ge 1) \tag{1.2.3}$$

where Ω is a compact domain in $\mathbb{R}^n$ that is Jordan measurable. We assume this for formula (1.2.3) to be well defined for any f, g being continuous on Ω. In applications, we use domains occupied by physical bodies that are of comparatively simple shape. We will always assume that such domains are Jordan measurable. Since the spaces $L^p(\Omega)$ and $W^{l,p}(\Omega)$ are auxiliary for our purposes (we use them to characterize physical objects) we shall assume the same Jordan measurability for Ω as well, without explicit mention.

Now let us show that (1.2.3) really represents a metric. The only nontrivial axiom to be verified is D4; its validity follows from the Minkowski inequality for integrals

$$\left(\int_\Omega |f_1(\mathbf{x}) + f_2(\mathbf{x})|^p \, d\Omega \right)^{1/p} \le \left(\int_\Omega |f_1(\mathbf{x})|^p \, d\Omega \right)^{1/p} + \left(\int_\Omega |f_2(\mathbf{x})|^p \, d\Omega \right)^{1/p} \tag{1.2.4}$$

which holds for any $p \ge 1$. With

$$f_1(\mathbf{x}) = f(\mathbf{x}) - h(\mathbf{x}), \qquad f_2(\mathbf{x}) = h(\mathbf{x}) - g(\mathbf{x}),$$

(1.2.4) becomes $d(f,g) \le d(f,h) + d(h,g)$, showing that $C(\Omega)$ is also a metric space under (1.2.3).

Although

$$\left(\int_\Omega |f(\mathbf{x}) - g(\mathbf{x})|^p \, d\Omega \right)^{1/p} \le (\operatorname{mes}\Omega)^{1/p} \max_{\mathbf{x}\in\Omega} |f(\mathbf{x}) - g(\mathbf{x})|,$$

we cannot find a constant m such that

$$\max_{\mathbf{x}\in\Omega} |f(\mathbf{x}) - g(\mathbf{x})| \le m \left(\int_\Omega |f(\mathbf{x}) - g(\mathbf{x})|^p d\Omega \right)^{1/p}$$

holds for any pair of continuous functions $f(\mathbf{x})$ and $g(\mathbf{x})$. (The reader should show this by constructing a counterexample.) Hence the metrics (1.2.1) and (1.2.3) are not equivalent on $C(\Omega)$.

Remark 1.2.1. For $0 < p < 1$, $d(f, g)$ in (1.2.3) is not a metric.

Another inequality for integrals, the Hölder inequality

$$\int_\Omega |f(\mathbf{x})g(\mathbf{x})|\, d\Omega \le \left(\int_\Omega |f(\mathbf{x})|^p\, d\Omega\right)^{1/p} \left(\int_\Omega |g(\mathbf{x})|^q\, d\Omega\right)^{1/q} \tag{1.2.5}$$

where $1/p+1/q = 1$, will be used frequently. Proofs of this and Minkowski's inequality can be found in [29].

Problem 1.2.2. Show that the function

$$d(f, g) = \int_0^1 |f'(x) - g'(x)|\, dx$$

is not a metric on the set of all functions that are continuous on $[0, 1]$. On what set is it a metric?

1.3 Energy Spaces

We have already introduced the energy space for a string. Let us consider other examples. In what follows, we shall employ only dimensionless variables, parameters, and functions of state of a body.

Bending of a Bar

In the Introduction we considered the problem of bending a clamped bar, which was governed by (4). The potential energy of the bar is

$$\mathcal{E}_1(y) = \frac{1}{2}\int_0^l B(x)(y'')^2\, dx.$$

On the set S consisting of all functions $y(x)$ that are twice continuously differentiable on $[0, l]$ and that satisfy

$$y(0) = y'(0) = y(l) = y'(l) = 0, \tag{1.3.1}$$

let us consider

$$d(y_1, y_2) = (2\mathcal{E}_1(y_1 - y_2))^{1/2} = \left(\int_0^l B(x)[y_1''(x) - y_2''(x)]^2\, dx\right)^{1/2}.$$

For this, D1 and D3 obviously hold. Satisfaction of D4 follows from the fact that $\mathcal{E}_1(y)$ is quadratic in y. To verify D2, we need only show that $d(y, z) = 0$ implies $y(x) = z(x)$. But $d(y, z) = 0$ implies $(y(x) - z(x))'' = 0$, hence $y(x) - z(x) = a_1 x + a_2$ where a_1, a_2 are constants; imposing (1.3.1), we arrive at $a_1 = a_2 = 0$. So $d(y_1, y_2)$ is indeed a metric on S.

Elastic Membranes

The potential energy of a membrane occupying a domain $\Omega \subset \mathbb{R}^2$ is proportional to

$$\mathcal{E}_2(u) = \int_\Omega \left[\left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 \right] dx\, dy.$$

So we can try

$$d(u, v) = (\mathcal{E}_2(u - v))^{1/2} \tag{1.3.2}$$

as a metric on the functions $u = u(x, y)$ that describe the normal displacements of the membrane. We first consider the case where the edge of the membrane is clamped, i.e.,

$$u\Big|_{\partial\Omega} = 0 \tag{1.3.3}$$

where $\partial\Omega$ is the boundary of Ω. The function $d(u, v)$ of (1.3.2) is a metric on the set $C^{(1)}(\Omega)$. Axioms D1 and D3 hold obviously; D2 holds by (1.3.3), and D4 holds by the quadratic nature of $\mathcal{E}_2(u)$. This space is appropriate for investigating the corresponding boundary value problem

$$\Delta u = -f, \qquad\qquad u\Big|_{\partial\Omega} = 0,$$

called the Dirichlet problem for Poisson's equation. This describes the behavior of the clamped membrane under a load $f = f(x, y)$.

Another main problem for Poisson's equation, the Neumann problem, is determined by the boundary condition

$$\frac{\partial u}{\partial n}\Big|_{\partial\Omega} = 0. \tag{1.3.4}$$

A result from the calculus of variations is that a minimizer of the functional

$$J(u) = \frac{1}{2} \int_\Omega \left[\left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 - 2fu \right] dx\, dy \tag{1.3.5}$$

is a solution to the Neumann problem that can be formulated as

Problem 1.3.1. Given $f(x, y) \in C(\Omega)$, find a minimizer $u(x, y)$ of $J(u)$ such that $u(x, y) \in C^{(1)}(\Omega)$.

The boundary condition (1.3.4) appears here as a natural one; we need not formulate it in advance. That is why we do not require any boundary conditions on functions constituting the energy space for the Neumann problem. If we take (1.3.2) as a metric for this energy space, we see that D2 is not fulfilled: from $d(u, v) = 0$ it follows that $u(x, y) - v(x, y) =$ constant. There are two ways in which we can make use of the energy metric for this problem. One is to introduce a space whose elements are actually equivalence classes of functions, two functions belonging to the same class

(and hence identified with each other) if their difference is a given constant on Ω. This approach takes into account the stress of the membrane, but not its displacements as a "rigid" whole. Another approach, which avoids "rigid motions," is to impose an additional integral-type restriction on all functions of the space, e.g.,

$$\int_\Omega u(x,y)\,dx\,dy = 0.$$

Both approaches permit us to use (1.3.2) as a metric on an energy space for a Neumann problem. We shall consider this in more detail later. To do the problem sensibly, we shall need to impose on the forces the balance condition $\int_\Omega f\,dx\,dy = 0$.

A Plate

For a linear elastic plate the potential energy is

$$\mathcal{E}_3(w) = \int_\Omega \frac{D}{2}\left\{(\Delta w)^2 + 2(1-\nu)\left[\left(\frac{\partial^2 w}{\partial x \partial y}\right)^2 - \frac{\partial^2 w}{\partial x^2}\frac{\partial^2 w}{\partial y^2}\right]\right\} dx\,dy \tag{1.3.6}$$

where D is the bending stiffness of the plate, ν is Poisson's ratio, and $w(x,y)$ is the normal displacement of the mid-surface of the plate, which is denoted by Ω in the xy-plane. If the edge of the plate is clamped we get

$$w\Big|_{\partial\Omega} = \frac{\partial w}{\partial n}\Big|_{\partial\Omega} = 0. \tag{1.3.7}$$

If $\mathcal{E}_3(w) = 0$, then $w = a + bx + cy$ and, from (1.3.7), $w = 0$. So D2 is fulfilled by the distance function

$$d(w_1, w_2) = \left(2\mathcal{E}_3(w_1 - w_2)\right)^{1/2}. \tag{1.3.8}$$

The remaining metric axioms are easily checked, and $d(w_1, w_2)$ is a metric on the subset of $C^{(2)}(\Omega)$ consisting of all functions satisfying (1.3.7). This is the energy space for the plate.

If the edge of the plate is free from geometrical fixing (clamping), the situation is similar to the Neumann problem of membrane theory: we must eliminate "rigid" motions of the plate. We shall consider this in detail later.

Linear Elasticity

Consider an elastic body occupying a bounded domain $\Omega \subset \mathbb{R}^3$. The potential energy functional of the body is

$$\mathcal{E}_4(\mathbf{u}) = \frac{1}{2}\int_\Omega c^{ijkl}\epsilon_{kl}\epsilon_{ij}\,d\Omega \tag{1.3.9}$$

where c^{ijkl} is a component of the tensor of elastic moduli; the strain tensor with components (ϵ_{ij}) is defined by

$$\epsilon_{ij} \equiv \epsilon_{ij}(\mathbf{u}) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right), \qquad \mathbf{u} = (u_1, u_2, u_3).$$

Here (x_1, x_2, x_3) are the Cartesian coordinates of a point of Ω. As is the usual case with tensors, the repeated index convention for summation is in effect.

From elasticity theory, the elastic moduli satisfy the following conditions:

(a) The tensor is symmetric, that is

$$c^{ijkl} = c^{klij} = c^{jikl}. \tag{1.3.10}$$

(b) The tensor is positive definite; that is, for any symmetric tensor (ϵ_{ij}) with $\epsilon_{ij} = \epsilon_{ji}$, the inequality

$$c^{ijkl}\epsilon_{kl}\epsilon_{ij} \geq c_0 \sum_{i,j=1}^{3} \epsilon_{ij}^2 \tag{1.3.11}$$

holds with a positive constant c_0 that does not depend on (ϵ_{ij}).

On the set of continuously differentiable vector functions $\mathbf{u}(\mathbf{x})$, representing displacements of points of the body, let us introduce a function

$$d(\mathbf{u}, \mathbf{v}) = \left(2\mathcal{E}_4(\mathbf{u} - \mathbf{v})\right)^{1/2}. \tag{1.3.12}$$

If $d(\mathbf{u}, \mathbf{v}) = 0$ then, from (1.3.11), $\epsilon_{ij}(\mathbf{u} - \mathbf{v}) = 0$ for all $i, j = 1, 2, 3$. As is known from the theory of elasticity,

$$\mathbf{u}(\mathbf{x}) - \mathbf{v}(\mathbf{x}) = \mathbf{a} + \mathbf{x} \times \mathbf{b}$$

where $\mathbf{a}$ and $\mathbf{b}$ are constant vectors. If we restrict the set of vector functions by the boundary condition

$$\mathbf{u}\Big|_{\partial\Omega} = 0 \tag{1.3.13}$$

(i.e., clamp the body edge) then we get $\mathbf{u}(\mathbf{x}) - \mathbf{v}(\mathbf{x}) = 0$. The other metric axioms hold, too. Thus we can impose the metric (1.3.12) on the set of all continuously differentiable vector functions $\mathbf{u}(\mathbf{x})$ satisfying (1.3.13); this is the energy space for the elastic body.

Later we shall consider other boundary conditions of the boundary value problems of the theory of elasticity.

We have not introduced special notation for the energy spaces discussed thus far, since they are not the spaces we shall actually use. They form only the basis of the actual energy spaces; to introduce these, we need the notions of the Lebesgue integral and generalized derivatives, which we shall introduce later.

1.4 Sets in a Metric Space

By analogy with Euclidean space, we may introduce a few concepts.

Definition 1.4.1. In a metric space X the set of points $x \in X$ such that $d(x_0, x) < \tau\ (\leq \tau)$ is called the *open* (*closed*) *ball* of radius τ about x_0.

This definition coincides with the definition of the ball in elementary geometry. However, even in Euclidean space, the use of a metric different from the Euclidean one can give quite different sets as balls. For example, in $\mathbb{R}^3$ with the metric $d(\mathbf{x}, \mathbf{y}) = \sup_i |x_i - y_i|$, a ball about zero $d(\mathbf{0}, \mathbf{x}) < 1$ is a cube having side length 2.

Definition 1.4.2. A subset S of a metric space X is said to be *open* if, together with any of its points x, S contains an open ball of radius $\tau(x)$ about x.

In a metric space we can introduce figures (e.g., ellipses) whose definitions require only a notion of distance. In a concrete metric space, we can introduce some sets using special properties of their elements. For example, in c (see page 11) a cube $\mathcal{C}$ may be defined by

$$\mathcal{C} = \{x = (x_1, x_2, \ldots) \in c : \ |x_k - x_{k0}| \leq a \text{ for each } k\}$$

where $x_0 = (x_{10}, x_{20}, \ldots)$ is a fixed point of c. Note that we call it a "cube" because this definition is similar to the definition of a cube in $\mathbb{R}^3$. However, by Definition 1.4.1, $\mathcal{C}$ is a ball.

Up to now we have not used the notion of linear space and, where possible in this chapter, we shall not exploit it. But the algebraic nature of a linear space X allows us to consider the straight line defined by

$$tx_1 + (1-t)x_2, \qquad x_1, x_2 \in X \tag{1.4.1}$$

where $t \in (-\infty, \infty)$ is a parameter. If we restrict $t \in [0, 1]$, then (1.4.1) yields a *segment* in X.

When necessary, we shall also use the notions of plane, subspace, etc.

Definition 1.4.3. A set in X is called *convex* if together with each pair of its points it contains the segment connecting those points.

Definition 1.4.4. A set in a metric space X is called *bounded* if there is a ball of a finite radius that contains all the elements of the set.

1.5 Convergence in a Metric Space

It is interesting to construct various geometrical figures in a metric space, but we are more interested in properties which, for spaces of functions, are

the usual subjects of calculus. First we introduce the notion of convergence of a sequence in a metric space.

In a metric space X, an infinite sequence $\{x_i\}$ has *limit* x if, for every positive number ε, there exists a number N dependent on ε such that whenever $i > N$ we have $d(x_i, x) < \varepsilon$. (In other words, for any $i > N$, all members of the sequence x_i belong to the ball of radius ε about x.) We write

$$x = \lim_{i \to \infty} x_i.$$

Alternatively, we may say that $x_i \to x$ as $i \to \infty$. We also say that $\{x_i\}$ is *convergent*. This notion generalizes one from calculus, and possesses similar properties:

1. There exists no more than one limit of a convergent sequence. To see this, suppose to the contrary that $\{x_i\}$ has two distinct limits x^1 and x^2. Then $d(x^1, x^2) = a \neq 0$, say. Take $\varepsilon = a/3$; by definition, there exists N such that for all $i \geq N$ we have $d(x_i, x^1) \leq a/3$ and $d(x_i, x^2) \leq a/3$. But $a = d(x^1, x^2) \leq d(x^1, x_i) + d(x_i, x^2) \leq a/3 + a/3 = 2a/3$, a contradiction.

2. A sequence which is convergent in a metric space is bounded.

The ease with which these and similar results are obtained might lead us to try to generalize other classical results — the Bolzano–Weierstrass theorem for example. However, as we mentioned before, many such results do not extend to spaces of infinite dimension.

A sequence $\{x_i\}$ is called a *Cauchy sequence* if for every positive number ε there exists a number N dependent on ε such that whenever $m, n \geq N$ we have $d(x_n, x_m) < \varepsilon$. That this is not in general equivalent to the notion of convergence is shown by the following exercise.

Problem 1.5.1. Construct a sequence of functions continuous on $[0, 1]$ such that the sequence converges to a discontinuous function in a space where the metric is

$$d(f, g) = \int_0^1 |f(x) - g(x)|\, dx. \tag{1.5.1}$$

1.6 Completeness

Definition 1.6.1. A metric space is said to be *complete* if every Cauchy sequence in the space has a limit in the space; otherwise, it is said to be *incomplete*.

The space $\mathbb{R}$ of all real numbers under the metric $d(x, y) = |x - y|$ gives us an example of a complete metric space. Its subset $\mathbb{Q}$ of all rational numbers gives us an example of an incomplete space; there exist Cauchy sequences of rational numbers whose limits are irrational.

Another example of a complete metric space is $C(\Omega)$ when Ω is compact. Its completeness is a consequence of Weierstrass' theorem that the limit of a uniformly convergent sequence of continuous functions on a compact set Ω is continuous on Ω. (The reader should verify that a Cauchy sequence in $C(\Omega)$ is uniformly convergent.)

Problem 1.5.1 shows that, depending on the kind of metric, the same set can underlie a complete or an incomplete metric space. The metric space of all continuous functions on a compact set Ω with the metric (1.5.1) is incomplete.

Definition 1.6.2. An element x of a metric space X is called an *accumulation point* of a set S if any ball centered at x contains a point of S different from x. Next, S is called a *closed set in* X if it contains all its points of accumulation.

It is clear that for x to be an accumulation point of S it suffices to establish the existence of a countable sequence of balls centered at x with radii $\varepsilon_n \to 0$ each of which contains a point of S different from x.

If X is a complete metric space then the definition of an accumulation point x in S states that there is a Cauchy sequence belonging to S, whose elements are different from x, for which x is a limit element. Conversely, if we have a Cauchy sequence belonging to S in a complete metric space then there is a limit point in X. There are only two possibilities for this point:

1. it is an accumulation point of S;

2. it is an isolated point belonging to S.

These facts bring us to another form of Definition 1.6.2 for a complete metric space that we shall use in what follows.

Definition 1.6.2′. A set S in a complete metric space X is called closed if any Cauchy sequence whose elements are in S has a limit belonging to S.

The next theorem is evident.

Theorem 1.6.1. A subset S of a complete metric space X supplied with the metric of X is a complete metric space if and only if S is closed in X.

Definition 1.6.3. A set A is said to be *dense* in a metric space X if for every $x \in X$ any ball of nonzero radius about x contains an element of A.

The Weierstrass theorem states that the set of all polynomials is dense in $C(\Omega)$, where Ω is any compact set in $\mathbb{R}^n$.

The property of completeness is of great importance since numerous passages to the limit are necessary to justify numerical methods, existence theorems, etc. The energy spaces of continuously differentiable functions introduced above are all incomplete. Because these spaces are so convenient in mechanics, we are led to consider the material of the next section.

1.7 The Completion Theorem

Definition 1.7.1. A one-to-one correspondence between metric spaces M_1 and M_2 with metrics d_1 and d_2 respectively is called a *one-to-one isometry* if the correspondence between the elements of these spaces preserves the distances between the elements; that is, if a pair of elements x, y belonging to M_1 corresponds to a pair u, v of M_2, then $d_1(x, y) = d_2(u, v)$.

Theorem 1.7.1. For a metric space M, there is a one-to-one isometry between M and a set $\tilde{M}$ which is dense in a complete metric space M^*; M^* is called the *completion* of M. If M is a linear space, the isometry preserves algebraic operations.

Remark 1.7.1. The elements of M differ in nature from those of $\tilde{M}$. However, in what follows we shall frequently identify them as part of our reasoning process.

Before we can prove the completion theorem, we need

Definition 1.7.2. Two sequences $\{x_n\}$ and $\{y_n\}$ in M are said to be *equivalent* if $d(x_n, y_n) \to 0$ as $n \to \infty$.

Proof of Theorem 1.7.1. The proof is constructive. First we show how to introduce the set M^*, then we verify that it has metric space properties as stated in the theorem.

Let $\{x_n\}$ be a Cauchy sequence in M. Collect all Cauchy sequences in M that are equivalent to $\{x_n\}$ and call the collection an equivalence class X. Any Cauchy sequence from X is called a *representative* of X. To any $x \in M$ there corresponds the equivalence class which contains the stationary sequence $(x, x, x, \ldots)$ and is called the *stationary equivalence class*. Denote all equivalence classes X by M^* and all stationary ones by $\tilde{M}$. Introducing on M^* the metric given by

$$d(X, Y) = \lim_{n \to \infty} d(x_n, y_n) \tag{1.7.1}$$

where $\{x_n\}$ and $\{y_n\}$ are representatives of the equivalence classes X and Y respectively, we obtain the needed M^*, $\tilde{M}$, and the correspondence.

First we must show that (a) (1.7.1) is actually a metric, i.e., it does not depend on the choice of representatives and satisfies the metric axioms, (b) M^* is complete, and (c) $\tilde{M}$ is dense in M^*.

(a) Validity of (1.7.1). Let us first establish that the limit $d(X, Y)$ exists and is independent of choice of representative sequences. From D4 we get

$$d(x_n, y_n) \leq d(x_n, x_m) + d(x_m, y_m) + d(y_m, y_n)$$

so that

$$d(x_n, y_n) - d(x_m, y_m) \leq d(x_n, x_m) + d(y_m, y_n),$$

and, interchanging m and n,

$$d(x_m, y_m) - d(x_n, y_n) \le d(x_m, x_n) + d(y_n, y_m),$$

hence

$$|d(x_n, y_n) - d(x_m, y_m)| \le d(x_n, x_m) + d(y_n, y_m) \to 0$$

as $n, m \to \infty$ because $\{x_n\}$ and $\{y_n\}$ are Cauchy sequences. So $\{d(x_n, y_n\}$ is a Cauchy sequence in $\mathbb{R}$ and the limit in (1.7.1) exists. Similarly, the reader can verify that this limit does not depend on the choice of representatives X, Y. We now verify the metric axioms for (1.7.1):

D1: $d(X, Y) = \lim_{n\to\infty} d(x_n, y_n) \ge 0$.

D2: If $X = Y$ then $d(X, Y) = 0$. Conversely, if $d(X, Y) = 0$ then X and Y contain the same set of equivalent Cauchy sequences.

D3: $d(X, Y) = \lim_{n\to\infty} d(x_n, y_n) = \lim_{n\to\infty} d(y_n, x_n) = d(Y, X)$.

D4: For $x_n, y_n, z_n \in M$,

$$d(x_n, y_n) \le d(x_n, z_n) + d(z_n, y_n).$$

Passage to the limit gives

$$d(X, Y) \le d(X, Z) + d(Z, Y)$$

for the equivalence classes X, Y, Z containing $\{x_n\}$, $\{y_n\}$, $\{z_n\}$, respectively.

(b) Completeness of M^*. Let $\{X^i\}$ be a Cauchy sequence in M^*. We shall show that there exists $X = \lim_{i\to\infty} X^i$. From each of the X^i we first choose a Cauchy sequence $\{x_j^{(i)}\}$ and from this an element denoted x_i such that $d(x_i, x_j^{(i)}) < 1/i$ for all $j > i$. (This is possible since $\{x_j^{(i)}\}$ is a Cauchy sequence.) Let us show that $\{x_i\}$ is a Cauchy sequence. Denote by X_i the equivalence class containing the stationary sequence $(x_i, x_i, \ldots)$. Then

$$\begin{aligned} d(x_i, x_j) &= d(X_i, X_j) \\ &\le d(X_i, X^i) + d(X^i, X^j) + d(X^j, X_j) \\ &\le \frac{1}{i} + d(X^i, X^j) + \frac{1}{j} \to 0 \text{ as } i, j \to \infty. \end{aligned}$$

Now let us denote by X the equivalence class containing the Cauchy sequence $\{x_i\}$. We shall show that $\lim_{i\to\infty} X^i = X$. We have

$$\begin{aligned} d(X^i, X) &\le d(X^i, X_i) + d(X_i, X) \\ &\le \frac{1}{i} + d(X_i, X) \\ &= \frac{1}{i} + \lim_{j\to\infty} d(x_i, x_j) \to 0 \text{ as } i \to \infty \end{aligned} \tag{1.7.2}$$

since $\{x_i\}$ is a Cauchy sequence. This completes the proof of (b).

(c) It is almost obvious that $\tilde{M}$ is dense in M^*. For a class X containing a representative sequence $\{x_n\}$, denoting by X_n the stationary class for the stationary sequence $(x_n, x_n, \ldots)$, we have

$$d(X_n, X) = \lim_{m\to\infty} d(x_n, x_m) \to 0 \text{ as } n \to \infty$$

since $\{x_n\}$ is a Cauchy sequence.

Finally, the equality $d(X, Y) = d(x, y)$ if X and Y are stationary classes corresponding to x and y, respectively, gives the one-to-one isometry between M and $\tilde{M}$. The preservation of algebraic operations in M is obvious, and this completes the proof of Theorem 1.7.1. □

It is worth noting what happens if M is complete. It is clear that we can determine a one-to-one correspondence between any equivalence class and the only element which is the limit of a representative sequence of this class. Thus we can identify a complete metric space with its completion.

Because Theorem 1.7.1 is of great importance to us, let us review its main points: M^* is a metric space whose elements are classes of all equivalent Cauchy sequences from M; M is isometrically identified with $\tilde{M}$, which is the set of all stationary equivalence classes; $\tilde{M}$ is dense in M^*.

We can sometimes establish a property of a limit of a representative sequence of X that does not depend on the particular choice of representative. In that case we shall say that the class X possesses this property. This is typical for energy and Sobolev spaces; the formulation of such properties is the basis of so-called imbedding theorems.

The following sections will provide examples of the application of Theorem 1.7.1.

1.8 The Lebesgue Integral and the Space $L^p(\Omega)$

By arguments similar to those we gave in Section 1.6, the set of all functions which are continuous on a closed and bounded domain $\Omega \subset \mathbb{R}^n$ with metric

$$d(f(\mathbf{x}), g(\mathbf{x})) = \left(\int_\Omega |f(\mathbf{x}) - g(\mathbf{x})|^p \, d\Omega \right)^{1/p}, \qquad p \geq 1, \tag{1.8.1}$$

is an incomplete metric space.

Let us apply Theorem 1.7.1 to this case. The corresponding space of equivalence classes is denoted by $L^p(\Omega)$. (In case $p = 1$ we usually omit the superscript and write $L(\Omega)$ instead.) An element of $L^p(\Omega)$ is the set of all Cauchy sequences of functions, continuous on Ω, that are equivalent to one another. Here $\{f_n(\mathbf{x})\}$ is a Cauchy sequence if

$$\int_\Omega |f_n(\mathbf{x}) - f_m(\mathbf{x})|^p \, d\Omega \to 0 \text{ as } n, m \to \infty$$

and two sequences $\{f_n(\mathbf{x})\}$ and $\{g_n(\mathbf{x})\}$ are equivalent if

$$\int_\Omega |f_n(\mathbf{x}) - g_n(\mathbf{x})|^p \, d\Omega \to 0 \text{ as } n \to \infty.$$

Remark 1.8.1. In the classical theory of functions of a real variable, it is shown that for any equivalence class in $L^p(\Omega)$ there is a function (or, more precisely, a class of equivalent functions) which is a limit, in a certain sense, of a representative sequence of the class; for this function, the so-called Lebesgue integral is introduced. Our constructions of $L^p(\Omega)$ and the Lebesgue integral are equivalent to those of the classical theory. In view of this, we shall sometimes refer to an equivalence class of $L^p(\Omega)$ as a "function."

Remark 1.8.2. In accordance with Weierstrass' theorem, any function continuous on Ω can be approximated by a polynomial with any accuracy in the metric of $C(\Omega)$, and hence in that of $L^p(\Omega)$. An interpretation is that any equivalence class of $L^p(\Omega)$ contains a Cauchy sequence whose elements are infinitely differentiable functions (moreover, polynomials), and we may thus obtain $L^p(\Omega)$ on the basis of only this subset of $C(\Omega)$.

Remark 1.8.3. In (1.8.1) we use Riemann integration. We must therefore exclude some "exotic" domains Ω which are allowed in the classical theory of Lebesgue integration. It is possible to extend the present approach to achieve the same degree of generality, but the applications we consider do not necessitate this. We therefore leave it to the reader to bridge this gap if he/she wishes to do so. We also remark that Ω need not be bounded in order to construct the theory.

The Lebesgue Integral

An element of $L^p(\Omega)$ (an equivalence class) is denoted by $F(\mathbf{x})$. To construct the Lebesgue integral, we use the Riemann integral. We first consider how to define $\int_\Omega |F(\mathbf{x})|^p \, d\Omega$ when $F(\mathbf{x}) \in L^p(\Omega)$. We take a representative Cauchy sequence $\{f_n(\mathbf{x})\}$ from $F(\mathbf{x})$ and consider the sequence $\{K_n\}$ given by

$$K_n = \left(\int_\Omega |f_n(\mathbf{x})|^p \, d\Omega \right)^{1/p}.$$

This is a Cauchy sequence of numbers: we have

$$\begin{aligned} |K_n - K_m| &= \left| \left(\int_\Omega |f_n(\mathbf{x})|^p \, d\Omega \right)^{1/p} - \left(\int_\Omega |f_m(\mathbf{x})|^p \, d\Omega \right)^{1/p} \right| \\ &\leq \left(\int_\Omega |f_n(\mathbf{x}) - f_m(\mathbf{x})|^p \, d\Omega \right)^{1/p} \to 0 \text{ as } m, n \to \infty, \end{aligned}$$

as a consequence of the inequality

$$\left(\int_\Omega |f-g+g|^p\,d\Omega\right)^{1/p} \le \left(\int_\Omega |f-g|^p\,d\Omega\right)^{1/p} + \left(\int_\Omega |g|^p\,d\Omega\right)^{1/p}$$

(which follows from the Minkowski inequality) and a similar inequality obtained by interchanging the roles of f and g. So there exists

$$K = \lim_{n\to\infty} K_n = \lim_{n\to\infty} \left(\int_\Omega |f_n(\mathbf{x})|^p\,d\Omega\right)^{1/p}.$$

To complete the construction, we must show that K is independent of the choice of representative sequence. We leave this to the reader as an easy application of Minkowski's inequality. The number K^p is called the Lebesgue integral of $|F(\mathbf{x})|^p$:

$$K^p \equiv \int_\Omega |F(\mathbf{x})|^p\,d\Omega = \lim_{n\to\infty} \int_\Omega |f_n(\mathbf{x})|^p\,d\Omega.$$

Let $F(\mathbf{x}) \in L^p(\Omega)$ where Ω is a bounded domain. Let us show that $F(\mathbf{x}) \in L^r(\Omega)$ whenever $1 \le r \le p$. By Hölder's inequality we have

$$\begin{aligned}\left|\int_\Omega 1\cdot|f(\mathbf{x})|^r\,d\Omega\right| &\le \left(\int_\Omega 1^q\,d\Omega\right)^{1/q}\left(\int_\Omega (|f(\mathbf{x})|^r)^{p/r}\,d\Omega\right)^{r/p}\\ &= (\operatorname{mes}\Omega)^{1/q}\left(\int_\Omega |f(\mathbf{x})|^p\,d\Omega\right)^{r/p}\end{aligned}$$

if $1/q + r/p = 1$. For any r such that $1 \le r \le p$, it follows that

$$\begin{aligned}&\left(\int_\Omega |f_n(\mathbf{x}) - f_m(\mathbf{x})|^r\,d\Omega\right)^{1/r} \le\\ &\qquad (\operatorname{mes}\Omega)^{1/r-1/p}\left(\int_\Omega |f_n(\mathbf{x}) - f_m(\mathbf{x})|^p\,d\Omega\right)^{1/p}.\end{aligned}$$

This means that a sequence of functions which is a Cauchy sequence in the metric (1.8.1) of $L^p(\Omega)$ is also a Cauchy sequence in the metric of $L^r(\Omega)$ whenever $1 \le r < p$. In similar fashion we can show that any two sequences equivalent in $L^p(\Omega)$ are also equivalent in $L^r(\Omega)$. Hence any element of $L^p(\Omega)$ also belongs to $L^r(\Omega)$ if $1 \le r < p$, and we can say that $L^p(\Omega)$ is a subset of $L^r(\Omega)$. Thus we can determine an integral

$$\int_\Omega |F(\mathbf{x})|^r\,d\Omega$$

for $1 \le r < p$. Moreover, passage to the limit shows that

$$\left(\int_\Omega |F(\mathbf{x})|^r\,d\Omega\right)^{1/r} \le (\operatorname{mes}\Omega)^{1/r-1/p}\left(\int_\Omega |F(\mathbf{x})|^p\,d\Omega\right)^{1/p}. \qquad (1.8.2)$$

Now we can introduce the Lebesgue integral for an element $F(\mathbf{x}) \in L^p(\Omega)$, $p \geq 1$. Take a representative sequence $\{f_n(\mathbf{x})\}$ of the class $F(\mathbf{x})$. That the sequence of numbers $\{\int_\Omega f_n(\mathbf{x})\, d\Omega\}$ is a Cauchy sequence follows from the inequality

$$\left| \int_\Omega f(\mathbf{x})\, d\Omega \right| \leq \int_\Omega |f(\mathbf{x})|\, d\Omega.$$

So the quantity

$$\int_\Omega F(\mathbf{x})\, d\Omega = \lim_{n\to\infty} \int_\Omega f_n(\mathbf{x})\, d\Omega$$

is uniquely determined for $F(\mathbf{x})$ and is called the *Lebesgue integral* of $F(\mathbf{x}) \in L^p(\Omega)$ over Ω. Note that for the Lebesgue integral we have

$$\left| \int_\Omega F(\mathbf{x})\, d\Omega \right| \leq (\operatorname{mes}\Omega)^{1/q} \left(\int_\Omega |F(\mathbf{x})|^p\, d\Omega \right)^{1/p} \tag{1.8.3}$$

if $1/q + 1/p = 1$.

In what follows, we shall frequently encounter integrals of the form

$$\int_\Omega F(\mathbf{x}) G(\mathbf{x})\, d\Omega.$$

For example, the work of external forces is of this form. Let us determine this integral when $F(\mathbf{x}) \in L^p(\Omega)$ and $G(\mathbf{x}) \in L^q(\Omega)$ where $1/p + 1/q = 1$. Consider

$$I_n = \int_\Omega f_n(\mathbf{x}) g_n(\mathbf{x})\, d\Omega$$

where $\{f_n(\mathbf{x})\}$ and $\{g_n(\mathbf{x})\}$ are representative sequences of $F(\mathbf{x})$ and $G(\mathbf{x})$, respectively. Then

$$\begin{aligned}
|I_n - I_m| &= \left| \int_\Omega [f_n(\mathbf{x}) g_n(\mathbf{x}) - f_m(\mathbf{x}) g_m(\mathbf{x})]\, d\Omega \right| \\
&= \left| \int_\Omega [(f_n - f_m) g_n + f_m (g_n - g_m)]\, d\Omega \right| \\
&\leq \int_\Omega |f_n - f_m|\, |g_n|\, d\Omega + \int_\Omega |f_m|\, |g_n - g_m|\, d\Omega \\
&\leq \left(\int_\Omega |f_n - f_m|^p\, d\Omega \right)^{1/p} \left(\int_\Omega |g_n|^q\, d\Omega \right)^{1/q} \\
&\quad + \left(\int_\Omega |f_m|^p\, d\Omega \right)^{1/p} \left(\int_\Omega |g_n - g_m|^q\, d\Omega \right)^{1/q} \to 0 \text{ as } n, m \to \infty
\end{aligned}$$

since $\{f_n(\mathbf{x})\}$ and $\{g_n(\mathbf{x})\}$ are Cauchy sequences in their respective metrics and, for large n,

$$\int_\Omega |f_n(\mathbf{x})|^p \, d\Omega \le \int_\Omega |F(\mathbf{x})|^p \, d\Omega + 1,$$
$$\int_\Omega |g_n(\mathbf{x})|^q \, d\Omega \le \int_\Omega |G(\mathbf{x})|^q \, d\Omega + 1.$$

So there exists $I = \lim_{n\to\infty} I_n$, which we call the Lebesgue integral

$$I = \int_\Omega F(\mathbf{x})G(\mathbf{x}) \, d\Omega.$$

(Convince yourself that it is independent of the choice of representatives, hence is well defined.)

Passage to the limit in

$$\left| \int_\Omega f_n(\mathbf{x}) g_n(\mathbf{x}) \, d\Omega \right| \le \left(\int_\Omega |f_n(\mathbf{x})|^p \, d\Omega \right)^{1/p} \left(\int_\Omega |g_n(\mathbf{x})|^q \, d\Omega \right)^{1/q}$$

shows that Hölder's inequality

$$\left| \int_\Omega F(\mathbf{x})G(\mathbf{x}) \, d\Omega \right| \le \left(\int_\Omega |F(\mathbf{x})|^p \, d\Omega \right)^{1/p} \left(\int_\Omega |G(\mathbf{x})|^q \, d\Omega \right)^{1/q} \tag{1.8.4}$$

holds for $F(\mathbf{x}) \in L^p(\Omega)$, $G(\mathbf{x}) \in L^q(\Omega)$, whenever $1/p + 1/q = 1$. Equality holds in Hölder inequality if and only if $F(\mathbf{x}) = \lambda G(\mathbf{x})$ for some number λ.

Remark 1.8.4. If Ω is unbounded, Hölder's inequality still holds; however, in this case it is not true in general that $L^p(\Omega)$ is a subset of $L^r(\Omega)$ for all $r < p$.

We conclude this section by asserting that the properties of the classes in $L^p(\Omega)$ introduced above permit us to deal with the Lebesgue integral as if its integrand were an ordinary function.

1.9 Banach and Hilbert Spaces

Most of the metric spaces we have considered have also been linear spaces. This implies that each pair $x, y \in X$ has a uniquely defined sum $x + y$ such that

1. $x + y = y + x$,
2. $x + (y + z) = (x + y) + z$, and
3. there is a zero element $\theta \in X$ such that $x + \theta = x$;

moreover, it implies that each $x \in X$ has a uniquely defined product by a real (or complex) scalar $\lambda \in \mathbb{R}$ (or $\mathbb{C}$) such that

4. $\lambda(\mu x) = (\lambda\mu)x$,
5. $1x = x$, $0x = \theta$,
6. $x + (-1x) = \theta$,
7. $\lambda(x + y) = \lambda x + \lambda y$,
8. $(\lambda + \mu)x = \lambda x + \mu x$.

In what follows, we shall denote the zero element of X by 0 instead of θ.

If multiplication by scalars is introduced as multiplication by purely real numbers, the space will be called a *real linear space*; if the scalars are in general complex numbers, the space will be called a *complex linear space.*

We could continue to consider the general properties of metric spaces, but all spaces of interest to us have some special and very convenient properties — namely, their metrics take a special form which can be denoted by

$$d(x, y) = \|x - y\|. \tag{1.9.1}$$

Definition 1.9.1. $\|x\|$ is called a *norm* on a linear space X if it is a real-valued function defined for every $x \in X$ and satisfies the following norm axioms:

N1. $\|x\| \geq 0$, and $\|x\| = 0$ if and only if $x = 0$;

N2. $\|\lambda x\| = |\lambda| \, \|x\|$ for any real (or complex) λ;

N3. $\|x + y\| \leq \|x\| + \|y\|$.

The reader should verify that any metric defined by (1.9.1) satisfies D1–D4, provided that $\|x\|$ satisfies N1–N3.

Definition 1.9.2. A linear space X is called a *normed space* if, for every $x \in X$, a norm of x satisfying N1–N3 is defined. A normed space X is said to be *real* (*complex*) if the scalars λ in the product λx are taken from $\mathbb{R}$ ($\mathbb{C}$).

A normed space is a metric space, but the converse is false; there are linear metric spaces which are not normed. For a normed space, we shall use the terminology of the corresponding metric space. We shall subsequently follow this practice for inner product spaces as well.

Problem 1.9.1. Two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ defined on X are said to be *equivalent* if there exist positive real numbers M and m such that for all $x \in X$,

$$m\|x\|_1 \leq \|x\|_2 \leq M\|x\|_1. \tag{1.9.2}$$

Show that on $\mathbb{R}^n$, all norms are equivalent.

Problem 1.9.2. Show that if x and y are elements of a real normed space X, then

$$\|x - y\| \geq |\,\|x\| - \|y\|\,|.$$

Definition 1.9.3. A complete normed space is called a *Banach space*.

Several of the spaces we examined previously are Banach spaces. $C(\Omega)$ with compact Ω is a linear space, and is clearly a normed space if we set

$$\|f(\mathbf{x})\| = \max_{\mathbf{x}\in\Omega} |f(\mathbf{x})|.$$

Because $C(\Omega)$ is complete as a metric space, it is a Banach space. We leave it to the reader to show that $L^p(\Omega)$ is a Banach space under the norm

$$\|F(\mathbf{x})\| = \left(\int_\Omega |F(\mathbf{x})|^p \, d\Omega\right)^{1/p}, \qquad p \geq 1.$$

Other familiar Banach spaces are c, m, and ℓ^p.

Let us consider a new example of a Banach space: $C^{(k)}(\Omega)$, where $\Omega \subset \mathbb{R}^n$ is a closed and bounded domain. This space consists of those functions that are defined and continuous on Ω and such that all their derivatives up to order k are continuous on Ω. The norm on $C^{(k)}(\Omega)$ is defined by

$$\|f(\mathbf{x})\| = \max_{\mathbf{x}\in\Omega} |f(\mathbf{x})| + \sum_{|\alpha|\leq k} \max_{\mathbf{x}\in\Omega} |D^\alpha f(\mathbf{x})|.$$

The reader should supply the routine but necessary steps to verify that N1–N3 are satisfied; we proceed to show that the resulting space is complete.

Let $\{f_i(\mathbf{x})\}$ be a Cauchy sequence in $C^{(k)}(\Omega)$. This implies that the sequence $\{f_i(\mathbf{x})\}$ as well as all the sequences $\{D^\alpha f_i(\mathbf{x})\}$ when $|\alpha| \leq k$ are Cauchy sequences in $C(\Omega)$. Being uniformly convergent on Ω, each of these sequences has a limit function:

$$\lim_{i\to\infty} f_i(\mathbf{x}) = f(\mathbf{x}), \qquad \lim_{i\to\infty} D^\alpha f_i(\mathbf{x}) = f^\alpha(\mathbf{x}), \quad |\alpha| \leq k,$$

where $f(\mathbf{x})$ and $f^\alpha(\mathbf{x})$ for each $|\alpha| \leq k$ are continuous. To complete the verification we must show that

$$D^\alpha f(\mathbf{x}) = f^\alpha(\mathbf{x}).$$

We check this only for $\partial f/\partial x_1$; for the other derivatives it can be done in a similar way. So let

$$\lim_{i\to\infty} \frac{\partial f_i(\mathbf{x})}{\partial x_1} = f^1(\mathbf{x}) = f^1(x_1, x_2, \ldots, x_n).$$

Consider

$$\Delta = f(x_1, x_2, \ldots, x_n) - f(a, x_2, \ldots, x_n) - \int_a^{x_1} f^1(t, x_2, \ldots, x_n)\, dt.$$

We have

$$\begin{aligned}\Delta = &[f(x_1,\dots,x_n) - f_i(x_1,\dots,x_n)] - \\ &- [f(a,x_2,\dots,x_n) - f_i(a,x_2,\dots,x_n)] - \\ &- \left[\int_a^{x_1}\left(f^1(t,x_2,\dots,x_n) - \frac{\partial f_i(t,x_2,\dots,x_n)}{\partial t}\right)dt\right].\end{aligned}$$

Each of the terms in square brackets tends to zero uniformly as $i \to \infty$, so $\Delta = 0$ since Δ does not depend on i, i.e.,

$$f(x_1,x_2,\dots,x_n) - f(a,x_2,\dots,x_n) = \int_a^{x_1} f^1(t,x_2,\dots,x_n)\,dt.$$

Thus

$$\frac{\partial f(\mathbf{x})}{\partial x_1} = f^1(\mathbf{x}).$$

Another example of a Banach space is the Hölder space $H^{k,\lambda}(\Omega)$, $0 < \lambda \le 1$, which consists of those functions of $C^{(k)}(\Omega)$ whose norms in $H^{k,\lambda}(\Omega)$, defined by

$$\|f\| = \sum_{0\le|\alpha|\le k}\max_{\mathbf{x}\in\Omega}|D^\alpha f(\mathbf{x})| + \sum_{|\alpha|=k}\sup_{\substack{\mathbf{x},\mathbf{y}\in\Omega\\ \mathbf{x}\ne\mathbf{y}}}\frac{|D^\alpha f(\mathbf{x}) - D^\alpha f(\mathbf{y})|}{|\mathbf{x}-\mathbf{y}|^\lambda},$$

are finite.

Just as the notion of distance can be widely extended, so can the notion of the vector dot product:

Definition 1.9.4. Let H be a linear space over $\mathbb{C}$. A function (x,y) defined uniquely for each pair $x, y \in H$ is called an *inner product* on H if it satisfies the following axioms:

P1. $(x,x) \ge 0$, and $(x,x) = 0$ if and only if $x = 0$;

P2. $(x,y) = \overline{(y,x)}$;

P3. $(\lambda x + \mu y, z) = \lambda(x,z) + \mu(y,z)$ whenever $\lambda, \mu \in \mathbb{C}$.

The space H, taken together with an inner product, is called a (complex) *inner product space*.

We can consider H over $\mathbb{R}$; then the inner product is real-valued, P2 is replaced by

P2′. $(x,y) = (y,x)$,

and H is called a real inner product space. If it is clear from the context, the designation "real" or "complex" shall be omitted.

Let us consider some properties of H. We introduce $\|x\|$ using

$$\|x\| = (x,x)^{1/2}.$$

To show that we really have a norm, we prove the *Schwarz inequality*:

Theorem 1.9.1. For any $x, y \in H$, the inequality

$$|(x, y)| \leq \|x\| \, \|y\| \tag{1.9.3}$$

holds. For $x, y \neq 0$, equality holds if and only if $x = \lambda y$.

Proof. If either x or y is zero, there is nothing to show. Let $y \neq 0$ and let λ be a scalar. By P1, $(x + \lambda y, x + \lambda y) \geq 0$. We have

$$(x + \lambda y, x + \lambda y) = (x, x) + \lambda(y, x) + \overline{\lambda}(x, y) + \lambda\overline{\lambda}(y, y) \equiv A(\lambda).$$

Put $\lambda_0 = -(x, y)/(y, y)$; then

$$A(\lambda_0) = \|x\|^2 - 2\frac{|(x, y)|^2}{\|y\|^2} + \frac{|(x, y)|^2\|y\|^2}{\|y\|^4} \geq 0.$$

Inequality (1.9.3) follows directly. □

Now we can verify that $\|x\|$ satisfies N1–N3. N1 is satisfied by virtue of P1; N2 is satisfied because

$$\|\lambda x\| = (\lambda x, \lambda x)^{1/2} = (\lambda\overline{\lambda})^{1/2}(x, x)^{1/2} = |\lambda| \, \|x\|;$$

N3 is satisfied because

$$\begin{aligned} \|x + y\|^2 &= (x + y, x + y) \\ &= (x, x) + (x, y) + (y, x) + (y, y) \\ &\leq \|x\|^2 + \|x\| \, \|y\| + \|x\| \, \|y\| + \|y\|^2 \\ &= (\|x\| + \|y\|)^2. \end{aligned}$$

We have shown that an inner product space is a normed space.

Definition 1.9.5. A complete inner product space is called a *Hilbert space.*

By analogy with Euclidean space, we shall say that x is *orthogonal to* y if $(x, y) = 0$.

Problem 1.9.3. Show that for all x, y in an inner product space, the *parallelogram equality*

$$\|x + y\|^2 + \|x - y\|^2 = 2(\|x\|^2 + \|y\|^2) \tag{1.9.4}$$

holds.

Let us consider some examples of Hilbert spaces.

1. *The space* ℓ^2. For $\mathbf{x}, \mathbf{y} \in \ell^2$, an inner product is defined by

$$(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{\infty} x_k\overline{y_k}. \tag{1.9.5}$$

The space ℓ^2 was the prototype for all Hilbert spaces, and as such it was instrumental in the development of functional analysis. It was introduced by Hilbert in a paper devoted to the justification of the Dirichlet principle. In ℓ^2 over $\mathbb{R}$, the inner product is given by

$$(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{\infty} x_k y_k. \tag{1.9.6}$$

2. *The space* $L^2(\Omega)$. Here the inner product is

$$(f(\mathbf{x}), g(\mathbf{x})) = \int_{\Omega} f(\mathbf{x})\overline{g(\mathbf{x})}\, d\Omega. \tag{1.9.7}$$

The axioms P1–P3 are readily verified for both of these spaces. The reader should take a moment to write down the Schwarz inequality in both cases, and to write down the inner product for $L^2(\Omega)$ over the reals.

Most importantly, the energy spaces we introduced earlier are all inner product spaces.

1.10 Some Energy Spaces

A Bar

Earlier we noted that the set S of all continuous functions $y(x)$ having continuous first and second derivatives on $[0, l]$ and which satisfy the boundary conditions

$$y(0) = y'(0) = y(l) = y'(l) = 0 \tag{1.10.1}$$

under the metric

$$d(y, z) = \left(\frac{1}{2} \int_0^l B(x)[y''(x) - z''(x)]^2\, dx \right)^{1/2} \tag{1.10.2}$$

is a metric space. We called this an energy space for the clamped bar. We can now introduce an inner product

$$(y, z) = \frac{1}{2} \int_0^l B(x) y''(x) z''(x)\, dx \tag{1.10.3}$$

and norm

$$\|y\| = \left(\frac{1}{2} \int_0^l B(x)[y''(x)]^2\, dx \right)^{1/2}$$

on this space. We have $d(y, z) = \|y - z\|$. But this space is not complete (it is clear that there are Cauchy sequences whose limits do not belong to

$C^{(2)}(0,l)$; the reader should construct an example). To have a complete space, we must apply the completion theorem. The real energy space denoted by E_B is the completion of S in the metric (1.10.2) (or, what amounts to the same thing, in the inner product (1.10.3)).

Let us consider some properties of the elements of E_B. An element of E_B is a set of Cauchy sequences equivalent in the metric (1.10.2). If we assume that

$$0 < m_1 \le B(x) \le m_2$$

then the sequence of second derivatives $\{y_n''(x)\}$ of a representative sequence is a Cauchy sequence in $L^2(0,l)$ since

$$m_1 \int_0^l [y_n''(x) - y_m''(x)]^2 \, dx \le \int_0^l B(x)[y_n''(x) - y_m''(x)]^2 \, dx.$$

So we can say that $\{y_n''(x)\}$ belongs to $L^2(0,l)$.

Now consider $\{y_n'(x)\}$. For any $y(x) \in S$ we get

$$y'(x) = \int_0^x y''(t) \, dt.$$

So for a representative $\{y_n(x)\}$ of a class $y(x) \in E_B$ we have

$$\begin{aligned} |y_n'(x) - y_m'(x)| &\le \int_0^x |y''(t) - y_m''(t)| \, dt \le \int_0^l 1 \cdot |y_n''(t) - y_m''(t)| \, dt \\ &\le l^{1/2} \left(\int_0^l [y_n''(x) - y_m''(x)]^2 \, dx \right)^{1/2} \\ &\le (l/m_1)^{1/2} \left(\int_0^l B(x)[y_n''(x) - y_m''(x)]^2 \, dx \right)^{1/2} \\ &\to 0 \text{ as } n, m \to \infty. \end{aligned} \tag{1.10.4}$$

It follows that $\{y_n'(x)\}$ converges uniformly on $[0,l]$; hence there exists a limit function $z(x)$ which is also continuous on $[0,l]$. This function does not depend on the choice of representative sequence (verify). The same holds for a sequence of functions $\{y_n(x)\}$: its limit is a function $y(x)$ continuous on $[0,l]$. Moreover,

$$y'(x) = z(x).$$

To prove this, it is necessary to repeat the arguments of Section 1.9 on the differentiability of the elements of $C^{(k)}(\Omega)$, with due regard for (1.10.4). From (1.10.4) and the similar inequality for $\{y_n(x)\}$ we get

$$\max_{x \in \Omega} (|y(x)| + |y'(x)|) \le m \left(\frac{1}{2} \int_0^l B(x)[y''(x)]^2 \, dx \right)^{1/2} \tag{1.10.5}$$

for some constant m independent of $y(x) \in E_B$. So each element $y(x) \in E_B$ can be identified with an element $y(x) \in C^{(1)}(0,l)$ in such a way that (1.10.5) is fulfilled. This correspondence is called an imbedding operator. In what follows, we shall interpret (1.10.5) as a statement that the imbedding operator from E_B to $C^{(1)}(0,l)$ is continuous (we shall see why in Section 1.12). From now on we refer to the elements of E_B as if they were continuously differentiable functions, attaching the properties of the uniquely determined limit functions to the corresponding elements of E_B themselves.

We are interested in analysis of all the terms that are included into the statement of the problem of equilibrium of a body as the problem of minimum of potential energy. So we will consider the functional of the work of external forces. For the bar it is

$$A = \int_0^l F(x)y(x)\,dx.$$

This is well defined on E_B if $F(x) \in L(\Omega)$; moreover, (1.10.5) shows that it can contain terms of the form

$$\sum_k F_k y(x_k) + M_k y'(x_k),$$

which can be interpreted as the work of point forces and point moments, respectively. This is a consequence of the continuity of the imbedding operator.

Remark 1.10.1. Alternatively we can define E_B on a base set S_1 of smoother functions, $C^{(4)}(0,l)$ say, satisfying (1.10.1). The result is the same since S_1 is dense in S with respect to the norm of $C^{(2)}(0,l)$. Sometimes such a definition is convenient.

Remark 1.10.2. Those readers familiar with the contemporary literature in this area may have noticed that Western authors usually deal with Sobolev spaces, studying the properties of operators corresponding to problems under consideration; we prefer instead to deal with energy spaces, studying first their properties and then those of the corresponding operators. Although these approaches lead to the same results, in our view the mechanics (physics) of a particular problem should play a principal role in the analysis — in this way the methodology seems simpler, clearer, and more natural. Why is it that in papers devoted to the study of elastic bodies we mainly find investigation of the case of a clamped boundary? Sometimes this is done on the principle that it is better to deal with homogeneous Dirichlet boundary conditions only, but more often it is an unfortunate consequence of the use of the Sobolev spaces $H^k(\Omega)$. The theory of these spaces is well developed, but is not amenable to the study of other boundary conditions. Indeed, success in the investigation of mechanics problems can be much more difficult without the benefit of the physical ideas that are brought out by the energy spaces.

Remark 1.10.3. In defining the energy space of the bar, we left aside the question of smoothness of the stiffness function $B(x)$. From a mathematical standpoint this is risky since, in principle, $B(x)$ can be nonintegrable. But in the case of an actual physical bar, $B(x)$ can have no more than a finite number of discontinuities and must be differentiable everywhere else. For simplicity, we shall continue to make realistic assumptions concerning physical parameters such as stiffness and elastic constants; in particular, we shall suppose whatever degree of smoothness as may be required for our purposes.

A Membrane (Clamped Edge)

The subset of $C^{(1)}(\Omega)$ consisting of all functions satisfying

$$u(x,y)\Big|_{\partial\Omega} = 0 \tag{1.10.6}$$

with the metric

$$d(u,v) = \left\{ \iint\limits_{\Omega} \left[\left(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} - \frac{\partial v}{\partial y} \right)^2 \right] dx\, dy \right\}^{1/2} \tag{1.10.7}$$

is an incomplete metric space. If we introduce an inner product

$$(u,v) = \iint\limits_{\Omega} \left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y} \right) dx\, dy \tag{1.10.8}$$

consistent with (1.10.7), we get an inner product space. Its completion in the metric (1.10.7) is the energy space for the clamped membrane, and is denoted E_{MC}. This is a real Hilbert space.

What can we tell about the elements of E_{MC}? It is obvious that the sequences of first derivatives $\{\partial u_n/\partial x\}$, $\{\partial u_n/\partial y\}$, of a representative sequence $\{u_n\}$ are Cauchy sequences in the norm on $L^2(\Omega)$. What about functions? If we extend each $u_n(x,y)$ by zero outside Ω, we can write

$$u_n(x,y) = \int_0^x \frac{\partial u_n(s,y)}{\partial s}\, ds$$

(assuming, without loss of generality, that Ω is confined to the band $0 \leq x \leq a$). Squaring both sides and integrating over Ω, we get

$$\begin{aligned}\iint_\Omega u_n^2(x,y)\,dx\,dy &= \iint_\Omega \left(\int_0^x \frac{\partial u_n(s,y)}{\partial s}\,ds\right) dx\,dy \\ &\leq \iint_\Omega \left(\int_0^a 1\cdot\left|\frac{\partial u_n(s,y)}{\partial s}\right| ds\right)^2 dx\,dy \\ &\leq \iint_\Omega a\left(\int_0^a \left(\frac{\partial u_n(s,y)}{\partial s}\right)^2 ds\right) dx\,dy \\ &\leq a^2 \iint_\Omega \left(\frac{\partial u_n(x,y)}{\partial x}\right)^2 dx\,dy.\end{aligned}$$

This means that if $\{\partial u_n/\partial x\}$ is a Cauchy sequence in the metric of $L^2(\Omega)$ then $\{u_n\}$ is also a Cauchy sequence in this metric. So we can consider elements $U(x,y)$ of E_{MC} to be such that $U(x,y)$, $\partial U/\partial x$, and $\partial U/\partial y$ belong to $L^2(\Omega)$. In the next section, we shall see how to interpret derivatives of $U(x,y)$.

As a consequence of the last chain of inequalities, we get *Friedrichs's inequality*

$$\iint_\Omega U^2(x,y)\,dx\,dy \leq m \iint_\Omega \left[\left(\frac{\partial U}{\partial x}\right)^2 + \left(\frac{\partial U}{\partial y}\right)^2\right] dx\,dy$$

which holds for any $U(x,y) \in E_{MC}$ and a constant m independent of $U(x,y)$.

A Membrane (Free Edge)

Although it is natural to introduce the energy space using the energy metric (1.10.7), we cannot distinguish between two states $u_1(x,y)$ and $u_2(x,y)$ of the membrane with free edge if

$$u_2(x,y) - u_1(x,y) = c = \text{constant}.$$

This difference is the so-called "rigid" motion of the membrane, the only form of rigid motion possible in this theory. We first show that no other rigid motions are possible. The proof is a consequence of *Poincaré's inequality*

$$\iint_\Omega u^2\,dx\,dy \leq m\left\{\left(\iint_\Omega u\,dx\,dy\right)^2 + \iint_\Omega \left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] dx\,dy\right\} \tag{1.10.9}$$

for a function $u(x,y) \in C^{(1)}(\Omega)$. The constant m does not depend on $u(x,y)$. Our proof follows Courant and Hilbert [8].

We first assume that Ω is the square $[0,a] \times [0,a]$, and begin with the identity

$$u(x_2,y_2) - u(x_1,y_1) = \int_{x_1}^{x_2} \frac{\partial u(s,y_1)}{\partial s}\, ds + \int_{y_1}^{y_2} \frac{\partial u(x_2,t)}{\partial t}\, dt. \qquad (1.10.10)$$

(We will use the Poincaré inequality for general domains. Note that a modification of (1.10.9) for more general domains will be established in Section 1.26.) Squaring both sides and then integrating over the square, first with respect to the variables x_1, y_1 and then with respect to x_2, y_2, we get

$$\begin{aligned}
&\iint_\Omega \iint_\Omega \left[u^2(x_2,y_2) - 2u(x_2,y_2)u(x_1,y_1) + u^2(x_1,y_1)\right]\, dx_1\, dy_1\, dx_2\, dy_2 \\
&= \iint_\Omega \iint_\Omega \left[\int_{x_1}^{x_2} \frac{\partial u(s,y_1)}{\partial s}\, ds + \int_{y_1}^{y_2} \frac{\partial u(x_2,t)}{\partial t}\, dt\right]^2 dx_1\, dy_1\, dx_2\, dy_2 \\
&\le \iint_\Omega \iint_\Omega \left[\int_0^a 1 \cdot \left|\frac{\partial u(s,y_1)}{\partial s}\right| ds + \int_0^a 1 \cdot \left|\frac{\partial u(x_2,t)}{\partial t}\right| dt\right]^2 dx_1\, dy_1\, dx_2\, dy_2 \\
&\le 2a \iint_\Omega \iint_\Omega \left[\int_0^a \left(\frac{\partial u(s,y_1)}{\partial s}\right)^2 ds + \int_0^a \left(\frac{\partial u(x_2,t)}{\partial t}\right)^2 dt\right] dx_1\, dy_1\, dx_2\, dy_2 \\
&\le 2a^4 \int_0^a \int_0^a \left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] dx\, dy.
\end{aligned}$$

The beginning of this chain of inequalities is

$$a^2 \iint_\Omega u^2(x,y)\, dx\, dy - 2\left(\iint_\Omega u(x,y)\, dx\, dy\right)^2 + a^2\left(\iint_\Omega u^2(x,y)\, dx\, dy\right)$$

so

$$2a^2 \iint_\Omega u^2\, dx\, dy \le 2\left(\iint_\Omega u\, dx\, dy\right)^2 + 2a^4 \iint_\Omega \left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] dx\, dy$$

and we obtain the needed inequality with a constant $m = \max(a^2, 1/a^2)$. It can be shown that Poincairé's inequality is valid for more general domains. A modification of the inequality on a more general domain will be proved in Section 1.26.

Let us return to the free membrane problem. Provided we consider only the membrane's state of stress, any two membrane states are identical if

they are described by functions $u_1(x, y)$ and $u_2(x, y)$ whose difference is constant. We gather all functions (such that the difference between any two is a constant) into a class denoted by $u_*(x, y)$. There is a unique representative of $u_*(x, y)$ denoted by $u_b(x, y)$ such that

$$\iint_{\Omega} u_b(x, y)\, dx\, dy = 0. \tag{1.10.11}$$

For this so-called *balanced representative* (or *balanced function*), Poincairé's inequality takes the form

$$\iint_{\Omega} u_b^2(x, y)\, dx\, dy \le m \iint_{\Omega} \left[\left(\frac{\partial u_b}{\partial x} \right)^2 + \left(\frac{\partial u_b}{\partial y} \right)^2 \right] dx\, dy. \tag{1.10.12}$$

Now it is clear that there are no "rigid" motions of the membrane other than $u(x, y) = c$.

Because (1.10.12) has the same form as Friedrichs's inequality, we can repeat our former arguments to construct the energy space E_{MF} for a free membrane using the balanced representatives of the classes $u_*(x, y)$. In what follows we shall use this space E_{MF}, remembering that its elements all satisfy (1.10.11).

The condition (1.10.11) is a geometrical constraint resulting from our mathematical technique. Solving the static free membrane problem, we must remember that nature does not impose this constraint — the membrane can move as a "rigid body." But if we consider only deformations, the results must be independent of such motions. Consider then the functional of the work of external forces

$$A = \iint_{\Omega} F(x, y) U(x, y)\, dx\, dy.$$

If we use the space E_{MF}, then A makes sense if $F(x, y) \in L^2(\Omega)$ (guaranteed by (1.10.12) together with the Schwarz inequality in $L^2(\Omega)$). This is the only restriction on external forces for a clamped membrane. However, in the case of a *static* free membrane the functional A must be invariant under transformations of the form $u(x, y) \mapsto u(x, y) + c$. This requires that

$$\iint_{\Omega} F(x, y)\, dx\, dy = 0 \tag{1.10.13}$$

be satisfied. Again, we consider the static problem where rigid motion, however, is possible. Since we did not introduce inertia forces, we have put formally the mass of the membrane to zero. In this situation of zero mass, any forces with nonzero resultant would make the membrane as a whole

move with infinite acceleration. Thus, (1.10.13) also precludes such physical nonsense.

There is another way to construct an energy space using the classes $u_*(x, y)$ as base elements. Here the zero element is the set of all constants. Then the completion of the set of all these classes in the metric (1.10.7) is the energy space, the set of equivalent Cauchy sequences each of the elements of which is determined to within a constant. (In algebra, such a construction is called a factor space by the space of constants, but we shall not use this terminology.) The inequality (1.10.12) remains valid for representatives of classes — elements satisfying (1.10.11) which we uniquely choose from every class.

The restriction (1.10.13) is necessary for the functional of external forces to be uniquely defined for an element $U_*(x, y)$. We shall use the same notation E_{MF} for this type of energy space since there is a one-to-one correspondence, preserving distances and inner products, between the two types of energy space for the free membrane. Moreover, we shall always make clear which version we mean.

Those familiar with the theory of the Neumann problem should note that the necessary condition for its solvability which arises in mathematical physics, as a mathematical consequence, is of the same nature as (1.10.13).

Finally, we note that Poisson's equation governs not only membranes, but also situations in electricity, magnetism, hydrodynamics, mathematical biology, and other fields. So we can consider spaces such as E_M in various other sciences. It is clear that the results will be the same.

We will proceed to introduce other energy spaces in a similar manner: they will be completions of corresponding metric (inner product) spaces consisting of smooth functions satisfying certain boundary conditions. The problem is to determine properties of the elements of those completions. As a rule, metrics must contain all terms of internal energy (we now discuss only linear systems). For example, we can consider a membrane whose edge is elastically supported; then we must include the energy of elastic support in the expression for the energy metric.

Bending a Plate

Here we begin with the work of internal forces on variations of displacements

$$(w_1, w_2) = \iint\limits_{\Omega} D^{\alpha\beta\gamma\delta}\rho_{\gamma\delta}(w_1)\rho_{\alpha\beta}(w_2)\,dx\,dy \tag{1.10.14}$$

where $w_1(x, y)$ is the normal displacement of the mid-surface Ω of the plate, $w_2(x, y)$ is its variation, $\rho_{\alpha\beta}(u)$ are components of the change-of-curvature

tensor,

$$\rho_{11}(u) = \frac{\partial^2 u}{\partial x^2}, \qquad \rho_{12} = \frac{\partial^2 u}{\partial x \partial y}, \qquad \rho_{22} = \frac{\partial^2 u}{\partial y^2},$$

$D^{\alpha\beta\gamma\delta}$ are components of the tensor of elastic constants of the plate such that

$$D^{\alpha\beta\gamma\delta} = D^{\gamma\delta\alpha\beta} = D^{\beta\alpha\gamma\delta} \tag{1.10.15}$$

and, for any tensor $\rho_{\alpha\beta}$ there exists a constant $m_0 > 0$ such that

$$D^{\alpha\beta\gamma\delta}\rho_{\gamma\delta}\rho_{\alpha\beta} \geq m_0 \sum_{\alpha,\beta=1}^{2} \rho_{\alpha\beta}^2. \tag{1.10.16}$$

(We suppose $D^{\alpha\beta\gamma\delta}$ to be constants but piecewise continuity of these parameters would be sufficient.)

For the theory of shells and plates, here and in what follows, Greek indices assume values from the set $\{1, 2\}$ while Latin indices assume values from the set $\{1, 2, 3\}$. The repeated index convention for summation is also in force:

$$a^{\alpha\beta}b_{\alpha\beta} \equiv \sum_{\alpha,\beta=1}^{2} a^{\alpha\beta}b_{\alpha\beta}.$$

We first consider a plate with clamped edge $\partial\Omega$:

$$w\Big|_{\partial\Omega} = \frac{\partial w}{\partial n}\Big|_{\partial\Omega} = 0. \tag{1.10.17}$$

(Of course, the variation of w must satisfy (1.10.17) too.) Let us show that on S_4, the subset of $C^{(4)}(\Omega)$ consisting of those functions which satisfy (1.10.17), the form (w_1, w_2) given in (1.10.14) is an inner product. We begin with the axiom P1:

$$\begin{aligned}(w, w) &= \iint_{\Omega} D^{\alpha\beta\gamma\delta}\rho_{\alpha\beta}(w)\rho_{\gamma\delta}(w)\,dx\,dy \\ &\geq m_0 \iint_{\Omega} \sum_{\alpha,\beta=1}^{2} \rho_{\alpha\beta}^2(w)\,dx\,dy \\ &= m_0 \iint_{\Omega} \left[\left(\frac{\partial^2 w}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 w}{\partial x \partial y}\right)^2 + \left(\frac{\partial^2 w}{\partial y^2}\right)^2\right] dx\,dy \geq 0.\end{aligned}$$

If $w = 0$ then $(w, w) = 0$. If $(w, w) = 0$ then, on Ω,

$$\frac{\partial^2 w}{\partial x^2} = 0, \qquad \frac{\partial^2 w}{\partial x \partial y} = 0, \qquad \frac{\partial^2 w}{\partial y^2} = 0.$$

It follows that

$$w(x,y) = a_1 + a_2 x + a_3 y$$

where the a_i are constants. From (1.10.17) then, $w(x,y) = 0$. Hence P1 is satisfied. Satisfaction of P2 follows from (1.10.15), and it is evident that P3 is also satisfied. Thus S_4 with inner product (1.10.14) is an inner product space; its completion in the corresponding metric is the energy space E_{PC} for a clamped plate.

Let us consider some properties of the elements of E_{PC}. It was shown that

$$m_0 \iint_\Omega \left[\left(\frac{\partial^2 w}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 w}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 w}{\partial y^2} \right)^2 \right]^2 dx\, dy$$
$$\leq \iint_\Omega D^{\alpha\beta\gamma\delta} \rho_{\gamma\delta}(w) \rho_{\alpha\beta}(w)\, dx\, dy \equiv (w,w). \tag{1.10.18}$$

From this and the Friedrichs inequality (written for the first derivatives of $w \in S_4$ as well) we get

$$\iint_\Omega w^2\, dx\, dy \leq m_1 \iint_\Omega \left[\left(\frac{\partial w}{\partial x} \right)^2 + \left(\frac{\partial w}{\partial y} \right)^2 \right] dx\, dy$$
$$\leq m_2 \iint_\Omega \left[\left(\frac{\partial^2 w}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 w}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 w}{\partial y^2} \right)^2 \right] dx\, dy$$
$$\leq m_3 \iint_\Omega D^{\alpha\beta\gamma\delta} \rho_{\gamma\delta}(w) \rho_{\alpha\beta}(w)\, dx\, dy \equiv m_3 (w,w). \tag{1.10.19}$$

This means that, for a Cauchy sequence $\{w_n\}$ in E_{PC}, $w_n \in S_4$, the sequences

$$\{w_n\}, \qquad \{\partial w_n/\partial x\}, \qquad \{\partial w_n/\partial y\},$$

as well as

$$\{\partial^2 w_n/\partial x^2\}, \qquad \{\partial^2 w_n/\partial x \partial y\}, \qquad \{\partial^2 w_n/\partial y^2\},$$

are Cauchy sequences in $L^2(\Omega)$. So we can say that an element W of the completion E_{PC} is such that $W(x,y)$ and all its derivatives up to order two are in $L^2(\Omega)$.

We now investigate $W(x,y)$ further. Let $w \in S_4$ and $w(x,y) \equiv 0$ outside Ω. Suppose Ω lies in the domain $\{(x,y) \mid x > 0, y > 0\}$. Then the representation

$$w(x,y) = \int_0^x \int_0^y \frac{\partial^2 w(s,t)}{\partial s \partial t}\, ds\, dt$$

holds. Using the Hölder inequality and (1.10.19) we get

$$|w(x,y)| \le \int_0^x \int_0^y \left|\frac{\partial^2 w(s,t)}{\partial s \partial t}\right| ds\, dt \le \iint\limits_\Omega \left|\frac{\partial^2 w(s,t)}{\partial s \partial t}\right| ds\, dt$$
$$\le (\operatorname{mes}\Omega)^{1/2} \left(\iint\limits_\Omega \left(\frac{\partial^2 w(s,t)}{\partial s \partial t}\right)^2 ds\, dt \right)^{1/2}$$
$$\le m_4(w,w)^{1/2}. \tag{1.10.20}$$

This means that the sequence $\{w_n\}$ which is a Cauchy sequence in the metric of E_{PC}, $w_n \in S_4$, converges uniformly on Ω. Hence there exists a limit function $w_0(x,y) = \lim_{n\to\infty} w_n(x,y)$ which is continuous on Ω; this function is identified, as above, with the corresponding element of E_{PC} and we shall say that E_{PC} is continuously imbedded into $C(\Omega)$.

The functional of the work of external forces

$$A = \iint\limits_\Omega F(x,y) W(x,y)\, dx\, dy$$

now makes sense if $F(x,y) \in L(\Omega)$; moreover, it can contain the work of point forces

$$\sum_k F(x_k,y_k) w_0(x_k,y_k)$$

and line forces

$$\int_\gamma F(x,y) w_0(x,y)\, ds$$

where γ is a line in Ω. (We assume that $w_0(x,y)$ is the corresponding limit function for $W(x,y)$.)

In modern books on partial differential equations, they require that $F(x,y) \in H^{-2}(\Omega)$. This is a complete characterization of external forces — however, it is difficult for an engineer to verify this property.

Now let us consider a plate with free edge. In this case, we also wish to use the inner product (1.10.14) to create an energy space. As in the case of a membrane with free edge, the axiom P1 is not fulfilled: we saw that from $(w,w) = 0$ it followed that

$$w = a_1 + a_2 x + a_3 y. \tag{1.10.21}$$

This admissible motion of the plate as a rigid whole is called a rigid motion, but still differs from real "rigid" motions of the plate as a three-dimensional body.

We shall first use Poincaré's inequality (1.10.9) to show that the zero element of the corresponding completion is composed of functions of the form

(1.10.21). For this, taking $w(x,y) \in C^{(4)}(\Omega)$ we write down the Poincaré inequality for $\partial w/\partial x$ and $\partial w/\partial y$:

$$\iint\limits_{\Omega} \left(\frac{\partial w}{\partial x}\right)^2 dx\,dy \le m \left\{ \left(\iint\limits_{\Omega} \frac{\partial w}{\partial x}\,dx\,dy \right)^2 + \right.$$
$$\left. + \iint\limits_{\Omega} \left[\left(\frac{\partial^2 w}{\partial x^2}\right)^2 + \left(\frac{\partial^2 w}{\partial x \partial y}\right)^2 \right] dx\,dy \right\},$$

and then the same inequality with the roles of x and y interchanged. From this and (1.10.9) we get

$$\iint\limits_{\Omega} \left[w^2 + \left(\frac{\partial w}{\partial x}\right)^2 + \left(\frac{\partial w}{\partial y}\right)^2 \right] dx\,dy \le m_1 \left\{ \left(\iint\limits_{\Omega} w\,dx\,dy \right)^2 + \right.$$
$$+ \left(\iint\limits_{\Omega} \frac{\partial w}{\partial x}\,dx\,dy \right)^2 + \left(\iint\limits_{\Omega} \frac{\partial w}{\partial y}\,dx\,dy \right)^2 +$$
$$\left. + \iint\limits_{\Omega} \left[\left(\frac{\partial^2 w}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 w}{\partial x \partial y}\right)^2 + \left(\frac{\partial^2 w}{\partial y^2}\right)^2 \right] dx\,dy \right\}$$

and from (1.10.18) it follows that

$$\iint\limits_{\Omega} \left[w^2 + \left(\frac{\partial w}{\partial x}\right)^2 + \left(\frac{\partial w}{\partial y}\right)^2 \right] dx\,dy \le m_2 \left\{ \left(\iint\limits_{\Omega} w\,dx\,dy \right)^2 + \right.$$
$$+ \left(\iint\limits_{\Omega} \frac{\partial w}{\partial x}\,dx\,dy \right)^2 + \left(\iint\limits_{\Omega} \frac{\partial w}{\partial y}\,dx\,dy \right)^2 +$$
$$\left. + \iint\limits_{\Omega} D^{\alpha\beta\gamma\delta} \rho_{\gamma\delta}(w) \rho_{\alpha\beta}(w)\,dx\,dy \right\}. \tag{1.10.22}$$

For any function $w(x,y) \in C^{(4)}(\Omega)$, we can take suitable constants a_i and find a function $w_b(x,y)$ of the form

$$w_b = w + a_1 + a_2 x + a_3 y \tag{1.10.23}$$

such that

$$\iint\limits_{\Omega} w_b\,dx\,dy = 0, \quad \iint\limits_{\Omega} \frac{\partial w_b}{\partial x}\,dx\,dy = 0, \quad \iint\limits_{\Omega} \frac{\partial w_b}{\partial y}\,dx\,dy = 0. \tag{1.10.24}$$

As for the membrane with free edge, we can now consider a subset S_{4b} of $C^{(4)}(\Omega)$ consisting of those functions which satisfy (1.10.24). We shall refer to such functions as the balanced functions as we did for the membrane. We construct an energy space E_{PF} for a plate with free edge as the completion of S_{4b} in the metric corresponding to the inner product (1.10.14).

From (1.10.24), (1.10.22), and (1.10.18), we see that an element $W(x, y)$ of E_{PF} is such that $W(x, y)$ and all its "derivatives" up to order two are in $L^2(\Omega)$. Here we could show the existence of a limit function $w_0(x, y) = \lim_{n\to\infty} w_n$, $w_0(x, y) \in C(\Omega)$, for any Cauchy sequence $\{w_n\}$, but in this case the technique is more complicated and, in what follows, we have this result as a particular case of the Sobolev imbedding theorem.

Note that the system of relations (1.10.24) can be replaced by the system

$$\iint_\Omega w(x, y)\,dx\,dy = 0, \quad \iint_\Omega xw(x, y)\,dx\,dy = 0, \quad \iint_\Omega yw(x, y)\,dx\,dy = 0,$$

since these also fix uniquely the a_i for a class of functions of the form (1.10.23). (This possibility follows from the general result by S.L. Sobolev [22] on equivalent norms in Sobolev spaces.)

The system (1.10.24) represents constraints which are absent in nature. For a static problem there must be a certain invariance of some objects under transformations of the form (1.10.23). In particular, the work of external forces does not depend on such transformations if the problem is stated correctly. This leads to the necessary conditions

$$\iint_\Omega F(x, y)\,dx\,dy = 0, \quad \iint_\Omega xF(x, y)\,dx\,dy = 0, \quad \iint_\Omega yF(x, y)\,dx\,dy = 0. \tag{1.10.25}$$

The mechanical sense of (1.10.25) is clear: the resultant force and moments vanish. This is the condition for a self-balanced force system.

Problem 1.10.1. What is the form of (1.10.25) if the external forces contain point and line forces?

An energy space, as for the membrane with free edge, can be introduced in another way: namely, we combine all elements of the form $w(x, y) + a_1 + a_2x + a_3y$ with different constants a_i into a class which we consider as a single element of the base space S_{4*}; the zero element of S_{4*} is the set of all polynomials of the form $a_1 + a_2x + a_3y$. The completion of S_{4*} in the metric of E_{PF} is an energy space of elements whose natures differ from those of E_{PF}. However, we can state a one-to-one correspondence between elements of both spaces, so for this space we retain the notation E_{PF}. We advise the reader to carry through in detail the construction of this space. For example, one may consider mixed boundary conditions: how must the treatment be modified if the plate is clamped only along a segment $AB \subset \Omega$

so that

$$w(x,y)\Big|_{AB} = 0,$$

with the rest of the boundary free of geometrical constraints?

Linear Elasticity

We return to the problem of linear elasticity, which was considered in Section 1.3. Let us introduce a functional of the work of internal forces on variations $\mathbf{v}(\mathbf{x})$ of the displacement field $\mathbf{u}(\mathbf{x})$:

$$(\mathbf{u},\mathbf{v}) = \frac{1}{2}\int_\Omega c^{ijkl}\epsilon_{kl}(\mathbf{u})\epsilon_{ij}(\mathbf{v})\,d\Omega. \tag{1.10.26}$$

Here the elastic moduli c^{ijkl} may be piecewise continuous functions satisfying (1.3.10) and (1.3.11), which guarantee that all inner product axioms shall be satisfied by $(\mathbf{u},\mathbf{v})$ except P1: from $(\mathbf{u},\mathbf{u}) = 0$ it follows that $\mathbf{u} = \mathbf{a} + \mathbf{b}\times\mathbf{x}$. Note that $(\mathbf{u},\mathbf{v})$ is consistent with the metric (1.3.12).

Let us consider boundary conditions prescribed by

$$\mathbf{u}(\mathbf{x})\Big|_{\partial\Omega} = 0. \tag{1.10.27}$$

If we use the form (1.10.26) on the set S_3 of vector-functions $\mathbf{u}(\mathbf{x})$ satisfying (1.10.27) and such that each of their components is of class $C^{(2)}(\Omega)$, then $(\mathbf{u},\mathbf{v})$ becomes an inner product and S_3 with this inner product becomes an inner product space. Its completion E_{EC} in the corresponding metric is the energy space of an elastic body with clamped boundary. To describe the properties of the elements of E_{EC}, we establish *Korn's inequality*:

Lemma 1.10.1. For a vector function $\mathbf{u}(\mathbf{x}) \in S_3$, we have

$$\int_\Omega \left[|\mathbf{u}|^2 + \sum_{i,j=1}^{3}\left(\frac{\partial u_i}{\partial x_j}\right)^2\right] d\Omega \le m\int_\Omega c^{ijkl}\epsilon_{kl}(\mathbf{u})\epsilon_{ij}(\mathbf{u})\,d\Omega \tag{1.10.28}$$

for some constant m which does not depend on $\mathbf{u}(\mathbf{x})$.

Proof. By (1.3.11) and Friedrichs's inequality, it is sufficient to show that

$$\int_\Omega \sum_{i,j=1}^{3}\left(\frac{\partial u_i}{\partial x_j}\right)^2 d\Omega \le m_1\int_\Omega \sum_{\substack{i,j=1\\ i\le j}}^{3}\epsilon_{ij}^2(\mathbf{u})\,d\Omega.$$

Consider the term on the right:

$$A \equiv \int_\Omega \sum_{\substack{i,j=1\\ i\le j}}^{3} \epsilon_{ij}^2(\mathbf{u})\, d\Omega = \frac{1}{4}\int_\Omega \sum_{\substack{i,j=1\\ i\le j}}^{3} \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)^2 d\Omega$$

$$= \int_\Omega \left\{ \sum_{i=1}^{3} \left(\frac{\partial u_i}{\partial x_i}\right)^2 + \frac{1}{4}\sum_{\substack{i,j=1\\ i<j}}^{3} \left[\left(\frac{\partial u_i}{\partial x_j}\right)^2 + \left(\frac{\partial u_j}{\partial x_i}\right)^2 + 2\frac{\partial u_i}{\partial x_j}\frac{\partial u_j}{\partial x_i} \right] \right\} d\Omega.$$

Integrating by parts (twice) the term

$$B = \frac{1}{2}\int_\Omega \sum_{\substack{i,j=1\\ i<j}}^{3} \frac{\partial u_i}{\partial x_j}\frac{\partial u_j}{\partial x_i}\, d\Omega = \frac{1}{2}\int_\Omega \sum_{\substack{i,j=1\\ i<j}}^{3} \frac{\partial u_i}{\partial x_i}\frac{\partial u_j}{\partial x_j}\, d\Omega$$

and using the elementary inequality $|ab| \le (a^2 + b^2)/2$,

$$B \le \frac{1}{4}\int_\Omega \sum_{\substack{i,j=1\\ i<j}}^{3} \left[\left(\frac{\partial u_i}{\partial x_i}\right)^2 + \left(\frac{\partial u_j}{\partial x_j}\right)^2 \right] d\Omega = \frac{1}{2}\int_\Omega \sum_{i=1}^{3} \left(\frac{\partial u_i}{\partial x_i}\right)^2 d\Omega;$$

we get

$$A \ge \frac{1}{4}\int_\Omega \sum_{i,j=1}^{3} \left(\frac{\partial u_i}{\partial x_j}\right)^2 d\Omega$$

which completes the proof. □

By Korn's inequality, we see that each component of an element $U \in E_{EC}$ belongs to E_{MC}, i.e., the u_i and their first derivatives belong to $L^2(\Omega)$.

Note that the construction of an energy space is the same if the boundary condition (1.10.27) is given only on some part $\partial\Omega_1$ of the boundary of Ω:

$$\mathbf{u}(\mathbf{x})\Big|_{\partial\Omega_1} = 0.$$

Korn's inequality is also valid but its proof is more complicated (see, for example, [19, 9]).

If we consider an elastic body with free boundary we meet difficulties similar to those for a membrane or plate with free edge: we must circumvent the difficulty with the zero element of the energy space. The restrictions

$$\int_\Omega \mathbf{u}\, d\Omega = 0, \qquad \int_\Omega \mathbf{x} \times \mathbf{u}(\mathbf{x})\, d\Omega = 0,$$

provide that the zero element is zero, and that Korn's inequality remains valid for corresponding vector functions. So we get an energy space with known properties. But we can also organize an energy space of classes in which the zero element is the set of all elements of the form $\mathbf{a} + \mathbf{b} \times \mathbf{x}$.

1.11 Sobolev Spaces

In a famous book [22], S.L. Sobolev introduced normed spaces which now bear his name; they are denoted by $W^{m,p}(\Omega)$. The norm in $W^{m,p}(\Omega)$ is

$$\|u\| = \left(\int_\Omega \sum_{|\alpha| \le m} |D^\alpha u|^p \, d\Omega \right)^{1/p}, \tag{1.11.1}$$

where m is an integer, $p \ge 1$, and Ω is compact in $\mathbb{R}^n$. (The reader may wish to review the D^α notation from page 13.) Indeed this is a norm on the set $C^{(m)}(\Omega)$: fulfillment of the axioms N1 and N2 is evident, and N3 is fulfilled by virtue of Minkowski's inequality (1.2.4). The completion of $C^{(m)}(\Omega)$ in the norm (1.11.1) gives us a Banach space $W^{m,p}(\Omega)$.

It is interesting to note that for Ω a segment $[a, b]$, the spaces $W^{m,p}(a, b)$ were introduced by S. Banach in his dissertation. Our interest in Sobolev spaces is clear, since the elements of each of our energy spaces belonged to $W^{m,2}(\Omega)$ for some m.

For $u \in L^p(\Omega)$, K.O. Friedrichs [10] introduced the notion of strong derivative, calling $v \in L^p(\Omega)$ a strong derivative $D^\alpha(u)$ if there exists a sequence $\{\varphi_n\}$, $\varphi_n \in C^{(\infty)}(\Omega)$, such that

$$\int_\Omega |u(\mathbf{x}) - \varphi_n(\mathbf{x})|^p \, d\Omega \to 0 \quad \text{and} \quad \int_\Omega |v(\mathbf{x}) - D^\alpha \varphi_n(\mathbf{x})|^p \, d\Omega \to 0 \text{ as } n \to \infty.$$

Since $C^{(\infty)}(\Omega)$ is dense in any $C^{(k)}(\Omega)$, we see that an element of $W^{m,p}(\Omega)$ has all strong derivatives up to the order m lying in $L^p(\Omega)$.

Another approach to introduce a generalized derivative was proposed by Sobolev. He used an idea of the classical calculus of variations: if

$$\int_\Omega u(\mathbf{x}) \varphi(\mathbf{x}) \, d\Omega = 0$$

for all finite infinitely differentiable functions $\varphi(\mathbf{x})$, then $u(\mathbf{x}) = 0$ almost everywhere (everywhere if $u(\mathbf{x})$ is to be continuous), along with the integration by parts formula. (We call $\varphi(\mathbf{x})$ "finite" on an open domain $\Omega \subset \mathbb{R}^n$ if $\varphi(\mathbf{x}) \in C^{(\infty)}(\Omega)$ and the closure of the set $M = \{\mathbf{x} \in \Omega \colon \varphi(\mathbf{x}) \ne 0\}$ is compact in Ω.) So $v \in L^p(\Omega)$ is called a *weak derivative* $D^\alpha u$ of $u \in L^p(\Omega)$ if for every finite function $\varphi(\mathbf{x})$ on Ω we have

$$\int_\Omega u(\mathbf{x}) D^\alpha \varphi(\mathbf{x}) \, d\Omega = (-1)^{|\alpha|} \int_\Omega v(\mathbf{x}) \varphi(\mathbf{x}) \, d\Omega. \tag{1.11.2}$$

The two definitions of generalized derivative are equivalent [22]. We shall not give the proof, as it would be beyond the scope of our presentation. The same is true for some other facts of this section.

The result we now present is a particular case of *Sobolev's imbedding theorem*.

For an element W of H_{PC}, a subset of $W^{2,2}(\Omega)$ consisting of those functions satisfying the boundary condition (1.10.17), we saw that there is a limit function w which we identified with W, which is continuous, i.e., has better smoothness properties, and

$$\|w\|_{C(\Omega)} \le m\|W\|_{H_{PC}}.$$

Such a situation is typical in Sobolev spaces. A correspondence between an element W of $W^{m,p}(\Omega)$ and its limit function w which belongs to a space Y is called the *imbedding operator* of $W^{m,p}(\Omega)$ to Y; this operator is continuous if for any $W \in W^{m,p}(\Omega)$ we have

$$\|w\|_Y \le m\|W\|_{W^{m,p}(\Omega)} \tag{1.11.3}$$

with a constant m independent of W. Here we use the notation $\|\cdot\|_X$ to emphasize that the norm under discussion is the norm on a certain space X.

We assume that the compact set $\Omega \subset \mathbb{R}^n$ satisfies the so-called *cone condition*. This means that there is a finite circular cone in $\mathbb{R}^n$ such that any point of the boundary of Ω can be touched by the vertex of the cone while the cone lies fully inside Ω. This is the condition under which Sobolev's imbedding theorem is proved. We denote by Ω_r an r-dimensional piecewise smooth hypersurface in Ω. (This means that, at any point of smoothness, in a local coordinate system, it is described by functions having all derivatives continuous up to order m locally, if we consider $W^{m,p}(\Omega)$.)

The theory of Sobolev spaces and their extensions is a substantial branch of mathematics (see Adams [1], Lions and Magenes [18], etc.). We formulate only what is needed for our purposes, using the notion of a compact operator which will be introduced later (Section 2.6). This is Sobolev's imbedding theorem with some extensions:

Theorem 1.11.1. The imbedding operator of $W^{m,p}(\Omega)$ to $L^q(\Omega_r)$ is continuous if one of the following conditions holds:

(i) $n > mp$, $r > n - mp$, $q \le pr/(n - mp)$;

(ii) $n = mp$, q is finite with $q \ge 1$.

If $n < mp$, then the space $W^{m,p}(\Omega)$ is imbedded into the Hölder space $H^\alpha(\overline{\Omega})$ when $\alpha \le (mp - n)/p$, and the imbedding operator is continuous.

The imbedding operator of $W^{m,p}(\Omega)$ to $L^q(\Omega_r)$ is compact (i.e., takes every bounded set of $W^{m,p}(\Omega)$ into a precompact set of the corresponding space) if

(i) $n > mp$, $r > n - mp$, $q < pr/(n - mp)$ or

(ii) $n = mp$ and q is finite with $q \geq 1$.

If $n < mp$ then the imbedding operator is compact to $H^{\alpha}(\overline{\Omega})$ when $\alpha < (mp - n)/p$.

Note that this theorem, the second part of which is known as the Sobolev–Kondrashov imbedding theorem, allows us to get imbedding properties not only for functions but also for their derivatives: if $u \in W^{m,p}(\Omega)$ then $D^{\alpha}u \in W^{m-k,p}(\Omega)$ when $|\alpha| = k$. Also available are stricter results on the imbedding of Sobolev spaces on Ω into the spaces of functions given on manifolds Ω_r of dimension less than n. They are known as the trace theorems. We shall not present them here, since they require an extended notion of Sobolev spaces.

Let us formulate some consequences of Theorem 1.11.1 that we shall frequently use.

Theorem 1.11.2. Let γ be a piecewise differentiable curve in a compact set $\Omega \subset \mathbb{R}^2$. For any finite $q \geq 1$, the imbedding operator of $W^{1,2}(\Omega)$ to the spaces $L^q(\Omega)$ and $L^q(\gamma)$ is continuous (and compact), i.e.,

$$\max\{\|u\|_{L^q(\Omega)}, \|u\|_{L^q(\gamma)}\} \leq m\|u\|_{W^{1,2}(\Omega)} \tag{1.11.4}$$

with a constant m which does not depend on $u(\mathbf{x})$.

Theorem 1.11.3. Let $\Omega \subset \mathbb{R}^2$ be compact. If $\alpha \leq 1$, the imbedding operator of $W^{2,2}(\Omega)$ to $H^{\alpha}(\overline{\Omega})$ is continuous; if $\alpha < 1$, it is compact. For the first derivatives, the imbedding operator to $L^q(\Omega)$ and $L^q(\gamma)$ is continuous (and compact) for any finite $q \geq 1$.

Theorem 1.11.4. Let γ be a piecewise smooth surface in a compact set $\Omega \subset \mathbb{R}^3$. The imbedding operator of $W^{1,2}(\Omega)$ to $L^q(\Omega)$ when $1 \leq q \leq 6$, and to $L^p(\gamma)$ when $1 \leq p \leq 4$, is continuous; if $1 \leq q < 6$ or $1 \leq p < 4$, respectively, then it is compact.

We merely indicate how such theorems are proved. We established similar results for the bar problem (see (1.10.5)) and for the clamped plate problem (see (1.10.20)). At that time, we used the integral representations of functions of certain classes. In like manner, the original proof of Sobolev is given for Ω a union of bounded star-shaped domains. (A domain is called star-shaped with respect to a ball B if any ray with origin in B intersects the boundary of the domain only once.) For a domain Ω which is bounded and star-shaped with respect to a ball B, a function $u(\mathbf{x}) \in C^{(m)}(\Omega)$ can be represented in the form

$$u(\mathbf{x}) = \sum_{|\alpha| \leq m-1} x_1^{\alpha_1} \cdots x_n^{\alpha_n} \int_B K_{\alpha}(\mathbf{y}) u(\mathbf{y})\, d\Omega + \\ + \int_{\Omega} \frac{1}{|\mathbf{x} - \mathbf{y}|^{n-m}} \sum_{|\alpha| = m} K_{\alpha}(\mathbf{x}, \mathbf{y}) D^{\alpha} u(\mathbf{y})\, d\Omega_{\mathbf{y}} \tag{1.11.5}$$

where $K_\alpha(\mathbf{y})$ and $K_\alpha(\mathbf{x},\mathbf{y})$ are continuous functions. Investigating properties of the integral terms on the right-hand side of the representation (1.11.5), Sobolev formulated his results; later they were extended to more general domains.

Another method is connected with the Fourier transformation of functions. In the case of $W^{m,2}(\Omega)$, it is necessary to extend functions of $C^{(m)}(\overline{\Omega})$ outside Ω in such a way that they belong to $C^m(\mathbb{R}^n)$ and $W^{m,2}(\mathbb{R}^n)$. Then using the Fourier transformation

$$\hat{u}(\mathbf{y}) = (2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{-i\mathbf{x}\cdot\mathbf{y}} u(\mathbf{x})\, dx_1 \cdots dx_n$$

along with the facts that

$$\|u(\mathbf{x})\|_{L^2(\mathbb{R}^n)} = \|\hat{u}(\mathbf{y})\|_{L^2(\mathbb{R}^n)}$$

and

$$\widehat{D^\alpha u}(\mathbf{x}) = (iy_1)^{\alpha_1} \cdots (iy_n)^{\alpha_n} \hat{u}(\mathbf{y})$$

for $u \in L^2(\mathbb{R}^n)$, we can present the norm in $W^{m,2}(\Omega)$ in the form

$$\|u(\mathbf{x})\|^2_{W^{m,2}(\mathbb{R}^n)} = \sum_{|\alpha|\le m} \|y_1^{\alpha_1} \cdots y_n^{\alpha_n} \hat{u}(\mathbf{y})\|^2_{L^2(\mathbb{R}^n)}. \tag{1.11.6}$$

We can then study the properties of the weighted space $L^2_w(\mathbb{R}^n)$; this transformed problem is simpler, as many of the problems involved are algebraic estimates of Fourier images.

Moreover, we can consider $W^{p,2}(\mathbb{R}^n)$ with fractional indices p. These lead to necessary and sufficient conditions for the trace problem: given $W^{m,2}(\Omega)$, find the space $W^{p,2}(\partial\Omega)$ in which $W^{m,2}(\Omega)$ is continuously imbedded. The inverse trace problem is, given $W^{p,2}(\Omega)$, find the minimal index m such that every element $u \in W^{p,2}(\partial\Omega)$ can be extended to $\overline{\Omega}$, $u^* \in W^{m,2}(\Omega)$, in such a way that

$$\|u^*\|_{W^{m,2}(\Omega)} \le c\|u\|_{W^{p,2}(\partial\Omega)}.$$

In this way, many results from the contemporary theory of elliptic (and other types of) equations and systems are obtained. We should mention that the trace theorems are formulated mostly for smooth manifolds, hence are not applicable to practical problems involving domains with corners.

1.12 Introduction to Operators

We have already used the notions of operator and functional, with the understanding that the reader has surely encountered these basic notions in other subject areas. At this point, we pause to give a formal definition valid for metric and more general spaces.

Definition 1.12.1. Let X and Y be metric spaces. A correspondence $x \mapsto y = A(x)$ where $x \in X$ and $y \in Y$ is called an *operator* (*from X to Y*) if to each $x \in X$ there corresponds no more than one $y \in Y$. The set of all x for which there exists a corresponding $y = A(x)$ is called the *domain* of A, denoted $D(A)$, whereas the image of $D(A)$ is the *range* of A, denoted $R(A)$.

Note that there is nothing to prevent us from having $Y = X$. A particular case of the concept of operator occurs when $R(A) \subset \mathbb{R}$ or $\mathbb{C}$: we then refer to A as a *real* or *complex functional*, respectively.

In accordance with the classical definition of continuity of a function, we say that A is continuous at $x_0 \in X$ if for any $\varepsilon > 0$ there exists $\delta > 0$ (dependent on ε) such that $d(A(x), A(x_0)) < \varepsilon$ whenever $d(x, x_0) < \delta$. If A is continuous at every point of an open domain M, then it is said to be continuous in M.

If X and Y are linear spaces we can consider a class of linear operators $A\colon X \to Y$. For all $x_1, x_2 \in X$, a linear operator A satisfies

$$A(\lambda_1 x_1 + \lambda_2 x_2) = \lambda_1 A(x_1) + \lambda_2 A(x_2)$$

where λ_1, λ_2 are any (real or complex) numbers. For a linear operator A, an image $A(x)$ is usually denoted by Ax.

In this section, from now on, we let X and Y be normed spaces. The definition of continuity of an operator is changed in an evident way. For a linear operator A we have $Ax - Ax_0 = A(x - x_0)$, so A is continuous in the whole space X if and only if it is continuous at a single point $x_0 \in X$, say $x_0 = 0$. This allows us to formulate the next result.

Theorem 1.12.1. A linear operator A from X to Y, X and Y being normed spaces, is continuous if and only if there is a constant c such that for every $x \in X$

$$\|Ax\| \le c\|x\|. \tag{1.12.1}$$

The infimum of all such constants c is called the *norm* of A, denoted $\|A\|$.

Proof. We need only show continuity of A at $x = 0$. If (1.12.1) holds, then this continuity is clear by definition. Conversely, suppose A is continuous at $x = 0$. Take $\varepsilon = 1$; by definition there exists $\delta > 0$ such that $\|Ax\| < 1$ whenever $\|x\| < \delta$. For every nonzero $x \in X$, the norm of $x^* = \delta x/(2\|x\|)$ is

$$\|x^*\| = \|\delta x/(2\|x\|)\| = \delta/2 < \delta,$$

so $\|Ax^*\| < 1$. Since A is linear, we have

$$\frac{\delta}{2\|x\|}\|Ax\| < 1 \quad \text{or} \quad \|Ax\| < \frac{2}{\delta}\|x\|.$$

This is (1.12.1) with $c = 2/\delta$. □

A linear operator A satisfying (1.12.1) is said to be *bounded*.

Theorem 1.12.2. A linear operator A from X to Y, X and Y being normed spaces, is continuous if and only if for every sequence $\{x_n\}$, $x_n \to 0$, the sequence $Ax_n \to 0$ as $n \to \infty$.

Proof. It is clear that for a continuous operator $Ax_n \to 0$ if $x_n \to 0$ as $n \to \infty$. We prove the converse. Let $Ax_n \to 0$ for every sequence $\{x_n\}$ such that $x_n \to 0$. Suppose to the contrary that A is not continuous (i.e., that it is not bounded). Then there exists a sequence $\{x_n\}$ such that $\|x_n\| \le 1$ but $\|Ax_n\| \to \infty$. We can assume (why?) that $\|Ax_n\| \ge n$. Consider the sequence $y_n = x_n/\sqrt{n}$. It is clear that $\|y_n\| \to 0$ but $\|Ay_n\| \ge \sqrt{n}$ so Ay_n does not tend to 0. This contradiction completes the proof. □

Thus, for a linear operator acting from X to Y we can introduce two other equivalent definitions of continuity: the first uses the property of boundedness (inequality (1.12.1)), and the second, defined by Theorem 1.12.2, uses the notion of *sequential continuity*.

Let us consider some examples of operators:

1. The operator d/dx is the operator of differentiation. It is clear that it is bounded from $C^{(1)}(-\infty, \infty)$ to $C(-\infty, \infty)$.

2. A differential operator $Au = \sum_{|\alpha| \le m} a_\alpha D^\alpha u(\mathbf{x})$ with constant coefficients a_α is bounded (hence continuous) from $C^{(m)}(\overline{\Omega})$ to $C(\overline{\Omega})$ and from $W^{m,2}(\Omega)$ to $L^2(\Omega)$. The reader should take a moment to construct an example where it is not continuous.

3. The operator of integration defined by

$$B(u)(x) = \int_0^x u(s)\, ds$$

is continuous from $C(0,1)$ to $C(0,1)$. It is left as an exercise for the reader to determine whether it is continuous from $C(0,1)$ to $C^{(1)}(0,1)$.

Lastly, we mention that various authors employ other terms instead of "operator." We sometimes find instead the terms transformation, map, mapping, or simply function.

1.13 Contraction Mapping Principle

In the particular case $X = Y$, an operator A is said to be acting in X. Many problems of mechanics can be formulated as equations of the form

$$x = Ax \tag{1.13.1}$$

where A acts in a metric space X. A solution to (1.13.1) is called a *fixed point* of A. In the Introduction we saw two different problems whose solutions were in a certain sense similar. There are many other such problems; their general similarity is captured in the following definition.

Definition 1.13.1. An operator A acting in a metric space X is called a *contraction* in X if for every pair $x, y \in X$ there is a number q, $0 < q < 1$, such that

$$d(A(x), A(y)) \le q\, d(x, y). \tag{1.13.2}$$

The following central theorem is known as the *contraction mapping principle* or *Banach's principle of successive approximations*:

Theorem 1.13.1. Let A be a contraction operator (with a constant $q < 1$) in a complete metric space X. Then:

(i) A has a unique fixed point $x_* \in X$;

(ii) For any initial approximation $x_0 \in X$, the sequence of successive approximations

$$x_{k+1} = A(x_k), \qquad k = 0, 1, 2, \dots \tag{1.13.3}$$

converges to x_*; the rate of convergence is estimated by

$$d(x_k, x_*) \le \frac{q^k}{1-q} d(x_0, x_1). \tag{1.13.4}$$

Remark 1.13.1. X need not be a linear space. Banach's principle works if X is a closed subset of a metric space such that $A(X) \subset X$ and A is a contraction operator in X.

Proof of Theorem 1.13.1. We first show uniqueness of a fixed point of A. Assuming the existence of two such points x_1, x_2 with $x_1 = A(x_1)$, $x_2 = A(x_2)$, we get

$$d(x_1, x_2) = d(A(x_1), A(x_2)) \le q\, d(x_1, x_2);$$

as $q < 1$, this implies $d(x_1, x_2) = 0$. Hence $x_1 = x_2$.

Now take an element $x_0 \in X$ and consider the iterative procedure (1.13.3). For $d(x_n, x_{n+m})$, we successively obtain

$$\begin{aligned} d(x_n, x_{n+m}) &= d(A(x_{n-1}), A(x_{n+m-1})) \\ &\le q\, d(x_{n-1}, x_{n+m-1}) \\ &= q\, d(A(x_{n-2}), A(x_{n+m-2})) \\ &\le q^2 d(x_{n-2}, x_{n+m-2}) \\ &\vdots \\ &\le q^n d(x_0, x_m). \end{aligned}$$

But

$$\begin{aligned} d(x_0, x_m) &\le d(x_0, x_1) + d(x_1, x_2) + d(x_2, x_3) + \cdots + d(x_{m-1}, x_m) \\ &\le d(x_0, x_1) + q\, d(x_0, x_1) + q^2 d(x_0, x_1) + \cdots + q^{m-1} d(x_0, x_1) \\ &= (1 + q + q^2 + \cdots + q^{m-1}) d(x_0, x_1) \\ &= \frac{1 - q^m}{1 - q} d(x_0, x_1) \le \frac{1}{1 - q} d(x_0, x_1) \end{aligned}$$

so

$$d(x_n, x_{n+m}) \le \frac{q^n}{1 - q} d(x_0, x_1). \tag{1.13.5}$$

It follows that $\{x_n\}$ is a Cauchy sequence. Since X is complete there is an element $x_* \in X$ such that

$$x_* = \lim_{n \to \infty} x_n = \lim_{n \to \infty} A(x_{n-1}).$$

Let us estimate $d(x_*, A(x_*))$:

$$\begin{aligned} d(x_*, A(x_*)) &\le d(x_*, x_n) + d(x_n, A(x_*)) \\ &= d(x_*, x_n) + d(A(x_{n-1}), A(x_*)) \\ &\le d(x_*, x_n) + q\, d(x_{n-1}, x_*) \\ &\to 0 \text{ as } n \to \infty. \end{aligned}$$

Thus

$$x_* = A(x_*)$$

and x_* is a fixed point of A.

Passing to the limit as $m \to \infty$ in (1.13.5) gives the estimate (1.13.4), and completes the proof. □

Problem 1.13.1. Use one of the intermediate estimates of the proof to establish that

$$d(x_k, x_*) \le q^k d(x_0, x_*).$$

Why is this estimate of less practical value than (1.13.4)?

By A^N we denote

$$A^N(x) = \underbrace{A(A(\cdots(A}_{N \text{ times}}(x))\cdots)).$$

Problem 1.13.2. Let $Ay(t) = \int_0^t g(t - \tau) y(\tau)\, d\tau$, with $g(t) \in C[0, T]$, be an operator acting in $C[0, T]$. (1) Prove that A^N is a contraction operator for some integer N. Similar operators appear in the theory of viscoelasticity. (2) Is the statement valid if $g(t) \in L^p(0, T)$, $p > 1$?

Corollary 1.13.1. Let A^N, for some N, be a contraction operator in a complete metric space X. Then the operator A has a unique fixed point x_* to which the sequence of successive approximations (1.13.3) converges independently of choice of initial approximation $x_0 \in X$ and with rate

$$d(x_i, x_*) \le \frac{q^{i/N-1}}{1-q} \max\{d(x_0, x_N), d(x_1, x_{N+1}), \ldots, d(x_{N-1}, x_{2N-1})\}.$$

Proof. The operator A^N meets all requirements of Theorem 1.13.1, so the equation

$$x = A^N(x) \tag{1.13.6}$$

has a unique solution x_* and we have $x_* = A^N(x_*)$. We can apply A to both sides of this latter equation to obtain

$$A(x_*) = A(A^N(x_*)) = A^N(A(x_*)).$$

This means that $A(x_*)$ is also a solution to (1.13.6). From uniqueness of solution of (1.13.6), it follows that

$$x_* = A(x_*),$$

i.e., equation (1.13.1) is solvable. Noting that any fixed point of A is a fixed point of A^N, we get uniqueness of solution of (1.13.1). Finally, the whole sequence of successive approximations can be constructed by taking elements from each of the N subsequences

$$\begin{aligned}
&\{x_0, A^N(x_0), A^{2N}(x_0), \ldots\},\\
&\{A(x_0), A^{N+1}(x_0), A^{2N+1}(x_0), \ldots\},\\
&\vdots\\
&\{A^{N-1}(x_0), A^{2N-1}(x_0), A^{3N-1}(x_0), \ldots\}.
\end{aligned}$$

For each of these subsequences, the estimate (1.13.4) is valid: we have

$$\begin{aligned}
d(A^{kN}(x_0), x_*) &\le \frac{q^k}{1-q} d(x_0, A^N(x_0)),\\
d(A^{kN+1}(x_0), x_*) &\le \frac{q^k}{1-q} d(A(x_0), A^{N+1}(x_0)),\\
&\vdots\\
d(A^{kN+N-1}(x_0), x_*) &\le \frac{q^k}{1-q} d(A^{N-1}(x_0), A^{2N-1}(x_0)).
\end{aligned}$$

Replacing kN by i in the first inequality, $kN+1$ by i in the second inequality, and so on, and remembering that $x_k = A^k(x_0)$, we get

$$\begin{aligned}
d(x_i, x_*) &\le \frac{q^{i/N}}{1-q} d(x_0, x_N) && (i = kN),\\
d(x_i, x_*) &\le \frac{q^{(i-1)/N}}{1-q} d(x_1, x_{N+1}) && (i = kN+1),\\
&\vdots\\
d(x_i, x_*) &\le \frac{q^{(i-N+1)/N}}{1-q} d(x_{N-1}, x_{2N-1}) && (i = kN + (N-1)).
\end{aligned}$$

The desired estimate follows. □

We consider some applications of the Banach principle to more general systems of linear algebraic equations than were dealt with in the Introduction. We wish to solve the system

$$x_i = \sum_{j=1}^{\infty} a_{ij} x_j + c_i. \tag{1.13.7}$$

The corresponding operator A is defined by

$$\mathbf{y} = A\mathbf{x}, \qquad \mathbf{y} = (y_1, y_2, \ldots), \qquad y_i = \sum_{j=1}^{\infty} a_{ij} x_j + c_i.$$

Our subsequent treatment of this system depends on the space in which we seek a solution. If we take $X = m$, the space of bounded sequences with metric

$$d(\mathbf{x}, \mathbf{y}) = \sup_i |x_i - y_i|,$$

then we find that A is a contraction operator if

$$q = \sup_i \sum_{j=1}^{\infty} |a_{ij}| < 1 \tag{1.13.8}$$

and $\mathbf{c} = (c_1, c_2, \ldots) \in m$. So we can find a solution to (1.13.7) by the method of successive approximations beginning with any initial approximation from m.

Another restriction on the infinite matrix (a_{ij}) appears if we consider successive approximations in ℓ^p, $p > 1$. Here we get

$$d(A\mathbf{x}, A\mathbf{y}) = \left(\sum_{i=1}^{\infty} \left| \sum_{j=1}^{\infty} a_{ij}(x_j - y_j) \right|^p \right)^{1/p}.$$

Applying the Hölder inequality we obtain

$$d(A\mathbf{x}, A\mathbf{y}) \le \left(\sum_{i=1}^{\infty} \left(\sum_{j=1}^{\infty} |a_{ij}|^r \right)^{p/r} \sum_{j=1}^{\infty} |x_j - y_j|^p \right)^{1/p}$$
$$= \left(\sum_{i=1}^{\infty} \left(\sum_{j=1}^{\infty} |a_{ij}|^r \right)^{p/r} \right)^{1/p} d(\mathbf{x}, \mathbf{y})$$

where $1/r + 1/p = 1$. So A is a contraction operator in ℓ^p if $\mathbf{c} \in \ell^p$ and

$$q = \left(\sum_{i=1}^{\infty} \left(\sum_{j=1}^{\infty} |a_{ij}|^r \right)^{p/r} \right)^{1/p} < 1, \qquad r = \frac{p}{p-1}. \tag{1.13.9}$$

Now we can solve the system (1.13.7) by an iterative procedure in ℓ^p.

The values of q from (1.13.8) and (1.13.9) are the corresponding operator norms in m and ℓ^p, respectively (prove this).

In a similar way we can extend the result of the Introduction for the system (2) to more general systems of integral equations in different spaces. We leave this to the reader as an exercise.

We shall see Banach's principle applied to problems in plasticity. When applicable to such problems, it is a convenient and useful tool.

What can we say about the method relative to its use in numerical computation? The advantages of iterative procedures are well known; in particular, numerical error at each iterative step does not degrade the solution as a whole. However, the convergence rate does depend on q: with $q = 0.5$ or greater, say, convergence may be too slow to carry out many iterations on a very complicated system of equations. In such cases it may be possible to transform the iterative procedure in some way in order to speed up convergence (e.g., Seidel's method).

1.14 Generalized Solutions in Mechanics

We now discuss how to introduce generalized solutions in mechanics. We begin with Poisson's equation

$$-\Delta u(x, y) = F(x, y), \qquad (x, y) \in \Omega, \tag{1.14.1}$$

where Ω is a bounded open domain in $\mathbb{R}^2$. The Dirichlet problem consists of this equation supplemented by the boundary condition

$$u\Big|_{\partial\Omega} = 0. \tag{1.14.2}$$

Let $u(x, y)$ be its classical solution; i.e., $u \in C^{(2)}(\overline{\Omega})$ satisfies (1.14.1) and (1.14.2). Let $\varphi(x, y)$ be a finite function in Ω. (Recall this means that $\varphi \in C^{(\infty)}(\overline{\Omega})$ and the closure of the set $M = \{(x, y) \in \Omega \mid \varphi(x, y) \neq 0\}$ lies in Ω.)

Multiplying both sides of (1.14.1) by $\varphi(x, y)$ and integrating over Ω, we get

$$-\int_{\Omega} \varphi(x, y)\Delta u(x, y)\, dx\, dy = \int_{\Omega} F(x, y)\varphi(x, y)\, dx\, dy. \tag{1.14.3}$$

If this equality holds for every function $\varphi(x, y)$ that is finite and infinitely differentiable in Ω and if $u \in C^{(2)}(\overline{\Omega})$ and satisfies (1.14.2), then, as is well known from the classical calculus of variations, $u(x, y)$ is the unique classical solution to the Dirichlet problem.

But using (1.14.3), we can pose this Dirichlet problem directly without (1.14.1); namely, $u(x, y)$ is a solution to the Dirichlet problem if, obeying (1.14.2), it satisfies (1.14.3) for every $\varphi(x, y)$ finite and infinitely differentiable in Ω. If $F(x, y)$ belongs to $L^p(\Omega)$ then we can take, as it seems, $u(x, y)$ having second derivatives in the space $L^p(\Omega)$; such a $u(x, y)$ is not a classical solution, and it is natural to call it a generalized solution.

We can go further by applying integration by parts to the left-hand side of (1.14.3) as follows:

$$\int_{\Omega} \left(\frac{\partial u}{\partial x}\frac{\partial \varphi}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial \varphi}{\partial y} \right) dx\, dy = \int_{\Omega} F(x, y)\varphi(x, y)\, dx\, dy. \tag{1.14.4}$$

In such a case we may impose weaker restrictions on a solution $u(x, y)$ and call it the generalized solution if it belongs to E_{MC}, the energy space for a clamped membrane. Equation (1.14.4) defines this solution if it holds for every $\varphi(x, y)$ that is finite in Ω. Note the disparity in requirements on $u(x, y)$ and $\varphi(x, y)$.

Further integration by parts on the left-hand side of (1.14.4) gives us the equation

$$-\int_{\Omega} u(x, y)\Delta\varphi(x, y)\, dx\, dy = \int_{\Omega} F(x, y)\varphi(x, y)\, dx\, dy. \tag{1.14.5}$$

Now we can formally consider solutions from the space $L(\Omega)$ and this is a new class of generalized solutions.

This approach leads to the so-called theory of distributions, originated by Schwartz [21]. He extended the notion of generalized solution to a class of linear continuous functionals, *distributions*, defined on the set $\mathcal{D}(\Omega)$ of all functions finite and infinitely differentiable in Ω. For this it is necessary to introduce the convergence and other structures of continuity in $\mathcal{D}(\Omega)$. Unfortunately $\mathcal{D}(\Omega)$ is not a normed space (see, for example, Yosida [29] — it is a so-called locally convex topological space) and its presentation would be beyond our present scope. This theory justifies, in particular, the use of

the so-called δ-function, which was introduced in quantum mechanics via the equality

$$\int \delta(x-a)f(x)\,dx = f(a), \tag{1.14.6}$$

valid for every continuous $f(x)$. Physicists considered $\delta(x)$ to be a function whose value is zero everywhere except at $x=0$, where its value is infinity. Any known theory of integration gave zero for the value of the integral on the left-hand side of (1.14.6), and the theory of distributions explained how to understand such strange functions. It is interesting to note that the δ-function was well known in classical mechanics, too; if we consider $\delta(x-a)$ as a unit point force applied at $x=a$, then the integral on the left-hand side of (1.14.6) is the work of this force on the displacement $f(a)$, which is indeed $f(a)$.

So we have several generalized statements of the Dirichlet problem, but which one is most natural from the viewpoint of mechanics?

From mechanics it is known that a solution to the problem is a minimizer of the functional of total energy

$$I(u) = \int_\Omega \left[\left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 \right] dx\,dy - 2\int_\Omega Fu\,dx\,dy. \tag{1.14.7}$$

According to the calculus of variations, a minimizer of $I(u)$ on the subset of $C^{(2)}(\overline{\Omega})$ consisting of all functions satisfying (1.14.2) is a classical solution to the Dirichlet problem. But we can consider $I(u)$ on the energy space E_{MC} if $F(x,y) \in L^p(\Omega)$, $p>1$. Indeed, the first term in $I(u)$ is well defined in E_{MC} and can be written in the form $\|u\|^2$; the second,

$$\Phi(u) = -\int_\Omega F(x,y)u(x,y)\,dx\,dy,$$

is a linear functional with respect to $u(x,y)$. It is also continuous in E_{MC}; by virtue of Hölder's inequality with exponents p and $q=p/(p-1)$, we have

$$\begin{aligned}\left|\int_\Omega Fu\,dx\,dy\right| &\le \left(\int_\Omega |F|^p\,dx\,dy\right)^{1/p}\left(\int_\Omega |u|^q\,dx\,dy\right)^{1/q}\\ &\le m_1\|F\|_{L^p(\Omega)}\|u\|_{W^{1,2}(\Omega)}\\ &\le m_2\|u\|_{E_{MC}}.\end{aligned}$$

(Here we have used the imbedding Theorem 1.11.2 and the Friedrichs inequality.) By Theorem 1.12.1, $\Phi(u)$ is continuous in E_{MC}, and therefore so is $I(u)$.

Thus $I(u)$ is of the form

$$I(u) = \|u\|^2 + 2\Phi(u). \tag{1.14.8}$$

Let $u_0 \in E_{MC}$ be a minimizer of $I(u)$, i.e.,

$$I(u_0) \le I(u) \text{ for all } u \in E_{MC}. \tag{1.14.9}$$

We try a method from the classical calculus of variations. Take $u = u_0 + \epsilon v$, v being an arbitrary element of E_{MC}. Then

$$\begin{aligned} I(u) &= I(u_0 + \epsilon v) \\ &= \|u_0 + \epsilon v\|^2 + 2\Phi(u_0 + \epsilon v) \\ &= (u_0 + \epsilon v, u_0 + \epsilon v) + 2\Phi(u_0 + \epsilon v) \\ &= \|u_0\|^2 + 2\epsilon(u_0, v) + \epsilon^2\|v\|^2 + 2\Phi(u_0) + 2\epsilon\Phi(v) \\ &= \|u_0\|^2 + 2\Phi(u_0) + 2\epsilon[(u_0, v) + \Phi(v)] + \epsilon^2\|v\|^2. \end{aligned}$$

From (1.14.9), we get

$$2\epsilon[(u_0, v) + \Phi(v)] + \epsilon^2\|v\|^2 \ge 0.$$

Since ϵ is an arbitrary real number, it follows that

$$(u_0, v) + \Phi(v) = 0. \tag{1.14.10}$$

In other words,

$$\int_\Omega \left(\frac{\partial u_0}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u_0}{\partial y}\frac{\partial v}{\partial y} \right) dx\, dy - \int_\Omega F(x, y)v(x, y)\, dx\, dy = 0. \tag{1.14.11}$$

This equality is valid for every $v \in E_{MC}$, and defines the minimizer $u_0 \in E_{MC}$. Note that (1.14.11) has the same form as (1.14.4).

So we have introduced the notion of generalized solution which has an explicit mechanical background.

Definition 1.14.1. An element $u \in E_{MC}$ is called the *generalized solution* to the Dirichlet problem if u satisfies (1.14.11) for any $v \in E_{MC}$.

We can also obtain (1.14.11) from the principle of virtual displacements (work). This asserts that in the state of equilibrium, on all virtual (admissible) displacements, the work of internal forces (which is now the variation of total energy) is equal to the work of external forces.

In the case under consideration, both approaches to introducing generalized energy solutions are equivalent. In general, however, this is not so, and the virtual work principle has wider applicability. If $F(x, y)$ is a nonconservative load depending on $u(x, y)$, we cannot use the principle of minimum of total energy; however, (1.14.11) remains valid since it has the mathematical form of the virtual work principle. In what follows, we shall frequently use this principle to pose problems in equation form.

Note that the part of the presentation from (1.14.8) up to (1.14.10) is general and does not depend on the specific form (1.14.11) of the functional $I(u)$. So we can formulate

Theorem 1.14.1. Let u_0 be a minimizer of a functional $I(u) = \|u\|^2 + 2\Phi(u)$ given in an inner product (Hilbert) space H, the functional $\Phi(u)$ being linear and continuous. Then u_0 satisfies (1.14.10) for every $v \in H$.

Equation (1.14.10) is a necessary condition for minimization of the functional $I(u)$, analogous to the condition for functions that the first derivative equal zero at a point of minimum.

We can obtain (1.14.10) formally by evaluating

$$\frac{d}{d\epsilon} I(u_0 + \epsilon v)\Big|_{\epsilon=0} = 0 \tag{1.14.12}$$

(verify). This is valid for the following reason. Given u_0 and v the functional $I(u_0+\epsilon v)$ is an ordinary function of the numerical variable ϵ, and assumes a minimum value at $\epsilon = 0$. The left-hand side of (1.14.12) can be interpreted as a partial derivative at $u = u_0$ in the direction v, and is called the Gâteaux derivative of $I(u)$ at $u = u_0$ in the direction of v. We shall return to this issue later.

The Dirichlet problem for a clamped membrane is a touchstone for all static problems. In a similar way we can introduce a natural notion of generalized solution for other problems under consideration. As we said, each of them can be represented as a problem of a minimum total energy functional of the form (1.14.10) in an energy space. For example, equation (1.14.11), a particular form of (1.14.10) for a clamped membrane, is the same for a free membrane — we need only replace E_{MC} by E_{MF}. $\Phi(u)$, to be a continuous linear functional in E_{MF}, must be supplemented with self balance condition (1.10.13) for the load.

Let us concretize equation (1.14.10) for each of the other problems we have under consideration.

A plate.

The definition of generalized solution $w_0 \in E_P$ is given by the equation

$$\iint\limits_{\Omega} D^{\alpha\beta\gamma\delta}\rho_{\gamma\delta}(w_0)\rho_{\alpha\beta}(w)\,dx\,dy - \iint\limits_{\Omega} F(x,y)w(x,y)\,dx\,dy - \\ - \sum_{k=1}^{m} F_k w(x_k, y_k) - \int_{\gamma} f(s)w(x,y)\,ds = 0 \tag{1.14.13}$$

(see the notation of Section 1.10) which must be valid for every $w \in E_P$. The equation is the same for any kind of homogeneous boundary conditions (i.e., for usual ones) but the energy space will change from one set of boundary conditions to another. If a plate can move as a rigid whole, the requirement that

$$F(x,y) \in L(\Omega), \qquad f(s) \in L(\gamma),$$

for Φ to be a continuous linear functional, must be supplemented with self-balanced conditions for the load:

$$\int_\Omega F(x,y)w_i(x,y)\,dx\,dy + \sum_{k=1}^{m} F_k w_i(x_k,y_k) + \int_\gamma f(s)w(x,y)\,ds = 0 \tag{1.14.14}$$

for $i = 1, 2, 3$, where $w_1(x,y) = 1$, $w_2(x,y) = x$, and $w_2(x,y) = y$.

Note that for each concrete problem we must specify the energy space. The same is true for the following problem.

Linear elasticity.

Here the generalized solution $\mathbf{u} \in E_E$ is defined by the integro-differential equation

$$\int_\Omega c^{ijkl}\epsilon_{kl}(\mathbf{u})\epsilon_{ij}(\mathbf{v})\,d\Omega - \int_\Omega \mathbf{F}(x,y,z)\cdot\mathbf{v}(x,y,z)\,d\Omega - \\ - \int_\Gamma \mathbf{f}(x,y,z)\cdot\mathbf{v}(x,y,z)\,dS = 0 \tag{1.14.15}$$

which must be valid for every $\mathbf{v} \in E_E$.

The load, thanks to Theorem 1.11.4 and Korn's inequality, is of the class

$$F_i(x,y,z) \in L^{6/5}(\Omega), \qquad f_i(x,y,z) \in L^{4/3}(\Gamma), \qquad i = 1,2,3,$$

Γ being a piecewise smooth surface in $\overline{\Omega}$. This provides continuity of $\Phi(w)$.

As above, for a body with free boundary we must require that the load be self-balanced:

$$\int_\Omega \mathbf{F}(\mathbf{x})\,d\Omega + \int_\Gamma \mathbf{f}(\mathbf{x})\,dS = 0, \\ \int_\Omega \mathbf{x}\times\mathbf{F}(\mathbf{x})\,d\Omega + \int_\Gamma \mathbf{x}\times\mathbf{f}(\mathbf{x})\,dS = 0. \tag{1.14.16}$$

We have argued that it is legitimate to introduce the generalized solution in such a way. Of course, full legitimacy will be assured when we prove that this solution exists and is unique in the corresponding space.

We emphasize once more that the definition of generalized solution arose in a natural way from the variational principle of mechanics.

Now we return to general properties of metric spaces.

1.15 Separability

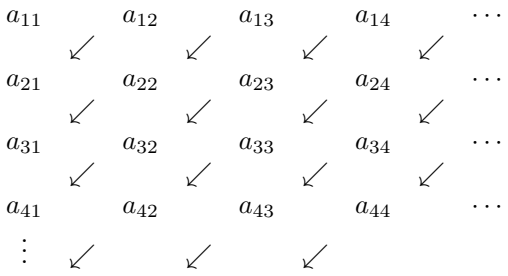
Two sets are said to be *of equal power* if there is a one-to-one correspondence between their elements. Of all sets having infinitely many elements, the set of least power is the set of positive integers.

Definition 1.15.1. A set which is of equal power with the set of positive integers is said to be *countable*.

Roughly speaking, each element of a countable set can be numbered by assigning it a positive-integer index.

Theorem 1.15.1. A countable union of countable sets is countable.

Proof. It suffices to show how we may enumerate the elements of the union. The method is clear from the diagram



where, for a fixed i, $\{a_{ij}\}$ is a sequence of enumerated elements of the ith set. The first element we choose to enumerate is a_{11}; then we go along the diagonal, enumerating a_{12} and a_{21}. The next diagonal gives a_{13}, a_{22}, a_{31}. Proceeding in this way, all the elements of all the sets are put into one-to-one correspondence with the sequence of positive integers. This completes the proof. □

Corollary 1.15.1. The set of all rational numbers is countable.

Proof. A rational number is represented in the form i/j where i, j are integers. Denoting $a_{ij} = i/j$, we obtain the sequence a_{ij} which meets the condition of the theorem. □

Problem 1.15.1. Show that the set of all polynomials with rational coefficients is countable.

Georg Cantor (1845–1918) proved

Theorem 1.15.2. The set of real numbers of the segment $[0, 1]$ is not countable.

(The proof can be found in any textbook on set theory or the theory of functions of a real variable.) So the set $[0, 1]$ is not of equal power with the set of positive integers; the points of $[0, 1]$ form a continuum.

It would be beyond our scope to discuss Cantor's theory of sets. Our interests lie in applying the notion of countability to metric spaces. Modern mechanics depends a great deal on computational ability. A computer can process only finite sets of numbers, hence can only approximate results to a given decimal accuracy. If x is an arbitrary element of an infinite set X

and we want to use a computer to find it, then we must be certain that every element of X can be approximated by elements of another set which is finite or, at least, countable. This leads to

Definition 1.15.2. X is called a *separable* metric space if it contains a countable subset which is dense in X.

In other words, X is separable if there is a countable set $M \subset X$ such that for every $x \in X$ there is a sequence $\{m_i\}$, $m_i \in M$, approximating x: $d(x, m_i) \to 0$ as $i \to \infty$.

An example of a separable metric space is the set of real numbers in $[a, b]$; here the dense set M is the set of all rational numbers in $[a, b]$.

The set of all polynomials on a closed and bounded domain Ω, equipped with the norm of the space $C(\Omega)$, is separable; the dense subset is the set P_r of all polynomials with rational coefficients. Indeed, approximating the coefficients a_α of an arbitrary polynomial $\sum_\alpha a_\alpha \mathbf{x}^\alpha$ by rational numbers a_{α_r} we can always obtain

$$\max_{\mathbf{x}\in\Omega} \left| \sum_\alpha a_\alpha \mathbf{x}^\alpha - \sum_\alpha a_{\alpha_r} \mathbf{x}^\alpha \right| < \varepsilon \qquad (\mathbf{x}^\alpha \equiv x_1^{\alpha_1} \cdots x_n^{\alpha_n})$$

for any given precision $\varepsilon > 0$.

A nontrivial example of a separable space is provided by the classical Weierstrass theorem, which can be formulated as

Theorem 1.15.3. The set P_r of all polynomials with rational coefficients is dense in $C(\Omega)$, where Ω is a closed and bounded domain in $\mathbb{R}^n$.

Since P_r is countable, the theorem states that $C(\Omega)$ is separable. Now we construct an example of a set which is not a separable metric space.

Lemma 1.15.1. The set of functions $f(x)$ bounded on $[0, 1]$ and equipped with the norm

$$\|f(x)\| = \sup_{x\in[0,1]} |f(x)|$$

is not separable.

Proof. It suffices to construct a subset M of the space whose elements cannot be approximated by functions from a countable set. Let α be an arbitrary point of $[0, 1]$. The set M is composed of functions defined as follows:

$$f_\alpha(x) = \begin{cases} 1, & x \geq \alpha, \\ 0, & x < \alpha. \end{cases}$$

The distance from $f_\alpha(x)$ to $f_\beta(x)$ is

$$\|f_\alpha(x) - f_\beta(x)\| = \sup_{x\in[0,1]} |f_\alpha(x) - f_\beta(x)| = 1 \text{ if } \alpha \neq \beta.$$

Take a ball B_α of radius 1/3 about $f_\alpha(x)$. If $\alpha \neq \beta$ then the intersection $B_\alpha \cap B_\beta$ is empty.

If a countable subset is dense in the space then each of the B_α must contain at least one element of this subset, but this contradicts Theorem 1.15.2 since the set of balls B_α is of equal power with the continuum. □

Now we show that the metric spaces we introduced are separable. We begin with

Theorem 1.15.4. Let Ω be a closed and bounded domain in $\mathbb{R}^n$. Then $L^p(\Omega)$ is separable for any $p \geq 1$.

Proof. It suffices to show that P_r, the set of all polynomials with rational coefficients, is dense in $L^p(\Omega)$. We saw (Theorem 1.15.3) that P_r is dense in $C(\Omega)$. Then P_r is dense in the set of all functions continuous on Ω which is equipped with the metric of $L^p(\Omega)$. Indeed, if $f(\mathbf{x}) \in C(\Omega)$ then, for a given $\varepsilon > 0$, we can find a polynomial $Q_\varepsilon(\mathbf{x})$ from P_r such that

$$\max_{\mathbf{x}\in\Omega} |f(\mathbf{x}) - Q_\varepsilon(\mathbf{x})| \leq \frac{\varepsilon}{(\operatorname{mes}\Omega)^{1/p}}.$$

Then

$$\begin{aligned} \|f(\mathbf{x}) - Q_\varepsilon(\mathbf{x})\|_{L^p(\Omega)} &= \left(\int_\Omega |f(\mathbf{x}) - Q_\varepsilon(\mathbf{x})|^p \, d\Omega\right)^{1/p} \\ &\leq \left(\frac{\varepsilon^p}{\operatorname{mes}\Omega} \int_\Omega 1 \, d\Omega\right)^{1/p} \\ &= \varepsilon. \end{aligned}$$

Now let $F(\mathbf{x})$ be an element of $L^p(\Omega)$ and $\{f_n(\mathbf{x})\}$ its representative Cauchy sequence. Each $f_n(\mathbf{x})$ can be approximated by a polynomial from P_r (since $f_n(\mathbf{x})$ is a continuous function) with any accuracy, say $1/n$:

$$\|f_n(\mathbf{x}) - Q_n(\mathbf{x})\|_{L^p(\Omega)} < 1/n.$$

It is easy to verify that $\{Q_n(\mathbf{x})\}$ is also a representative sequence of $F(\mathbf{x})$. By Theorem 1.15.1, the set of all Cauchy sequences constituted of elements of a countable set is also countable. This completes the proof. □

This proof is of a general nature. If P_r is replaced by a countable subset which is dense in a metric space X, and the metric of $L^p(\Omega)$ by the metric of X, the result is an abstract modification of Theorem 1.15.4:

Theorem 1.15.5. The completion of a separable metric space is separable.

This theorem allows us to show separability of energy spaces if we prove the following:

Theorem 1.15.6. Let Ω be compact in $\mathbb{R}^n$. Then for any positive integer k, $C^{(k)}(\Omega)$ is separable.

We give only a sketch of the proof. A function $f(\mathbf{x}) \in C^{(k)}(\Omega)$ can be approximated with any accuracy in the norm of $C^{(k)}(\Omega)$ by a function $f_1(\mathbf{x})$ that is infinitely differentiable. This can be done using the averaging technique. We consider the derivative

$$\frac{\partial^{kn} f_1(x_1, \ldots, x_n)}{\partial x_1^k \cdots \partial x_n^k}$$

as an element of $C(\Omega)$ and, within a prescribed accuracy, approximate it by a polynomial Q_{kn} belonging to P_r. This can be done by virtue of the Weierstrass theorem. Using Q_{kn}, we then construct a polynomial with rational coefficients that approximates $f(\mathbf{x})$ in $C^{(k)}(\Omega)$ within the prescribed accuracy. For this, we choose a point $\mathbf{x}_0 = (x_{10}, \ldots, x_{n0}) \in \Omega$ with rational coordinates. We then choose rational numbers a_α that approximate the values of $D^\alpha f_1(\mathbf{x}_0)$ with some prescribed accuracy. Using these numbers as initial data, we perform successive integrations on the polynomial Q_{kn}:

$$\begin{aligned} Q_{kn-1}(x_1, x_2, \ldots, x_n) &= a_{k-1,k,\ldots,k} + \int_{x_{10}}^{x_1} Q_{kn}(s, x_2, \ldots, x_n)\, ds, \\ Q_{kn-2}(x_1, x_2, \ldots, x_n) &= a_{k-2,k,\ldots,k} + \int_{x_{10}}^{x_1} Q_{kn-1}(s, x_2, \ldots, x_n)\, ds, \\ &\vdots \end{aligned}$$

At each stage of integration we get a polynomial with rational coefficients that approximates in $C(\Omega)$ one of the derivatives of $f_1(\mathbf{x})$ within prescribed accuracy. So the final polynomial approximates $f_1(\mathbf{x})$ and thus $f(\mathbf{x})$ in $C^{(k)}(\Omega)$. We leave it to the reader to complete the proof.

We also need the following almost trivial result:

Theorem 1.15.7. Any subspace E of a separable metric space X is separable.

Proof. Consider a countable set consisting of $(x_1, x_2, \ldots)$ which is dense in X. Let B_{ki} be a ball of radius $1/k$ about x_i. By Theorem 1.15.1, the set of all B_{ki} is countable.

For any fixed k the union $\cup_i B_{ki}$ covers X and thus E. For every B_{ki}, take an element of E which lies in B_{ki} (if it exists). Denote this element by e_{ki}. For any $e \in B_{ki} \cap E$, the distance $d(e, e_{ki})$ is less than $2/k$. It follows that the set of all e_{ki} is, on the one hand, countable, and, on the other hand, dense in E. □

This theorem is of great importance. We have a limited selection of countable sets of functions with which to demonstrate separability of certain

spaces: they are the space P_r of polynomials with rational coefficients, the space of trigonometric polynomials with rational coefficients, and a few others. As a rule, the elements of these spaces do not meet the boundary conditions imposed on functions of, for example, energy spaces. We can circumvent this difficulty by taking a wider space, containing the space of interest, whose separability may be shown. The needed separability is then a consequence of the theorem.

As a particular case of Theorems 1.15.5 and 1.15.6, we have

Lemma 1.15.2. The Sobolev spaces $W^{m,p}(\Omega)$, $p \geq 1$, are separable.

Because we showed that all of the introduced energy spaces were subspaces of certain Sobolev spaces, we have

Lemma 1.15.3. The energy spaces introduced above are all separable.

In what follows, we shall introduce other energy spaces and show that they are also subspaces of Sobolev spaces; hence their separability shall be established.

1.16 Compactness, Hausdorff Criterion

The classical Bolzano theorem states that every bounded sequence of $\mathbb{R}^n$ contains a Cauchy subsequence. How does this property depend on the dimension of a space?

Consider, for example, a sequence of elements of ℓ^2:

$$\begin{aligned} x_1 &= (1,0,0,0,\ldots),\\ x_2 &= (0,1,0,0,\ldots),\\ x_3 &= (0,0,1,0,\ldots),\\ &\vdots \end{aligned}$$

Since $\|x_i\| = 1$ the sequence is bounded in ℓ^2, but for any pair of distinct elements we get

$$\|x_i - x_j\| = (1^2 + 1^2)^{1/2} = \sqrt{2};$$

hence $\{x_i\}$ does not contain a Cauchy subsequence.

Let us introduce

Definition 1.16.1. A set in a metric space is called *precompact* if every sequence consisting of elements of the set contains a Cauchy subsequence. If the limit elements of these subsequences all belong to the set, then the set is called *compact*.

A compact set is closed. In these terms, Bolzano's theorem can be reformulated as follows: a bounded subset of $\mathbb{R}^n$ is precompact; a closed and bounded subset of $\mathbb{R}^n$ is compact.

Note that $\mathbb{R}^n$ itself is not compact if we assume a usual metric. Is there a metric on $\mathbb{R}^n$ such that the whole of $\mathbb{R}^n$ is precompact in this metric? (The answer is "yes." Why?)

To establish a criterion for compactness of a set we need

Definition 1.16.2. A finite set E of elements of a metric space X is called a *finite ε-net* of a set $M \subset X$ if for every $x \in M$ there is an element $e \in E$ such that $d(x, e) < \varepsilon$.

This definition means that every element of M lies in one of a finite collection of balls of radius ε if M has a finite ε-net. It is clear that any finite set has a finite ε-net for any $\varepsilon > 0$. In particular, we may have $M = X$.

Now we formulate Hausdorff's criterion for compactness.

Theorem 1.16.1. A subset of a metric space is precompact if and only if for every $\varepsilon > 0$ there is a finite ε-net for this set.

Proof. (a) *Necessity.* Let M be a precompact subset of a metric space X. The existence of a finite ε-net for any $\varepsilon > 0$ for M is proved by contradiction. Let $\varepsilon_0 > 0$ be such that there is no finite ε_0-net for M: this means that a union of any finite number of balls of radius ε_0 cannot contain all elements of M. Take an element $x_1 \in M$ and a ball B_1 of radius ε_0 about x_1. Since there is no finite ε_0-net for M, there is an element x_2 of M such that $x_2 \notin B_1$. Construct the ball B_2 of radius ε_0 about x_2. Outside $B_1 \cup B_2$ there is a third element x_3 of M — otherwise x_1 and x_2 form an ε_0-net. Continuing to construct a sequence of elements and corresponding balls, we get a sequence $\{x_n\}$ satisfying the condition $d(x_n, x_m) \geq \varepsilon_0$ for $n \neq m$. Therefore $\{x_n\}$ does not contain a Cauchy subsequence, and this contradicts the definition of precompactness.

(b) *Sufficiency.* Suppose that for every $\varepsilon > 0$, there is a finite ε-net of a set $M \subset X$. We must show that M is precompact. Let $\{x_n\}$ be an arbitrary sequence from M: we shall show that we can select a Cauchy subsequence from $\{x_n\}$. For this, take $\varepsilon_1 = 1/2$ and construct a finite ε_1-net for M. One of the balls, say B_1, of radius ε_1 about an element of this finite net must contain an infinite number of elements of $\{x_n\}$. Take one of the latter elements and denote it x_{i_1}. Next construct a finite ε_2-net with $\varepsilon_2 = 1/2^2$. One of the balls, say B_2, of radius ε_2 about the elements of this net contains an infinite number of elements of $\{x_n\}$ which belong to B_1. We choose such an element and denote it x_{i_2}. Continuing this procedure, we obtain an infinite sequence $\{x_{i_k}\}$ which is a subsequence of $\{x_n\}$. Since x_{i_k} and $x_{i_{k+1}}$, by construction, are in the ball B_k of radius $\varepsilon_k = 1/2^k$, we get $d(x_{i_k}, x_{i_{k+1}}) \leq 1/2^{k-1}$. Then the triangle inequality gives

$$\begin{aligned} d(x_{i_k}, x_{i_{k+m}}) &\leq d(x_{i_k}, x_{i_{k+1}}) + d(x_{i_{k+1}}, x_{i_{k+2}}) + \cdots + d(x_{i_{k+m-1}}, x_{i_{k+m}}) \\ &\leq \frac{1}{2^{k-1}} + \frac{1}{2^k} + \cdots + \frac{1}{2^{k+m-2}} < \frac{1}{2^{k-2}}. \end{aligned}$$

This means that $\{x_{i_k}\}$ is a Cauchy sequence, as desired. □

Note that a closed precompact set is compact. How do we formulate the Hausdorff theorem for M to be compact?

Corollary 1.16.1. A precompact subset S of a metric space M is bounded.

Proof. Take $\varepsilon = 1$ and construct a 1-net. The union of the finite number of unit balls with centers at the nodes of the net covers S. So there is a finite ball of M that contains S. □

In a certain way, the property of compactness is close to the property of separability; namely, we have

Theorem 1.16.2. A precompact metric space M (or a precompact subset of a metric space) is separable.

Proof. Using Theorem 1.16.1 we construct a countable set E which is dense in M, as follows. Take the sequence $\varepsilon_k = 1/k$; let the set $(x_{k1}, x_{k2}, \ldots, x_{kN})$ be a finite ε_k-net of M (N certainly depends on k). The collection of all x_{ki} for all possible k, i, being a countable set, is the needed E since, for every $x \in M$ and any $\varepsilon > 0$ there is a ball B of radius $\varepsilon_k < \varepsilon$ about an x_{ki} such that $x \in B$. This means there is a sequence in E which converges to x. □

Now we examine an extension of Bolzano's theorem.

Theorem 1.16.3. Every closed and bounded subset of a Banach space X is compact if and only if X has finite dimension.

(We recall that X has finite dimension if there is a finite set of elements $x_1, \ldots, x_n$ such that any $x \in X$ can be represented in the form $x = \sum_{i=1}^{n} \alpha_i x_i$. The least such n is called the dimension of X.)

Sufficiency of the theorem is proved in a manner similar to that for Bolzano's theorem and is omitted. Necessity is a consequence of the following

Lemma 1.16.1 (Riesz). Let M be a closed subspace of a normed space X and suppose $M \neq X$. Then for any ε, $0 < \varepsilon < 1$, there is an element x_ε such that

$$x_\varepsilon \notin M, \qquad \|x_\varepsilon\| = 1, \qquad \text{and} \qquad \inf_{y \in M} \|y - x_\varepsilon\| > 1 - \varepsilon.$$

Proof. Since $M \neq X$ there is an element $x_0 \in X$ such that $x_0 \notin M$. Denote $d = \inf_{y \in M} \|x_0 - y\|$. First we show that $d > 0$. If on the contrary $d = 0$, then there is a sequence $\{y_k\}$, $y_k \in M$, such that $\|x_0 - y_k\| \to 0$ as $k \to \infty$; this means $\lim_{k \to \infty} y_k = x_0$ and so $x_0 \in M$ because M is closed. Thus $d > 0$. By definition of infimum, for any $\varepsilon > 0$ there exists

$y_\varepsilon \in M$ such that $d \le \|x_0 - y_\varepsilon\| \le d/(1-\varepsilon/2)$. The needed element is $x_\varepsilon = (x_0 - y_\varepsilon)/\|x_0 - y_\varepsilon\|$. Indeed $\|x_\varepsilon\| = 1$ and for any $y \in M$ we have

$$\|x_\varepsilon - y\| = \left\| \frac{x_0 - y_\varepsilon}{\|x_0 - y_\varepsilon\|} - y \right\| = \frac{\|x_0 - (y_\varepsilon + \|x_0 - y_\varepsilon\| y)\|}{\|x_0 - y_\varepsilon\|}$$
$$\ge d \Big/ \frac{d}{1-\varepsilon/2} = 1 - \frac{\varepsilon}{2}.$$

(Here we used the fact that $y_\varepsilon + \|x_0 - y_\varepsilon\| y \in M$.) □

Proof of necessity for Theorem 1.16.3. It suffices to prove that the unit ball about zero is compact only in a finite dimensional Banach space.

Take an element y_1 such that $\|y_1\| = 1$ and denote by E_1 the space spanned by y_1, i.e., the set of all elements of the form αy_1, α being in $\mathbb{C}$. If $E_1 \ne X$ then, by Lemma 1.16.1, there is an element $y_2 \notin E_1$ such that $\|y_2\| = 1$ and $\|y_1 - y_2\| > 1/2$. Denote by E_2 a linear space spanned by y_1 and y_2. If $E_2 \ne X$ then, by the same lemma, we can find y_3 such that

$$\|y_3\| = 1, \qquad \|y_3 - y_1\| > 1/2, \qquad \|y_3 - y_2\| > 1/2.$$

If X is infinite dimensional then this process goes on indefinitely and we get a sequence $\{y_k\}$ such that

$$\|y_i - y_j\| > 1/2 \text{ if } i \ne j.$$

This sequence lying in the unit ball about zero cannot contain a Cauchy subsequence, which contradicts the hypothesis. So the process must terminate and thus $X = E_k$ for some k, i.e., is finite dimensional. □

In the next section we consider a widely applicable theorem on compactness.

1.17 Arzelà's Theorem and Its Applications

Theorem 1.17.1. Let Ω be a closed and bounded (i.e., compact) domain in $\mathbb{R}^n$. A set M of functions continuous on Ω is precompact in $C(\Omega)$ if and only if M satisfies the following pair of conditions:

(i) M is *uniformly bounded*. There is a constant c such that for every $f(\mathbf{x}) \in M$,

$$|f(\mathbf{x})| \le c$$

for all $\mathbf{x} \in \Omega$.

(ii) M is *equicontinuous*. For any $\varepsilon > 0$ there exists $\delta > 0$, dependent on ε, such that whenever $|\mathbf{x} - \mathbf{y}| < \delta$, $\mathbf{x}, \mathbf{y} \in \Omega$, then

$$|f(\mathbf{x}) - f(\mathbf{y})| < \varepsilon$$

holds for every $f(\mathbf{x}) \in M$.

Problem 1.17.1. Does (i) mean that M is bounded in $C(\Omega)$?

Proof of Theorem 1.17.1. (a) *Necessity.* Let M be precompact in $C(\Omega)$. By Theorem 1.16.1 there is a finite 1-net for M; i.e., there is a finite set of continuous functions $g_i(\mathbf{x})$, $i = 1, \dots, k$, such that to any $f(\mathbf{x})$ there corresponds $g_i(\mathbf{x})$ for which

$$\|f(\mathbf{x}) - g_i(\mathbf{x})\| = \max_{\mathbf{x}\in\Omega} |f(\mathbf{x}) - g_i(\mathbf{x})| \le 1.$$

Since the $g_i(\mathbf{x})$ are continuous there is a constant c_1 such that $|g_i(\mathbf{x})| < c_1$ for $i = 1, \dots, k$, and thus

$$|f(\mathbf{x})| \le c_1 + 1.$$

So condition (i) is fulfilled.

Let us show (ii). Let $\varepsilon > 0$ be given. By precompactness of M, there is a finite $\varepsilon/3$-net, say $g_i(\mathbf{x})$, $i = 1, \dots, m$. Since the number of $g_i(\mathbf{x})$ is finite and they are equicontinuous on Ω, we can find a positive number δ such that whenever $|\mathbf{x} - \mathbf{y}| < \delta$ then

$$|g_i(\mathbf{x}) - g_i(\mathbf{y})| < \varepsilon/3 \text{ for all } i = 1, \dots, m.$$

For an arbitrary function $f(\mathbf{x})$ from M, there exists $g_r(\mathbf{x})$ such that

$$|f(\mathbf{x}) - g_r(\mathbf{x})| < \varepsilon/3 \text{ for all } \mathbf{x} \in \Omega.$$

Let $\mathbf{x}, \mathbf{y} \in \Omega$ be such that $|\mathbf{x} - \mathbf{y}| < \delta$. Then

$$\begin{aligned} |f(\mathbf{x}) - f(\mathbf{y})| &\le |f(\mathbf{x}) - g_r(\mathbf{x})| + |g_r(\mathbf{x}) - g_r(\mathbf{y})| + |g_r(\mathbf{y}) - f(\mathbf{y})| \\ &< \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon \end{aligned}$$

and thus condition (ii) is fulfilled too.

(b) *Sufficiency.* Let M satisfy conditions (i) and (ii) of the theorem. We must show that from any sequence of functions lying in M we can choose a subsequence which is uniformly convergent. Since Ω is compact in $\mathbb{R}^n$ we can find a finite δ-net for Ω for any $\delta > 0$, say "cubic" close packing. Take $\delta_k = 1/k$ and construct a corresponding finite δ_k-net for Ω. We enumerate all the nodes of these nets successively: first all points of δ_1-net, then all points of δ_2-net, and so on. As a result, we get a countable set of points $\{\mathbf{x}_k\}$ which is dense in Ω.

Take a sequence of functions $\{f_k(\mathbf{x})\}$ from M and consider it at $\mathbf{x} = \mathbf{x}_1$. We can choose a convergent subsequence $\{f_{k_1}(\mathbf{x}_1)\}$ from it because the numerical sequence $\{f_k(\mathbf{x}_1)\}$ is bounded. Considering the numerical sequence $\{f_{k_1}(\mathbf{x}_2)\}$, by the same reasoning we can choose a convergent subsequence $\{f_{k_2}(\mathbf{x}_2)\}$ from the latter. The same can be done for $\mathbf{x} = \mathbf{x}_3$, $\mathbf{x} = \mathbf{x}_4$, and so on. On the kth step of this procedure we get a subsequence which is convergent (Cauchy) at $\mathbf{x} = \mathbf{x}_i$, $i = 1, \dots, k$.

Consider a sequence consisting of diagonal elements of these sequences, namely, $\{f_{n_n}(\mathbf{x})\}$. By construction this sequence is convergent at every $\mathbf{x} = \mathbf{x}_i$, $i = 1, 2, 3, \ldots$. Let us show that it is uniformly convergent. First, by equicontinuity of M, for any $\varepsilon > 0$ we can find $\delta > 0$ such that whenever $|\mathbf{x} - \mathbf{y}| < \delta$ then for every n

$$|f_{n_n}(\mathbf{x}) - f_{n_n}(\mathbf{y})| < \varepsilon/3.$$

Take some finite δ_1-net of Ω, $\delta_1 < \delta$, with nodes denoted by $\mathbf{z}_i$, $i = 1, \ldots, r$. Since r is finite, for the ε, we can find a number N such that for all $n, m > N$,

$$|f_{n_n}(\mathbf{z}_i) - f_{m_m}(\mathbf{z}_i)| < \varepsilon/3, \quad i = 1, \ldots, r.$$

Let $\mathbf{x}$ be an arbitrary point of Ω and $\mathbf{z}_k$ be the point of the δ_1-net nearest to $\mathbf{x}$, i.e., $|\mathbf{x} - \mathbf{z}_k| < \delta_1$. For the above m, n we get

$$\begin{aligned} |f_{n_n}(\mathbf{x}) - f_{m_m}(\mathbf{x})| &\le |f_{n_n}(\mathbf{x}) - f_{n_n}(\mathbf{z}_k)| + |f_{n_n}(\mathbf{z}_k) - f_{m_m}(\mathbf{z}_k)| + \\ &\quad + |f_{m_m}(\mathbf{z}_k) - f_{m_m}(\mathbf{x})| \\ &< \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon. \end{aligned}$$

Therefore $\{f_{n_n}(\mathbf{x})\}$ is uniformly convergent and the proof is complete. □

Corollary 1.17.1. A set bounded in the space $C^{(1)}(\Omega)$ is precompact in $C(\Omega)$. It is compact if it is closed.

Proof. A set bounded in $C^{(1)}(\Omega)$ is bounded in $C(\Omega)$, i.e., the condition (i) of the theorem is fulfilled. Fulfillment of (ii) follows from the elementary inequality

$$|f(\mathbf{x}) - f(\mathbf{y})| = \left| \int_0^1 \frac{df(s\mathbf{x} + (1-s)\mathbf{y})}{ds}\, ds \right| \le C|\mathbf{x} - \mathbf{y}|.$$

□

A famous application of Arzelà's theorem is the following local existence theorem due to Peano for the Cauchy problem for a system of ordinary differential equations

$$\mathbf{y}' = f(t, \mathbf{y}(t)), \qquad \mathbf{y}(t_0) = \mathbf{y}_0, \qquad \mathbf{y} \in \mathbb{R}^n. \tag{1.17.1}$$

Denote

$$Q(t_0, a, b) = \{(t, \mathbf{y}) \mid t_0 \le t \le t_0 + a,\ |\mathbf{y} - \mathbf{y}_0| \le b,\ \mathbf{y} \in \mathbb{R}^n\}.$$

Theorem 1.17.2 (Peano). Let $f(t, \mathbf{y})$ be continuous on $Q(t_0, a, b)$ and such that $|f(t, \mathbf{y})| \le m$ on this domain, and let $\alpha = \min\{a, b/m\}$. Then there is a continuous solution $\mathbf{y} = \mathbf{y}(t)$ to the Cauchy problem (1.17.1) on the interval $[t_0, t_0 + \alpha]$.

Proof. On $[t_0 - 1, t_0]$ we define a function $\mathbf{y} = \mathbf{y}_\varepsilon(t)$ by

$$\mathbf{y}_\varepsilon(t) = \mathbf{y}_0 + (t - t_0) f(t_0, \mathbf{y}_0).$$

Its continuation onto $[t_0, t_0 + \alpha]$ is determined by the equation

$$\mathbf{y}_\varepsilon(t) = \mathbf{y}_0 + \int_{t_0}^{t} f(s, \mathbf{y}_\varepsilon(s - \varepsilon))\, ds. \tag{1.17.2}$$

The process of continuation is done successively onto the segment $[t_0, t_0 + \alpha_1]$, $\alpha_1 = \min\{\alpha, \varepsilon\}$, then onto $[t_0 + \alpha_1, t_0 + 2\alpha_1]$ and so on.

If $\varepsilon \le b/m$, then on $[t_0 - \varepsilon, t_0]$ we get

$$|\mathbf{y}_\varepsilon(t) - \mathbf{y}_0| \le b.$$

By the conditions of the theorem this inequality also holds on $[t_0, t_0 + \alpha]$. Moreover, on this latter segment $|\mathbf{y}'_\varepsilon(t)| \le m$, i.e., the set of all functions $\{\mathbf{y}_\varepsilon(t)\}$ when $\varepsilon \le b/m$, considered on $[t_0, t_0 + \alpha]$, satisfies the conditions of Corollary 1.17.1. Thus we can find a sequence $\{\varepsilon_k\}$ such that $\varepsilon_k \to 0$ as $k \to \infty$ and the sequence $\{\mathbf{y}_{\varepsilon_k}(t)\}$ converges uniformly, $\mathbf{y}(t) = \lim_{k\to\infty} \mathbf{y}_{\varepsilon_k}(t)$ on $[t_0, t_0 + \alpha]$.

By equicontinuity of $f(t, \mathbf{y})$, the sequence $\{f(t, y_{\varepsilon_k}(t)\}$ converges uniformly to $f(t, \mathbf{y}(t))$ on the same segment. Therefore we can take the limit under the integral sign on the right-hand side of (1.17.2); passage to the limit in (1.17.2) gives

$$\mathbf{y}(t) = \mathbf{y}_0 + \int_{t_0}^{t} f(s, \mathbf{y}(s))\, ds, \qquad t \in [t_0, t_0 + \alpha].$$

This means $\mathbf{y}(t)$ is a continuous solution to the problem (1.17.1). □

Our subject is now to demonstrate how compactness can be applied to the justification of a finite difference procedure to solve ordinary differential equations. We consider the simplest of these methods, due to Euler, supposing the requirements of Theorem 1.17.2 to be fulfilled.

Let $\Delta > 0$ be a step of Euler's method for the solution of (1.17.1). Euler's method is defined by the system of equations

$$\begin{aligned} \frac{\mathbf{z}_{k+1} - \mathbf{z}_k}{\Delta} &= f(t_0 + k\Delta, \mathbf{z}_k), \quad k = 0, 1, \dots \\ \mathbf{z}_0 &= \mathbf{y}_0. \end{aligned} \tag{1.17.3}$$

Denote by $\mathbf{y} = \mathbf{y}_\Delta(t)$ the linear interpolation function of the set of pairs $(t_0 + k\Delta, \mathbf{z}_k)$; for $0 \le s \le \Delta$, it is defined by

$$\mathbf{y}_\Delta(t_0 + k\Delta + s) = \frac{\Delta - s}{\Delta}\mathbf{z}_k + \frac{s}{\Delta}\mathbf{z}_{k+1}.$$

On each of the segments $[t_0 + k\Delta, t_0 + (k+1)\Delta]$, $k < \alpha/n$, we get the relation

$$|\mathbf{y}'_\Delta(t)| = |f(t_0 + k\Delta, \mathbf{z}_k)| \le m \tag{1.17.4}$$

from which it follows that

$$|\mathbf{z}_k - \mathbf{z}_0| \equiv |\mathbf{z}_k - \mathbf{y}_0| \le k\Delta m \le \alpha m \le b$$

and thus the system (1.17.3) is solvable until $k < \alpha/\Delta$. Therefore, (1.17.4) holds for all $t \in [t_0, t_0 + \alpha]$ (except t of the form $t_0 + k\Delta$ which, however, does not break the fulfillment of the conditions of Arzelà's theorem) and

$$|\mathbf{y}_\Delta(t) - \mathbf{y}_0| \le b \qquad \text{when} \qquad t \in [t_0, t_0 + \alpha].$$

So we see that the set of all functions $\{\mathbf{y}_\Delta(t)\}$ is precompact in $C(t_0, t_0+\alpha)$ if $\Delta \le \alpha$ since both conditions of Arzelà's theorem are fulfilled. Now we can formulate

Theorem 1.17.3. Assume that all conditions of Theorem 1.17.2 are fulfilled and the Cauchy problem (1.17.1) has the unique solution $\mathbf{y} = \mathbf{y}(t)$ in $Q(t_0, a, b)$. Then a sequence $\{\mathbf{y}_{\Delta_k}(t)\}$ converges uniformly to $\mathbf{y} = \mathbf{y}(t)$ on $[t_0, t_0 + \alpha]$ as $\Delta_k \to 0$.

Proof. It suffices to show that for any $\varepsilon > 0$ there is only a finite number of functions from the sequence $\{\mathbf{y}_{\Delta_k}(t)\}$ which do not satisfy

$$|\mathbf{y}_{\Delta_k}(t) - \mathbf{y}(t)| \le \varepsilon, \text{ for all } t \in [t_0, t_0 + \alpha]. \tag{1.17.5}$$

Assume to the contrary that there are infinitely many functions from the sequence which do not satisfy (1.17.5) for some $\varepsilon > 0$. Then, using a standard technique from calculus, we can find a point t_1, $t_1 \in [t_0, t_0 + \alpha]$, and a subsequence $\Delta_{k_1} \to 0$ such that

$$|\mathbf{y}_{\Delta_k}(t_1) - \mathbf{y}(t_1)| > \varepsilon$$

and the number sequence $\{\mathbf{y}_{\Delta_{k_1}}(t)\}$ is convergent.

As the set $\{\mathbf{y}_{\Delta_{k_1}}(t)\}$ is precompact in $C(t_0, t_0+\alpha)$ we can choose from it a subsequence $\{\mathbf{y}_{\Delta_{k_2}}(t)\}$ uniformly converging to $\mathbf{y} = \mathbf{z}(t)$; it is clear that

$$\mathbf{z}(t) \not\equiv \mathbf{y}(t). \tag{1.17.6}$$

Rewriting (1.17.3) in the form

$$\frac{d\mathbf{y}_\Delta(t_0 + k\Delta + s)}{ds} = f(t_0 + k\Delta, \mathbf{y}_\Delta(t_0 + k\Delta)), \qquad 0 \le s \le \Delta$$

and integrating this with regard to (1.17.3), we get

$$\mathbf{y}_\Delta(t) - \mathbf{y}_0 = \Delta \sum_{i=0}^{k-1} f(t_0 + i\Delta, \mathbf{y}_\Delta(t_0 + i\Delta)) + sf(t_0 + k\Delta, \mathbf{y}_\Delta(t_0 + k\Delta)), \tag{1.17.7}$$

where $t = t_0 + k\Delta + s$, $0 \le s \le \Delta$. The expression on the right-hand side of (1.17.7) is a finite Riemann sum which, under the present conditions, converges as $\Delta = \Delta_{k2} \to 0$ to the integral

$$\int_{t_0}^{t} f(s, \mathbf{z}(s))\, ds$$

and, therefore, $\mathbf{z}(s)$ satisfies the equation

$$\mathbf{z}(t) - \mathbf{y}_0 = \int_{t_0}^{t} f(s, \mathbf{z}(s))\, ds$$

which is equivalent to the Cauchy problem (1.17.1). By uniqueness of solution of this problem we have $\mathbf{z}(t) = \mathbf{y}(t)$, which contradicts (1.17.6). □

What can we say about convergence of the first derivatives of $\{\mathbf{y}_\Delta(t)\}$? At nodes of the Δ-net $\mathbf{y}_\Delta(t)$ has no derivative, but it has a right-hand derivative at every point (which is discontinuous at nodes) and it can be shown that the sequence of right-hand derivatives of $\{\mathbf{y}_{\Delta_k}(t)\}$ converges uniformly on $[t_0, t_0 + \alpha)$.

The Euler finite-difference procedure is not used for computer solution of differential equations but there are various finite difference methods, frequently used, for which the problem of convergence is open. A harder question is the justification of a finite difference procedure in a boundary value problem, as it is connected with solvability of the problem and uniqueness of solution. Boundary value problems for partial differential equations and systems are of great interest with respect to the application of finite difference methods, but justification of this applicability is in large part an open problem. Most of the achievements here are for so-called variational difference methods which are close to the finite element method; to justify them one uses the modified technique of energy spaces which is under consideration in this book.

Finally, we state (without proof) the criterion for compactness in $L^p(\Omega)$ where Ω is a closed and bounded domain in $\mathbb{R}^n$.

Theorem 1.17.4. A set M of elements of $L^p(\Omega)$, $1 < p < \infty$, is precompact in $L^p(\Omega)$ if and only if M satisfies the following pair of conditions:

(i) M is bounded in $L^p(\Omega)$ (i.e., there is a constant m such that for every function $f(\mathbf{x})$ from M we have $\|f(\mathbf{x})\|_{L^p(\Omega)} \le m$);

(ii) M is equicontinuous in $L^p(\Omega)$ (i.e., for any $\varepsilon > 0$ we can find $\delta > 0$, dependent on ε, such that whenever $|\boldsymbol{\Delta}| < \delta$ then, for every $f(\mathbf{x}) \in M$, $\|f(\mathbf{x}+\boldsymbol{\Delta}) - f(\mathbf{x})\|_{L^p(\Omega)} < \varepsilon$. Here $f(\mathbf{x})$ is extended by zero outside Ω.)

The proof can be found in Kantorovich [16].

1.18 The Theory of Approximation in a Normed Space

In what follows, we shall consider some problems of minimum of a functional — as a rule, a functional of energy. We begin with one of the simple problems of this class: the so-called general problem of approximation in a normed space X. It can be stated as follows:

> Given $x \in X$ and elements $g_1, g_2, \ldots, g_n$ with each $g_i \in X$, find numbers $\lambda_1, \lambda_2, \ldots, \lambda_n$ such that the value of the function
>
> $$\phi(\lambda_1, \lambda_2, \ldots, \lambda_n) = \|x - \lambda_1 g_1 - \lambda_2 g_2 - \cdots - \lambda_n g_n\|$$
>
> is minimal.

We shall write down this and similar problems in the form

$$\phi(\lambda_1, \ldots, \lambda_n) \to \min_{\lambda_1, \ldots, \lambda_n} .$$

This form is shared by the problem of best approximation of a continuous function by a polynomial of nth order, or by a trigonometric polynomial, or by other special functions. Its solution depends on the norm which is used to pose the problem.

We suppose that $g_1, g_2, \ldots, g_n$ are linearly independent. This of course means that from the equation

$$\lambda_1 g_1 + \lambda_2 g_2 + \cdots + \lambda_n g_n = 0$$

it follows that $\lambda_1 = \lambda_2 = \cdots = \lambda_n = 0$.

Denote by X_n the linear subspace of X spanned by $g_1, g_2, g_3, \ldots, g_n$.

Theorem 1.18.1. For any $x \in X$ there exists x^*, dependent on x, such that $x^* = \sum_{i=1}^n \lambda_i^* g_i$ and

$$\|x - x^*\| = \inf_{\lambda_1, \ldots, \lambda_n} \left\| x - \sum_{i=1}^n \lambda_i g_i \right\| . \tag{1.18.1}$$

Proof. Consider $\phi(\lambda_1, \ldots, \lambda_n)$ as a function of n variables. The continuity on $\mathbb{R}^n$ (or $\mathbb{C}^n$ if X is a complex space) of this function, and of the function

$$\psi(\lambda_1, \ldots, \lambda_n) = \left\| \sum_{i=1}^n \lambda_i g_i \right\|$$

follows from the chain of inequalities

$$\begin{aligned}
&|\phi(\lambda_1+\Delta_1,\lambda_2+\Delta_2,\ldots,\lambda_n+\Delta_n)-\phi(\lambda_1,\lambda_2,\ldots,\lambda_n)|\\
&\quad=\left|\left\|x-\sum_{i=1}^n(\lambda_i+\Delta_i)g_i\right\|-\left\|x-\sum_{i=1}^n\lambda_i g_i\right\|\right|\\
&\quad\le\left\|\left[x-\sum_{i=1}^n(\lambda_i+\Delta_i)g_i\right]-\left[x-\sum_{i=1}^n\lambda_i g_i\right]\right\|\\
&\quad=\left\|\sum_{i=1}^n\Delta_i g_i\right\|\le\sum_{i=1}^n|\Delta_i|\,\|g_i\|.
\end{aligned}$$

(Here we have used

$$\|x-y\|\ge|\|x\|-\|y\||, \tag{1.18.2}$$

a consequence of the triangle inequality.) By continuity $\psi(\lambda_1,\ldots,\lambda_n)$ assumes a minimum value on the sphere $\sum_{i=1}^n|\lambda_i|^2=1$ at some point $\lambda_{10},\ldots,\lambda_{n0}$ with $\sum_{i=1}^n|\lambda_{i0}|^2=1$. As the set $(g_1,\ldots,g_n)$ is linearly independent we get

$$\left\|\sum_{i=1}^n\lambda_{i0}g_i\right\|=\min_{\sum_{i=1}^n|\lambda_i|^2=1}\left\|\sum_{i=1}^n\lambda_i g_i\right\|=d>0.$$

By (1.18.2),

$$\phi(\lambda_1,\ldots,\lambda_n)\ge\left\|\sum_{i=1}^n\lambda_i g_i\right\|-\|x\|$$

and, on the domain

$$\left(\sum_{i=1}^n|\lambda_i|^2\right)^{1/2}\ge 3\|x\|/d,$$

we have

$$\phi(\lambda_1,\ldots,\lambda_n)\ge(3\|x\|/d)d-\|x\|=2\|x\|.$$

Since $\phi(0,0,\ldots,0)=\|x\|$, we find that $\phi(\lambda_1,\lambda_2,\ldots,\lambda_n)$ has a minimal value at a point $(\lambda_1^*,\lambda_2^*,\ldots,\lambda_n^*)$ of the ball $\left(\sum_{i=1}^n|\lambda_i|^2\right)^{1/2}\le 3\|x\|/d$. □

When we referred to this general problem of approximation as relatively simple, we meant that it was simple in principle — not that its concrete applications are simple.

What about uniqueness of the best approximation in a normed space? This is attained in spaces of the following type:

Definition 1.18.1. A normed space is called *strictly normed* if from the equality

$$\|x+y\|=\|x\|+\|y\|,\quad x\ne 0,$$

it follows that $y=\lambda x$ and $\lambda\ge 0$.

The spaces ℓ^p, $L^p(\Omega)$, $W^{k,p}(\Omega)$, when $1 < p \le \infty$, are strictly normed. This follows from the properties of the Minkowski inequality (see, for example, Hardy et.al. [12]).

We establish a uniqueness theorem under more general conditions than the existence theorem. For this, we introduce

Definition 1.18.2. A set S which contains the whole segment

$$\lambda x + (1-\lambda)y, \qquad 0 \le \lambda \le 1,$$

for every pair of elements $(x, y) \in S$ is said to be *convex*.

Theorem 1.18.2. For any element x of a strictly normed space X there is at most one element of a closed convex set $M \subset X$ which minimizes the functional $F(y) = \|x - y\|$ on the set M.

Proof. Suppose to the contrary that there are two minimizers y_1 and y_2 of $F(y)$:

$$\|x - y_1\| = \|x - y_2\| = \inf_{y \in M} \|x - y\| \equiv d. \tag{1.18.3}$$

If $x \in M$ then $x = y_1 = y_2$. Let $x \notin M$ so that $d > 0$. By convexity of M, an element $(y_1 + y_2)/2$ belongs to M. So $\|x - (y_1 + y_2)/2\| \ge d$. On the other hand, we have

$$\left\| x - \frac{y_1 + y_2}{2} \right\| = \left\| \frac{x - y_1}{2} + \frac{x - y_2}{2} \right\| \le \frac{1}{2}\|x - y_1\| + \frac{1}{2}\|x - y_2\| = d.$$

Therefore

$$\left\| x - \frac{y_1 + y_2}{2} \right\| = \left\| \frac{x - y_1}{2} \right\| + \left\| \frac{x - y_2}{2} \right\|.$$

As X is strictly normed, it follows that

$$x - y_1 = \lambda(x - y_2), \qquad \lambda \ge 0,$$

so $\|x - y_1\| = \lambda\|x - y_2\|$. From (1.18.3) we get $\lambda = 1$, hence $y_1 = y_2$. □

Lemma 1.18.1. An inner product space is strictly normed.

Proof. Let $\|x+y\| = \|x\| + \|y\|$, $x \ne 0$. Then $\|x+y\|^2 = (\|x\| + \|y\|)^2$. This can be rewritten (for a complex space) in the form

$$\|x\|^2 + 2\,\mathrm{Re}(x, y) + \|y\|^2 = \|x\|^2 + 2\|x\|\,\|y\| + \|y\|^2 \tag{1.18.4}$$

so $\mathrm{Re}(x, y) = \|x\|\,\|y\|$. By the Schwarz inequality, we obtain $\mathrm{Im}(x, y) = 0$ and thus $(x, y) = \|x\|\,\|y\|$ (in the real case, this equality comes directly from (1.18.4)). By Theorem 1.9.1 we have $y = \lambda x$ and, placing this into the last equality, $\lambda \ge 0$. □

In a Hilbert space, we can combine Theorems 1.18.2 and 1.18.1 as follows:

Theorem 1.18.3. For any element x of a Hilbert space H, there is a unique element of a closed convex set M which is a minimizer of the functional $F(y) = \|x - y\|$ on M.

Proof. Uniqueness was proved in Theorem 1.18.2. We show existence of a minimizer. Let $\{y_k\}$ be a minimizing sequence of $F(y)$, i.e.,

$$\lim_{k\to\infty} F(y_k) = \lim_{k\to\infty} \|x - y_k\| = \inf_{y\in M} \|x - y\|.$$

(By definition of infimum, such a sequence exists.) As M is closed it suffices to show that $\{y_k\}$ is a Cauchy sequence. For this, we write down the parallelogram equality for a pair $x - y_i$ and $x - y_j$:

$$\|2x - y_i - y_j\|^2 + \|y_i - y_j\|^2 = 2\left(\|x - y_i\|^2 + \|x - y_j\|^2\right),$$

so

$$\|y_i - y_j\|^2 = 2\left(\|x - y_i\|^2 + \|x - y_j\|^2\right) - 4\left\|x - \frac{y_i + y_j}{2}\right\|^2. \qquad (1.18.5)$$

Since $\|x - y_j\|^2 = d^2 + \varepsilon_j$, $\varepsilon_j \to 0$ as $j \to \infty$, from (1.18.5) it follows that

$$\|y_i - y_j\|^2 \le 2(d^2 + \varepsilon_i + d^2 + \varepsilon_j) - 4d^2 = 2(\varepsilon_i + \varepsilon_j) \to 0$$

as $i, j \to \infty$. □

All requirements of Theorem 1.18.3 are fulfilled if M is a closed linear subspace of H. But this case is so important that we treat it separately as follows.

1.19 Decomposition Theorem, Riesz Representation

Let x be an arbitrary element of a Hilbert space H, M be a closed linear subspace of H, and m be the unique (by Theorem 1.18.3) minimizer of $F(y)$ on M:

$$\|x - m\| = \inf_{y\in M} \|x - y\|.$$

Taking an element v of M, we consider a real-valued function $f(\alpha) = \|x - m - \alpha v\|^2$ of a real variable α. This function takes its minimum value at $\alpha = 0$, so

$$\frac{df}{d\alpha}\Big|_{\alpha=0} = 0.$$

Direct calculation gives

$$\frac{df}{d\alpha}\Big|_{\alpha=0} = \frac{d}{d\alpha}(x - m - \alpha v, x - m - \alpha v)\Big|_{\alpha=0} = -2\,\mathrm{Re}(x - m, v) = 0.$$

Replacing v by iv, we get $\mathrm{Im}(x-m,v)=0$ so that

$$(x-m,v)=0. \tag{1.19.1}$$

It follows that $x-m$ is orthogonal to every $v\in M$.

Definition 1.19.1. An element n of a Hilbert space H is said to be *orthogonal to* M, a subspace of H, if n is orthogonal to every element of M. Two subspaces M and N of H are *mutually orthogonal* (denoted $M\perp N$) if any $n\in N$ is orthogonal to M and any $m\in M$ is orthogonal to N.

We say that there is an *orthogonal decomposition* of a Hilbert space H into M and N if M and N are mutually orthogonal subspaces of H and any element $x\in H$ can be uniquely represented in the form

$$x=m+n, \qquad m\in M, n\in N. \tag{1.19.2}$$

Now we can state the above result as the so-called *decomposition theorem* for a Hilbert space.

Theorem 1.19.1. Assume that M is a closed subspace of a Hilbert space H. Then there is a closed subspace N of H that is orthogonal to M and such that H can be uniquely decomposed into the orthogonal sum of M and N, i.e., any $x\in H$ can be uniquely represented in the form (1.19.2).

Proof. Denote by N the set of all elements of H such that any $n\in N$ is orthogonal to every $m\in M$ (we suppose that $M\neq H$). As was shown, N is not empty. It is seen that N is a subspace of H; indeed, if $n_1,n_2\in N$, i.e.,

$$(n_1,m)=(n_2,m)=0 \text{ for every } m\in M$$

then $(\lambda_1 n_1+\lambda_2 n_2,m)=0$ for any numbers λ_1,λ_2 and any $m\in M$. Moreover, N is closed: every Cauchy sequence $\{n_k\}$, $n_k\in N$, has a limit element $y=\lim_{k\to\infty} n_k$ and

$$(y,m)=\lim_{k\to\infty}(n_k,m)=0 \text{ for all } m\in M$$

so we have $y\in N$.

At the beginning of this section, we constructed for an arbitrary element $x\in H$ its projection m on M in such a way that $n=x-m$ is, by (1.19.1), orthogonal to M. So the representation (1.19.2) is proven. It remains to show uniqueness of this decomposition. Assume that for some x there are two such representations:

$$x=m_1+n_1 \text{ and } x=m_2+n_2, \qquad m_i\in M, n_i\in N.$$

Then $m_1+n_1=m_2+n_2$ or

$$m_1-m_2=n_1-n_2.$$

Multiplying both sides of this equality by $m_1 - m_2$ and then by $n_1 - n_2$ in H, we get $\|m_1 - m_2\|^2 = 0$ and $\|n_1 - n_2\|^2 = 0$. This completes the proof. □

This theorem has widespread applications. One of them is the following *Riesz representation theorem* in a Hilbert space, which is of great importance in what follows.

Theorem 1.19.2. Let $F(x)$ be a continuous linear functional given on a Hilbert space H. There is a unique element $f \in H$ such that

$$F(x) = (x, f) \quad \text{for every } x \in H. \tag{1.19.3}$$

Moreover, $\|F\| = \|f\|$.

Proof. Consider the so-called kernel of F, defined as the set M of all elements satisfying the equation

$$F(x) = 0. \tag{1.19.4}$$

By linearity of $F(x)$, any finite linear combination $\sum_{i=1}^{n} \lambda_i m_i$ of elements m_i from M belongs to M; if $\{m_k\}$, $m_k \in M$, is a Cauchy sequence in H, then by continuity of $F(x)$ we see that $y = \lim_{k\to\infty} m_k$ satisfies (1.19.4), i.e., $y \in M$. Therefore M is a closed subspace of H.

By Theorem 1.19.1 there is a closed subspace N of H which is orthogonal to M, and any element $x \in H$ can be uniquely represented in the form $x = m + n$ where $m \in M$, $n \in N$, and $(m, n) = 0$.

Let us show that N is one-dimensional, i.e., that any of its elements n has the form $n = \alpha n^*$ where n^* is a fixed element of N. Let n_1 and n_2 be two arbitrary elements of N. Then the element $n_3 = F(n_1)n_2 - F(n_2)n_1$ belongs to N. On the other hand

$$F(n_3) = F(n_1)F(n_2) - F(n_2)F(n_1) = 0,$$

which means that $n_3 \in M$. So $n_3 = 0$, hence N is one-dimensional.

Take an element $n \in N$ and define n_0 by

$$n_0 = n/\|n\|.$$

Any element of H can be represented as

$$x = m + \alpha n_0, \qquad m \in M,$$

where $\alpha = (x, n_0)$. It follows that

$$\begin{aligned} F(x) &= F(m + \alpha n_0) = F(m) + \alpha F(n_0) = \alpha F(n_0) \\ &= F(n_0)(x, n_0) = (x, \overline{F(n_0)} n_0). \end{aligned}$$

Denoting $\overline{F(n_0)}n_0$ by f, we obtain the needed representation (1.19.3).

Suppose there are two representers f_1 and f_2, i.e.,

$$F(x) = (x, f_1) = (x, f_2).$$

The choice $x = f_1 - f_2$ in the last equality gives $\|f_1 - f_2\|^2 = 0$, hence $f_1 = f_2$. Finally, the equality $\|F\| = \|f\|$ follows from the definition of $\|F\|$ and the Schwarz inequality. □

This proof is carried out in a complex Hilbert space, but remains valid for a real Hilbert space. The sense of the theorem is that any continuous linear functional given on a Hilbert space can be uniquely identified with an element of the same space. Since for a fixed $f \in H$ the inner product (x, f) is a continuous linear functional on H, we have a one-to-one correspondence between the set H' of all continuous linear functionals given on a Hilbert space H (called the *conjugate space* to H) and H itself.

Now let us consider some applications of the Riesz representation theorem. One of them is the *Lax–Milgram theorem*:

Theorem 1.19.3. Let $a(u, v)$ be a bilinear form in $u, v \in H$, a real Hilbert space (i.e., $a(u, v)$ is linear in each variable u, v), such that for all $u, v \in H$

(i) $|a(u, v)| \le M\|u\|\,\|v\|$,

(ii) $|a(u, u)| \ge \alpha\|u\|^2$,

with positive constants M, α that do not depend on u, v. Then there exists a continuous linear operator A having the properties

(i) the range of A is the whole of H;

(ii) A has a continuous inverse and $\|A^{-1}\| \le 1/\alpha$;

(iii) for all $u, v \in H$,

$$a(u, v) = (Au, v). \tag{1.19.5}$$

Proof. Fix $u \in H$. By (i), the form $a(u, v)$ is a linear continuous functional in v. By the Riesz representation theorem, there is a uniquely defined element g such that $a(u, v) = (v, g) = (g, v)$. Element u defines g uniquely, which means that there is a correspondence $u \mapsto g$ defining an operator A: $g = Au$. Thus we have established the representation (1.19.5).

Let us prove the stated properties of this operator. The linearity of $a(u, v)$ in u implies the linearity of A.

By (i) of the theorem, we have

$$|a(u, v)| = |(Au, v)| \le M\|u\|\,\|v\|.$$

Putting $v = Au$ we get

$$|(Au, Au)| = \|Au\|^2 \le M\|u\|\,\|Au\|$$

and so $\|Au\| \le M\|u\|$. This means that A is continuous.

By the Schwarz inequality and (ii) of the theorem, it follows that

$$\|Au\|\,\|u\| \ge |(Au,u)| \ge \alpha\|u\|^2$$

so

$$\|Au\| \ge \alpha\|u\|. \tag{1.19.6}$$

This means that on the range $R(A)$ operator A has a continuous inverse A^{-1} such that $\|A^{-1}\| \le 1/\alpha$.

It remains to demonstrate that $R(A) = H$. First of all, $R(A)$ is a closed subspace of H. Indeed, if $\{Ax_n\}$ is a Cauchy sequence then, by (1.19.6), $\{x_n\}$ is also a Cauchy sequence converging to x_0. By continuity of A we have $Ax_n \to Ax_0$, hence $R(A)$ is closed. If $R(A) \ne H$, then there is an element $v_0 \ne 0$ such that $v_0 \perp R(A)$, which means that $(Au, v_0) = 0$ for all $u \in H$. In particular $(Av_0, v_0) = 0$ so, by (ii), we get $v_0 = 0$. This is a contradiction, and the proof is completed. □

Corollary 1.19.1. Suppose $\Phi(v)$ is a continuous linear functional in H, and $a(u,v)$ is the bilinear form of the theorem. Then there is a unique element $u_0 \in H$ which satisfies the equation

$$a(u_0, v) = \Phi(v) \tag{1.19.7}$$

for all $v \in H$.

Proof. By the Riesz representation theorem, there is a unique representation $\Phi(v) = (v,f) = (f,v)$. So, by the theorem, we can rewrite equation (1.19.7) as $(Au, v) = (f, v)$. Since v is arbitrary it follows that $Au = f$. Denoting $u_0 = A^{-1}f$, we define the element with the needed properties. □

The Lax–Milgram theorem is used traditionally to demonstrate existence and uniqueness of weak solutions of boundary value problems. To establish the same theorems we shall use another approach that is based on the Riesz representation theorem and energy norming of spaces. Both approaches are equivalent, but we consider the energy approach to be preferable as it relates with the nature of the problems more deeply.

1.20 Existence of Energy Solutions to Some Mechanics Problems

We recall that in Section 1.14 we introduced generalized solutions for several mechanics problems and reduced those problems to a problem of finding a solution to the abstract equation

$$(u,v) + \Phi(v) = 0 \tag{1.20.1}$$

on an energy (Hilbert) space. We obtained some restrictions on the forces to provide continuity of the linear functional $\Phi(v)$ in the energy space. The following theorem guarantees solvability of those mechanics problems in a generalized sense.

Theorem 1.20.1. Assume $\Phi(v)$ is a continuous linear functional given on a Hilbert space H. Then there is a unique element $u \in H$ that satisfies (1.20.1) for every $v \in H$.

Proof. By the Riesz representation theorem there is a unique $u_0 \in H$ such that the continuous linear functional $\Phi(v)$ is represented in the form $\Phi(v) = (v, u_0) \equiv (u_0, v)$, and so (1.20.1) takes the form

$$(u, v) + (u_0, v) = 0. \tag{1.20.2}$$

We need to find $u \in H$ which satisfies (1.20.2) for every $v \in H$. Rewriting it in the form

$$(u + u_0, v) = 0,$$

we see that its unique solution is $u = -u_0$. □

This theorem answers the question of solvability, in the generalized sense, of the problems of Section 1.14. To demonstrate this, we rewrite Theorem 1.20.1 in concrete terms for a pair of problems.

Theorem 1.20.2. Assume that

$$F(x, y) \in L(\Omega), \qquad f(x, y) \in L(\gamma),$$

Ω being compact in $\mathbb{R}^2$ and γ being a piecewise smooth curve in Ω. Then the problem of equilibrium of a plate with clamped edge has a unique generalized solution, namely, there is a unique $w_0 \in E_{PC}$ which satisfies (1.14.13) for every $w \in E_{PC}$.

Changes for a plate which is free of clamping are evident: it is necessary to add the self-balance condition (1.14.14) for forces, and the space E_{PC} in the statement must be replaced by E_{PF}.

Theorem 1.20.3. Assume that all Cartesian components of the volume forces $\mathbf{F}(x, y, z)$ are in $L^{6/5}(\Omega)$, and that those of the surface forces $\mathbf{f}(x, y, z)$ are in $L^{4/3}(S)$, where Ω is compact in $\mathbb{R}^3$ and S is a piecewise smooth surface in Ω. Then the problem of equilibrium of an elastic body occupying Ω, with clamped boundary, has a unique generalized solution $\mathbf{u} \in E_{EC}$; namely, $\mathbf{u}(x, y, z)$ satisfies (1.14.15) for every $\mathbf{v} \in E_{EC}$.

In both theorems, restrictions on the load are to provide continuity of the corresponding functionals Φ, the work of external forces.

Problem 1.20.1. Formulate existence theorems for other problems of Section 1.14.

Now let us consider another application of the Riesz representation theorem.

A generalized solution to the eigenvalue problem for a clamped membrane is defined by the integro-differential equation

$$\int_\Omega \left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y} \right) dx\, dy = \lambda \int_\Omega uv\, dx\, dy \tag{1.20.3}$$

and is stated as follows: find an element $u \in E_{MC}$ and a corresponding number λ such that $u \neq 0$ satisfies (1.20.3) for every $v \in E_{MC}$.

First we reformulate this eigenvalue problem in the form of the operator equation

$$u = \lambda K u \tag{1.20.4}$$

in the space E_{MC}. For this, consider the term

$$F(v) = \int_\Omega uv\, dx\, dy$$

as a functional in E_{MC} with respect to v when u is a fixed element of E_{MC}. It is seen that $F(v)$ is a linear functional. By the Schwarz inequality and the Friedrichs inequality we get

$$\begin{aligned} |F(v)| &= \left| \int_\Omega uv\, dx\, dy \right| \\ &\leq \left(\int_\Omega u^2\, dx\, dy \right)^{1/2} \left(\int_\Omega v^2\, dx\, dy \right)^{1/2} \\ &\leq m\|u\|\,\|v\| \\ &= m_1\|v\| \end{aligned} \tag{1.20.5}$$

(hereafter the norm $\|\cdot\|$ and the scalar product are taken in E_{MC}) where m is a constant defined by the Friedrichs inequality; hence $F(v)$ is a continuous linear functional acting in the Hilbert space E_{MC}. By the Riesz representation theorem, $F(v)$ has the unique representation

$$F(v) \equiv \int_\Omega uv\, dx\, dy = (v, f) = (f, v). \tag{1.20.6}$$

What do we have? For every $u \in E_{MC}$, by this representation, there is a unique element $f \in E_{MC}$, hence the correspondence $u \mapsto f$ is an operator $f = K(u)$ from E_{MC} to E_{MC}.

Let us show some properties of this operator. First we show that it is linear. Let

$$f_1 = K(u_1), \qquad f_2 = K(u_2).$$

Then

$$\int_\Omega (\lambda_1 u_1 + \lambda_2 u_2) v\, dx\, dy = (K(\lambda_1 u_1 + \lambda_2 u_2), v)$$

while on the other hand,

$$\begin{aligned}\int_\Omega (\lambda_1 u_1 + \lambda_2 u_2) v \, dx \, dy &= \lambda_1 \int_\Omega u_1 v \, dx \, dy + \lambda_2 \int_\Omega u_2 v \, dx \, dy \\ &= \lambda_1 (K(u_1), v) + \lambda_2 (K(u_2), v) \\ &= (\lambda_1 K(u_1) + \lambda_2 K(u_2), v).\end{aligned}$$

Combining these we have

$$(K(\lambda_1 u_1 + \lambda_2 u_2), v) = (\lambda_1 K(u_1) + \lambda_2 K(u_2), v),$$

hence $K(\lambda_1 u_1 + \lambda_2 u_2) = \lambda_1 K(u_1) + \lambda_2 K(u_2)$ because $v \in E_{MC}$ is arbitrary. Therefore linearity is proven.

Now let us rewrite (1.20.5) in terms of this representation:

$$|(K(u), v)| \le m \|u\| \, \|v\|.$$

Take $v = K(u)$; then

$$\|K(u)\|^2 \le m \|u\| \, \|K(u)\|,$$

so

$$\|K(u)\| \le m \|u\|. \tag{1.20.7}$$

Hence K is a continuous operator in E_{MC}.

Equation (1.20.3) can now be written in the form

$$(u, v) = \lambda (K(u), v).$$

Since v is an arbitrary element of E_{MC}, this equation is equivalent to the operator equation

$$u = \lambda K(u)$$

with a continuous linear operator K.

By (1.20.7), we get

$$\|\lambda K(u) - \lambda K(v)\| = |\lambda| \|K(u - v)\| \le m |\lambda| \, \|u - v\|.$$

If $m|\lambda| < 1$, then λK is a contraction operator in E_{MC} and, by the contraction mapping principle, there is a unique fixed point of λK which clearly is $u = 0$. So the set $|\lambda| < 1/m$ does not contain real eigenvalues of the problem. Further, we shall see (and this is well known in mechanics) that eigenvalues in this problem must be real. The fact that the set $|\lambda| < 1/m$ does not contain real eigenvalues, and so any eigenvalues of the problem, has a clear mechanical sense: the lowest eigenfrequency of oscillation of a bounded clamped membrane is strictly positive. (From (1.20.3), when $v = u$ it follows that an eigenvalue must be positive.)

In a similar way, we can introduce eigenvalue problems for plates and elastic bodies. Here we can obtain corresponding equations of the form (1.20.4) with continuous linear operators and can also show that the corresponding lowest eigenvalues are strictly positive. All this we leave to the reader; later we shall consider eigenvalue problems in more detail.

In what follows, we shall see that, using the Riesz representation theorem, one can also introduce operators and operator equations for nonlinear problems of mechanics. One of them is presented in the next section.

1.21 The Problem of Elastico-Plasticity; Small Deformations

Following the lines of a paper by I.I. Vorovich and Yu.P. Krasovskij [25] that was published in a sketchy form, we consider a variant of the theory of elastico-plasticity (Il'yushin [13]), and justify the so-called method of elastic solutions for corresponding boundary value problems.

The system of partial differential equations describing the behavior of an elastic-plastic body occupying a bounded volume is

$$\left(\frac{\nu}{\nu-2}-\frac{\omega}{3}\right)\frac{\partial\theta}{\partial x_k}+(1-\omega)\Delta u_k- \\ -\frac{2}{3}e_I\frac{d\omega}{de_I}\sum_{s,t=1}^{3}\epsilon^*_{ks}\sum_{l=1}^{3}\epsilon^*_{lt}\frac{\partial^2 u_l}{\partial x_s\partial x_t}+\frac{F_k}{G}=0 \tag{1.21.1}$$

where ν is Poisson's ratio, G is the shear modulus, $\mathbf{F}=(F_1,F_2,F_3)$ are the volume forces, $\omega(e_i)$ is a function of the variable e_I, an intensity of the tensor of strains which defines plastic properties of the material with hardening; $\omega(e_I)$ must satisfy the following condition:

$$0\le\omega(e_I)\le\omega(e_I)+e_I\frac{d\omega(e_I)}{de_I}\le\lambda<1. \tag{1.21.2}$$

Other bits of notation are

$$\theta\equiv\theta(\mathbf{u})=\epsilon_{11}(\mathbf{u})+\epsilon_{22}(\mathbf{u})+\epsilon_{33}(\mathbf{u}),$$

$$\epsilon^*_{ks}=\begin{cases}\left(\frac{\partial u_k}{\partial x_s}-\frac{\theta}{3}\right)\frac{\sqrt{2}}{e_I}, & k=s,\\ \left(\frac{\partial u_k}{\partial x_s}+\frac{\partial u_s}{\partial x_k}\right)\frac{1}{\sqrt{2}e_I}, & k\ne s,\end{cases}$$

$$e_I=\frac{\sqrt{2}}{3}\left[(\epsilon_{11}-\epsilon_{22})^2+(\epsilon_{11}-\epsilon_{33})^2+(\epsilon_{22}-\epsilon_{33})^2+6(\epsilon_{12}^2+\epsilon_{13}^2+\epsilon_{23}^2)\right]^{1/2},$$

and

$$\epsilon_{ij}=\frac{1}{2}\left(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\right).$$

If $\omega(e_I) \equiv 0$ we get the equations of linear elasticity (for an isotropic homogeneous body). By analogy with elasticity problems, to pose a boundary value problem for (1.21.1) we must supplement it with boundary conditions. We consider a mixed boundary value problem: a part S_0 of the boundary $\partial\Omega$ of a body occupying the domain Ω is fixed,

$$\mathbf{u}\Big|_{S_0} = 0, \tag{1.21.3}$$

and the remainder $S = \partial\Omega \setminus S_0$ is subjected to surface forces $\mathbf{f}(\mathbf{x})$:

$$\begin{aligned}&\left[K\theta\cos(\mathbf{n},\mathbf{x}_k^0) + \sqrt{2}Ge_I\sum_{m=1}^{3}\epsilon_{km}^*\cos(\mathbf{n},\mathbf{x}_k^0)\right]\Bigg|_S\\ &= \sum_{k=1}^{3} f_k\cos(\mathbf{n},\mathbf{x}_k^0) + G\sqrt{2}\omega e_I\sum_{m=1}^{3}\epsilon_{km}^*\cos(\mathbf{n},\mathbf{x}_m^0)\end{aligned} \tag{1.21.4}$$

where K is the modulus of volume compressibility and $\cos(\mathbf{n},\mathbf{x}_k^0)$ is the cosine of the angle between the direction of the outward normal $\mathbf{n}$ to the boundary at a point and the direction $\mathbf{x}_k^0$ of the Cartesian axis OX_k.

When $\omega(e_I)$ is small (as it is if e_I is small) we have a nonlinear boundary value problem which is, in a certain way, a perturbation of a corresponding boundary value problem of linear elasticity. It leads to the idea of using an iterative procedure, the so-called method of elastic solutions, to solve the former. This procedure looks like that of the contraction mapping principle if we can make the problem take the corresponding operator form. Then it remains to show that the operator of the problem is a contraction. Now we begin to carry out the program.

Let us introduce the notation

$$\begin{aligned}\langle\mathbf{u},\mathbf{v}\rangle = \frac{2}{9}\{&[\epsilon_{11}(\mathbf{u}) - \epsilon_{22}(\mathbf{u})][\epsilon_{11}(\mathbf{v}) - \epsilon_{22}(\mathbf{v})] +\\ &+ [\epsilon_{11}(\mathbf{u}) - \epsilon_{33}(\mathbf{u})][\epsilon_{11}(\mathbf{v}) - \epsilon_{33}(\mathbf{v})] +\\ &+ [\epsilon_{22}(\mathbf{u}) - \epsilon_{33}(\mathbf{u})][\epsilon_{22}(\mathbf{v}) - \epsilon_{33}(\mathbf{v})] +\\ &+ 6[\epsilon_{12}(\mathbf{u})\epsilon_{12}(\mathbf{v}) + \epsilon_{13}(\mathbf{u})\epsilon_{13}(\mathbf{v}) + \epsilon_{23}(\mathbf{u})\epsilon_{23}(\mathbf{v})]\}\end{aligned} \tag{1.21.5}$$

If we consider the terms on the right-hand side of (1.21.5) as coordinates of vectors $\mathbf{a} = (a_1,\dots,a_6)$, $\mathbf{b} = (b_1,\dots,b_6)$,

$$a_i = c_i(\mathbf{u}), \qquad b_i = c_i(\mathbf{v}), \qquad i = 1,\dots,6,$$

$$\begin{aligned}c_1(\mathbf{w}) &= \frac{\sqrt{2}}{3}[\epsilon_{11}(\mathbf{w}) - \epsilon_{22}(\mathbf{w})], & c_2(\mathbf{w}) &= \frac{\sqrt{2}}{3}[\epsilon_{11}(\mathbf{w}) - \epsilon_{33}(\mathbf{w})],\\ c_3(\mathbf{w}) &= \frac{\sqrt{2}}{3}[\epsilon_{22}(\mathbf{w}) - \epsilon_{33}(\mathbf{w})], & c_4(\mathbf{w}) &= \frac{2}{\sqrt{3}}\epsilon_{12}(\mathbf{w}),\\ c_5(\mathbf{w}) &= \frac{2}{\sqrt{3}}\epsilon_{13}(\mathbf{w}), & c_6(\mathbf{w}) &= \frac{2}{\sqrt{3}}\epsilon_{23}(\mathbf{w}),\end{aligned}$$

then the form $\langle \mathbf{u}, \mathbf{v} \rangle$ is a scalar product of $\mathbf{a}$ by $\mathbf{b}$ in $\mathbb{R}^6$:

$$\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{i=1}^{6} a_i b_i.$$

Besides,

$$\langle \mathbf{u}, \mathbf{u} \rangle = \sum_{i=1}^{6} c_i^2(\mathbf{u}) = e_I^2(\mathbf{u}) \tag{1.21.6}$$

and by the Schwarz inequality we get

$$|\langle \mathbf{u}, \mathbf{v} \rangle| = \left| \sum_{i=1}^{6} c_i(\mathbf{u}) c_i(\mathbf{v}) \right| \leq e_I(\mathbf{u}) e_I(\mathbf{v}). \tag{1.21.7}$$

On the set C_2 of vector functions satisfying the boundary condition (1.21.3) and such that each of their components is of class $C^{(2)}(\Omega)$, let us now introduce an inner product

$$(\mathbf{u}, \mathbf{v}) = \int_\Omega \left(\frac{3}{2} G \langle \mathbf{u}, \mathbf{v} \rangle + \frac{1}{2} K \theta(\mathbf{u}) \theta(\mathbf{v}) \right) d\Omega. \tag{1.21.8}$$

This coincides with a special case of the inner product (1.10.26) of the linear theory of elasticity. So the completion of C_2 in the metric corresponding to (1.21.8) is the energy space of linear elasticity E_{EM} (M for "mixed") if we suppose that the condition (1.21.3) provides $\mathbf{u} = 0$ if

$$\|\mathbf{u}\|^2 = \int_\Omega \left(\frac{3}{2} G e_I^2(\mathbf{u}) + \frac{1}{2} K \theta^2(\mathbf{u}) \right) d\Omega = 0.$$

The norm of E_{EM} is equivalent to one of $W^{1,2}(\Omega) \times W^{1,2}(\Omega) \times W^{1,2}(\Omega)$ (see Section 1.10 and Fichera [9]). (By $H_1 \times H_2$ we denote the so-called Cartesian product of Hilbert spaces H_1 and H_2, the elements of which are pairs (x, y), $x \in H_1$, $y \in H_2$. The scalar product in $H_1 \times H_2$ is defined by the relation

$$(x_1, x_2)_1 + (y_1, y_2)_2$$

where $x_1, x_2 \in H_1$ and $y_1, y_2 \in H_2$.)

By the principle of virtual displacements, the integro-differential equation of equilibrium of an elasto-plastic body is

$$(\mathbf{u}, \mathbf{v}) - \frac{3}{2} G \int_\Omega \omega(e_I(\mathbf{u})) \langle \mathbf{u}, \mathbf{v} \rangle \, d\Omega - \\ - \sum_{i=1}^{3} \int_\Omega F_i v_i \, d\Omega - \sum_{i=1}^{3} \int_S f_i v_i \, dS = 0. \tag{1.21.9}$$

This equation can be obtained using the equations (1.21.1) and the boundary conditions (1.21.3) and (1.21.4). Conversely, using the technique of the

classical calculus of variations we can get (1.21.1) and the natural boundary conditions (1.21.4). Thus, in a certain way, (1.21.9) is equivalent to the above statement of the problem. So we can introduce

Definition 1.21.1. A vector function $\mathbf{u} \in E_{EM}$ is called the generalized solution of a problem of elastico-plasticity if it satisfies (1.21.9) for every $\mathbf{v} \in E_{EM}$.

For correctness of this definition we must impose some restrictions on external forces. It is evident that they coincide with those for linear elasticity. So we assume that

$$F_i(x_1, x_2, x_3) \in L^{6/5}(\Omega), \qquad f_i(x_1, x_2, x_3) \in L^{4/3}(S). \tag{1.21.10}$$

Consider the form

$$A[\mathbf{u}, \mathbf{v}] = \frac{3}{2} G \int_\Omega \omega(e_I(\mathbf{u}))\langle \mathbf{u}, \mathbf{v}\rangle \, d\Omega + \sum_{i=1}^{3} \int_\Omega F_i v_i \, d\Omega + \sum_{i=1}^{3} \int_S f_i v_i \, dS$$

as a functional in E_{EM} with respect to $\mathbf{v}(x_1, x_2, x_3)$ when $\mathbf{u}(x_1, x_2, x_3) \in E_{EM}$ is fixed. As in linear elasticity, the load terms, thanks to (1.21.10), are continuous linear functionals with respect to $\mathbf{v} \in E_{EM}$. Finally, in accordance with (1.21.5) and (1.21.2), we get

$$\left| \frac{3}{2} G \int_\Omega \omega(e_I(\mathbf{u}))\langle \mathbf{u}, \mathbf{v}\rangle \, d\Omega \right| \le \lambda \frac{3}{2} G \int_\Omega |\langle \mathbf{u}, \mathbf{v}\rangle| \, d\Omega \le \lambda \|\mathbf{u}\| \, \|\mathbf{v}\|,$$

so this part of the functional is also continuous.

Therefore we can apply the Riesz representation theorem to $A[\mathbf{u}, \mathbf{v}]$, which gives

$$A[\mathbf{u}, \mathbf{v}] = (\mathbf{v}, \mathbf{f}) \equiv (\mathbf{f}, \mathbf{v}).$$

This representation uniquely defines a correspondence $u \mapsto f$, where $u, f \in E_{EM}$, we obtain an operator $\mathbf{f} = A(\mathbf{u})$ acting in E_{EM}. Equation (1.21.9) is now equivalent to

$$(\mathbf{u}, \mathbf{v}) - (A(\mathbf{u}), \mathbf{v}) = 0 \tag{1.21.11}$$

or, since $\mathbf{v} \in E_{EM}$ is arbitrary,

$$\mathbf{u} = A(\mathbf{u}). \tag{1.21.12}$$

The operator A is nonlinear. We shall show that it is a contraction operator. For this, take arbitrary elements $\mathbf{u}, \mathbf{v}, \mathbf{w} \in E_{EM}$ and consider

$$(A(\mathbf{u}) - A(\mathbf{v}), \mathbf{w}) = \frac{3}{2} G \int_\Omega [\omega(e_I(\mathbf{u}))\langle \mathbf{u}, \mathbf{w}\rangle - \omega(e_I(\mathbf{v}))\langle \mathbf{v}, \mathbf{w}\rangle] \, d\Omega. \tag{1.21.13}$$

First, let $\mathbf{u}, \mathbf{v}, \mathbf{w}$ be in C_2. At every point of Ω, by (1.21.7), we can estimate the integrand from (1.21.13) as follows. We have

$$\begin{aligned}\text{Int} &= |\omega(e_I(\mathbf{u}))\,\langle \mathbf{u}, \mathbf{w}\rangle - \omega(e_I(\mathbf{v}))\,\langle \mathbf{v}, \mathbf{w}\rangle| \\ &= \left|\omega(e_I(\mathbf{u}))\sum_{i=1}^{6} c_i(\mathbf{u})\,c_i(\mathbf{w}) - \omega(e_I(\mathbf{v}))\sum_{i=1}^{6} c_i(\mathbf{v})\,c_i(\mathbf{w})\right|.\end{aligned}$$

Let us introduce a real-valued function $f(t)$ of a real variable t by the relation

$$f(t) = \sum_{i=1}^{6} \omega(e_I(t\mathbf{u} + (1-t)\mathbf{v}))c_i(t\mathbf{u} + (1-t)\mathbf{v})c_i(\mathbf{w}).$$

It is seen that

$$\text{Int} = |f(1) - f(0)|.$$

As $f(t)$ is continuously differentiable, the classical mean value theorem gives

$$f(1) - f(0) = f'(z)(1-0) = f'(z) \quad \text{for some } z \in [0,1],$$

or, in the above terms, we get

$$\begin{aligned}\text{Int} &= \left|\frac{d}{dt}\left\{\sum_{i=1}^{6} \omega(e_I(t\mathbf{u} + (1-t)\mathbf{v}))\,c_i(t\mathbf{u} + (1-t)\mathbf{v})\,c_i(\mathbf{w})\right\}_{t=z}\right| \\ &= \left|\left\{\frac{d\omega(e_I(t\mathbf{u} + (1-t)\mathbf{v}))}{de_I}\,\frac{de_I(t\mathbf{u} + (1-t)\mathbf{v})}{dt}\cdot\right.\right. \\ &\quad \left.\left.\cdot \sum_{i=1}^{6} c_i(t\mathbf{u} + (1-t)\mathbf{v})\,c_i(\mathbf{w}) + \omega\sum_{i=1}^{6} c_i(\mathbf{u} - \mathbf{v})\,c_i(\mathbf{w})\right\}_{t=z}\right|.\end{aligned}$$

(Here we have used the linearity of $c_i(\mathbf{u})$ in $\mathbf{u}$ and, thus, in t.) Let us consider the term

$$\begin{aligned}T &= \sum_{i=1}^{6} \frac{de_I(t\mathbf{u} + (1-t)\mathbf{v})}{dt}\,c_i(t\mathbf{u} + (1-t)\mathbf{v})\,c_i(\mathbf{w}) \\ &= \sum_{i=1}^{6} \frac{d}{dt}\left(\sum_{j=1}^{6} c_j^2(t\mathbf{u} + (1-t)\mathbf{v})\right)^{1/2} c_i(t\mathbf{u} + (1-t)\mathbf{v})\,c_i(\mathbf{w}) \\ &= \sum_{i=1}^{6} \frac{2\sum_{j=1}^{6} c_j(t\mathbf{u} + (1-t)\mathbf{v})\,c_j(\mathbf{u} - \mathbf{v})}{2\left(\sum_{j=1}^{6} c_j^2(t\mathbf{u} + (1-t)\mathbf{v})\right)^{1/2}}\,c_i(t\mathbf{u} + (1-t)\mathbf{v})\,c_i(\mathbf{w}).\end{aligned}$$

Applying the Schwarz inequality, we obtain

$$
\begin{aligned}
|T| &\le \sum_{i=1}^{6} \frac{\left(\sum_{j=1}^{6} c_j^2(t\mathbf{u}+(1-t)\mathbf{v})\right)^{1/2}\left(\sum_{j=1}^{6} c_j^2(\mathbf{u}-\mathbf{v})\right)^{1/2}}{\left(\sum_{j=1}^{6} c_j^2(t\mathbf{u}+(1-t)\mathbf{v})\right)^{1/2}}\cdot \\
&\qquad \cdot |c_i(t\mathbf{u}+(1-t)\mathbf{v})|\,|c_i(\mathbf{w})| \\
&= \left(\sum_{j=1}^{6} c_j^2(\mathbf{u}-\mathbf{v})\right)^{1/2} \sum_{i=1}^{6} |c_i(t\mathbf{u}+(1-t)\mathbf{v})|\,|c_i(\mathbf{w})| \\
&\le e_I(\mathbf{u}-\mathbf{v}) \left(\sum_{i=1}^{6} c_i^2(t\mathbf{u}+(1-t)\mathbf{v})\right)^{1/2} \left(\sum_{i=1}^{6} c_i^2(\mathbf{w})\right)^{1/2} \\
&= e_I(\mathbf{u}-\mathbf{v})\, e_I(t\mathbf{u}+(1-t)\mathbf{v})\, e_I(\mathbf{w})
\end{aligned}
$$

(here we also used (1.21.6)). Similarly,

$$
\begin{aligned}
\left|\sum_{i=1}^{6} c_i(\mathbf{u}-\mathbf{v})\, c_i(\mathbf{w})\right| &\le \left(\sum_{i=1}^{6} c_i^2(\mathbf{u}-\mathbf{v})\right)^{1/2} \left(\sum_{i=1}^{6} c_i^2(\mathbf{w})\right)^{1/2} \\
&= e_I(\mathbf{u}-\mathbf{v})\, e_I(\mathbf{w}).
\end{aligned}
$$

Combining all these, we get

$$
\begin{aligned}
\mathrm{Int} &\le \left\{ \frac{d\omega(e_I(t\mathbf{u}+(1-t)\mathbf{v}))}{de_I}\, e_I(t\mathbf{u}+(1-t)\mathbf{v}))\, e_I(\mathbf{u}-\mathbf{v})\, e_I(\mathbf{w}) + \right. \\
&\qquad \left. + \omega(e_I(t\mathbf{u}+(1-t)\mathbf{v}))\, e_I(\mathbf{u}-\mathbf{v})\, e_I(\mathbf{w}) \right\}_{t=z} \\
&= \left\{ \omega(e_I(t\mathbf{u}+(1-t)\mathbf{v})) + \frac{d\omega(e_I(t\mathbf{u}+(1-t)\mathbf{v}))}{de_I}\cdot \right. \\
&\qquad \left. \cdot\, e_I(t\mathbf{u}+(1-t)\mathbf{v})) \right\}_{t=z} e_I(\mathbf{u}-\mathbf{v})\, e_I(\mathbf{w}).
\end{aligned}
$$

By the condition (1.21.2), we have

$$
\mathrm{Int} \le \lambda e_I(\mathbf{u}-\mathbf{v})\, e_I(\mathbf{w}) \tag{1.21.14}
$$

at every point of Ω.

Returning to (1.21.13) we have, using (1.21.14),

$$
|(A(\mathbf{u})-A(\mathbf{v}),\mathbf{w})| \le \lambda \int_{\Omega} \frac{3}{2} G\, e_I(\mathbf{u}-\mathbf{v})\, e_I(\mathbf{w})\, d\Omega.
$$

In accordance with the norm of E_{EM} it follows that

$$
|(A(\mathbf{u})-A(\mathbf{v}),\mathbf{w})| \le \lambda \|\mathbf{u}-\mathbf{v}\|\, \|\mathbf{w}\|
$$

or, putting $\mathbf{w} = A(\mathbf{u}) - A(\mathbf{v})$, we get

$$\|A(\mathbf{u}) - A(\mathbf{v})\| \le \lambda \|\mathbf{u} - \mathbf{v}\|, \qquad \lambda = \text{const} < 1. \tag{1.21.15}$$

Being obtained for $\mathbf{u}, \mathbf{v}, \mathbf{w} \in C_2$, this inequality holds for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in E_{EM}$ since in this inequality we can pass to the limit for corresponding Cauchy sequences in E_{EM}.

Inequality (1.21.15) states that A is a contraction operator in E_{EM}; hence we can apply the contraction mapping principle, which states that (1.21.12) has a unique solution that can be found using the iterative procedure

$$\mathbf{u}_{k+1} = A(\mathbf{u}_k), \qquad k = 0, 1, 2, \ldots.$$

This procedure begins with an arbitrary element $\mathbf{u}_0 \in E_{EM}$; when $\mathbf{u}_0 = 0$, the procedure is called the method of elastic solutions since at each step we must solve a problem of linear elasticity with given load terms. In practical terms, the method works best when the constant λ is small.

So we can formulate

Theorem 1.21.1. Assume S_0 is a piecewise smooth surface of nonzero area and that the conditions (1.21.2) and (1.21.10) are fulfilled. Then a mixed boundary value problem of elastico-plasticity has a unique generalized solution in the sense of Definition 1.21.1; the iterative procedure (1.21.15) defines a sequence of successive approximations $\mathbf{u}_k \in E_{EM}$ which converges to the solution $\mathbf{u} \in E_{EM}$ and

$$\|\mathbf{u}_k - \mathbf{u}\| \le \frac{\lambda^k}{1 - \lambda} \|\mathbf{u}_0 - \mathbf{u}_1\|. \tag{1.21.16}$$

It is clear that we cannot apply this theorem when, say, $S = \partial\Omega$. In such a case, we must add the self-balance conditions (1.14.16). These guarantee that we can repeat the above method for a free elastic-plastic body, and so can formulate

Theorem 1.21.2. Assume that all the requirements of Theorem 1.21.1 and the self-balance conditions (1.14.16) are met. Then there is a unique generalized solution of the boundary value problem for a bounded elastic-plastic body, and it can be found by an iterative procedure of the form (1.21.15).

Problem 1.21.1. Is an estimate of the type (1.21.16) valid in Theorem 1.21.2?

We recommend that the reader prove Theorem 1.21.2 in detail, in order to gain experience with the technique.

Remark 1.21.1. We would like to call attention to the way in which we obtained the main inequality of this section. First it was proved for smooth functions and then was extended to the general case. This is a standard technique in the treatment of nonlinear problems of mechanics.

1.22 Bases and Complete Systems

If a linear space Y has finite dimension n, then there are n linearly independent elements $g_1, g_2, g_3, \ldots, g_n$ called a *basis* of Y such that every $y \in Y$ has a unique representation

$$y = \sum_{k=1}^{n} \alpha_k g_k,$$

the α_k being scalars. We now consider an infinite dimensional normed space X.

Definition 1.22.1. A system of elements $e_1, e_2, \ldots,$ is called a *basis* of X if any element $x \in X$ has an unique representation

$$x = \sum_{k=1}^{\infty} \alpha_k e_k$$

where the α_k are scalars.

It is clear that a basis $e_1, e_2, \ldots$ is a linearly independent system since the equation

$$0 = \sum_{k=1}^{\infty} \alpha_k e_k$$

has the unique solution $0 = \alpha_1 = \alpha_2 = \alpha_3 = \cdots$.

If a normed space has a basis, then the space is separable: a countable set of all linear combinations $\sum_{k=1}^{n} c_k e_k$ (n arbitrary) with rational coefficients c_k is dense in the space.

We are familiar with some systems of functions which could be bases in certain spaces: for example, $\left\{ \frac{1}{\sqrt{2\pi}} e^{ikx} \right\}$ in $L^2(0, 2\pi)$. Later, we confirm this example.

Now we consider the system $\{x^k\}$, $k = 1, 2, \ldots$, in $C(0, 1)$. If it is a basis, then any function $f(x) \in C(0, 1)$ could be represented in the form

$$f(x) = \sum_{k=0}^{\infty} \alpha_k x^k,$$

where the series converges uniformly on $[0, 1]$. This representation means that the function is analytic, but we know that there are continuous functions on $[0, 1]$ which are not analytic. Hence the system $\{x^k\}$ is not a basis. On the other hand, the Weierstrass theorem states that this system possesses properties similar to those of a basis. To generalize this similarity, we introduce

Definition 1.22.2. A countable system $g_1, g_2, g_3, \ldots$ of elements from a normed space X is said to be *complete in* X if for any $x \in X$ and any positive number ε there is a finite linear combination $\sum_{i=1}^{n(\varepsilon)} \alpha_i g_i$ such that

$$\left\| x - \sum_{i=1}^{n(\varepsilon)} \alpha_i g_i \right\| < \varepsilon.$$

By Definition 1.22.2, the system $\{x^k\}$ is complete in $C(0,1)$. Because $C(0,1)$ is dense in $L^p(0,1)$, $p \geq 1$, this system is also complete in $L^p(0,1)$.

Problem 1.22.1. Which systems are complete in $L^p(\Omega)$, $W^{k,p}(\Omega)$?

If a normed space has a countable complete system, then the space is separable. The reader should be able to name a countable dense set to verify this.

The problem of existence of a basis in a certain normed space is difficult, but there is a special case where it is fully solved: a separable Hilbert space. The reader will find here the theory of Fourier series largely repeated in abstract terms. We begin with

Definition 1.22.3. A system $\{x_k\}$ of elements of a Hilbert space H is said to be *orthonormal* if for all integers m, n,

$$(x_m, x_n) = \delta_{mn} = \begin{cases} 1, & m = n, \\ 0, & m \neq n. \end{cases}$$

We know, at least for $\mathbb{R}^n$, that there are some advantages in using an orthonormal system of vectors as a basis.

Suppose we have an arbitrary linearly independent system of elements of a Hilbert space H, say $f_1, f_2, f_3, \ldots, f_n$, and let H_n be the subspace of H spanned by this system. We would like to use this system to construct an orthonormal system $g_1, g_2, g_3, \ldots, g_n$ that is also a basis of H_n. This can be accomplished by the *Gram–Schmidt procedure*:

(1) The first element of the new system is $g_1 = f_1/\|f_1\|$, $\|g_1\| = 1$.

(2) Take $e_2 = f_2 - (f_2, g_1)g_1$; then $(e_2, g_1) = (f_2, g_1) - (f_2, g_1)\|g_1\|^2 = 0$, so the second element is $g_2 = e_2/\|e_2\|$.

(3) Take $e_3 = f_3 - (f_3, g_1)g_1 - (f_3, g_2)g_2$; then $(e_3, g_1) = 0$ and $(e_3, g_2) = 0$. Since $e_3 \neq 0$, we get the third element as $g_3 = e_3/\|e_3\|$.

$\vdots$

(i) Let $e_i = f_i - (f_i, g_1)g_1 - \cdots - (f_i, g_{i-1})g_{i-1}$. It is seen that $(e_i, g_k) = 0$ for $k = 1, \ldots, i-1$, hence we set $g_i = e_i/\|e_i\|$.

This process can be continued ad infinitum since all $e_k \neq 0$ (why?). So we obtain an orthonormalized system $g_1, g_2, g_3, \ldots$. The process is, however, found to be unstable for numerical computation.

As is known from linear algebra, a system $f_1, f_2, \ldots, f_n$ is linearly independent in an inner product space if and only if the Gram determinant

$$\begin{vmatrix} (f_1, f_1) & (f_1, f_2) & \cdots & (f_1, f_n) \\ (f_2, f_1) & (f_2, f_2) & \cdots & (f_2, f_n) \\ \vdots & \vdots & \ddots & \vdots \\ (f_n, f_1) & (f_n, f_2) & \cdots & (f_n, f_n) \end{vmatrix}$$

is not equal to zero. For an orthonormal system of elements the Gram determinant, being the determinant of the identity matrix, equals $+1$; hence an orthonormal system is linearly independent.

Problem 1.22.2. Show directly the linear independence of an orthonormal system.

Let $g_1, g_2, g_3, \ldots$ be an orthonormal system in a complex Hilbert space H. For an element $f \in H$, the numbers α_k, $k = 1, 2, 3, \ldots$ defined by the equality $\alpha_k = (f, g_k)$ are called the *Fourier coefficients* of f. Now we prove

Theorem 1.22.1. A complete orthonormal system $g_1, g_2, g_3, \ldots$ of elements of a Hilbert space H is a basis of H; any element $f \in H$ has unique representation

$$f = \sum_{k=1}^{\infty} \alpha_k g_k, \tag{1.22.1}$$

called the *Fourier series* of f, where $\alpha_k = (f, g_k)$ are the Fourier coefficients of f.

Proof. First we consider the problem of the best approximation of an element $f \in H$ by elements of a subspace H_n spanned by $g_1, g_2, g_3, \ldots, g_n$. In Section 1.18 we showed that this problem has a unique solution. Now we show that it is $\sum_{k=1}^{n} \alpha_k g_k$. Indeed, consider an arbitrary linear combination $\sum_{k=1}^{n} c_k g_k$. Then

$$\begin{aligned} \left\| f - \sum_{k=1}^{n} c_k g_k \right\|^2 &= \left(f - \sum_{k=1}^{n} c_k g_k, f - \sum_{k=1}^{n} c_k g_k \right) \\ &= \|f\|^2 - \left(f, \sum_{k=1}^{n} c_k g_k \right) - \left(\sum_{k=1}^{n} c_k g_k, f \right) + \left\| \sum_{k=1}^{n} c_k g_k \right\|^2 \\ &= \|f\|^2 - \sum_{k=1}^{n} \overline{c_k} \alpha_k - \sum_{k=1}^{n} c_k \overline{\alpha_k} + \sum_{k=1}^{n} c_k \overline{c_k} \\ &= \|f\|^2 - \sum_{k=1}^{n} |\alpha_k|^2 + \sum_{k=1}^{n} |c_k - \alpha_k|^2. \end{aligned}$$

It is seen that $\|f - \sum_{k=1}^{n} c_k g_k\|^2$ takes its minimum value when $c_k = \alpha_k$:

$$\left\| f - \sum_{k=1}^{n} \alpha_k g_k \right\|^2 = \min_{c_1,\ldots,c_n} \left\| f - \sum_{k=1}^{n} c_k g_k \right\|^2 = \|f\|^2 - \sum_{k=1}^{n} |\alpha_k|^2 \geq 0; \tag{1.22.2}$$

moreover, we obtain *Bessel's inequality*

$$\sum_{k=1}^{n} |(f, g_k)|^2 \leq \|f\|^2. \tag{1.22.3}$$

Denote

$$f_n = \sum_{k=1}^{n} \alpha_k g_k,$$

f_n being the nth partial sum of the Fourier series for f. Let us show that $\{f_n\}$ is a Cauchy sequence in H. By Bessel's inequality,

$$\sum_{k=1}^{n} |\alpha_k|^2 \leq \|f\|^2; \tag{1.22.4}$$

hence

$$\|f_n - f_{n+m}\|^2 = \left\| \sum_{k=n+1}^{n+m} \alpha_k g_k \right\|^2 = \sum_{k=n+1}^{n+m} |\alpha_k|^2 \to 0 \text{ as } n \to \infty.$$

By completeness of the system $g_1, g_2, g_3, \ldots$ in H, for any $\varepsilon > 0$ we can find a number N and coefficients $c_k(\varepsilon)$ such that

$$\left\| f - \sum_{k=1}^{N} c_k(\varepsilon) g_k \right\|^2 < \varepsilon.$$

By (1.22.2)

$$\|f - f_N\|^2 = \left\| f - \sum_{k=1}^{N} \alpha_k g_k \right\|^2 \leq \left\| f - \sum_{k=1}^{N} c_k(\varepsilon) g_k \right\|^2 < \varepsilon,$$

so the sequence $\{f_N\}$ converges to f:

$$f = \lim_{n\to\infty} f_n. \tag{1.22.5}$$

This completes the proof. □

From (1.22.5) we can obtain *Parseval's equality*

$$\sum_{k=1}^{\infty} |(f, g_k)|^2 = \|f\|^2, \tag{1.22.6}$$

which holds whenever $g_1, g_2, g_3, \ldots$ is a complete orthonormal system in H. Indeed, by (1.22.2),

$$0 = \lim_{n\to\infty} \left\| f - \sum_{k=1}^{n} \alpha_k g_k \right\|^2 = \lim_{n\to\infty} \left(\|f\|^2 - \sum_{k=1}^{n} |\alpha_k|^2 \right).$$

Now we introduce

Definition 1.22.4. A system $e_1, e_2, e_3, \ldots$ of elements of a Hilbert space H is said to be *closed in* H if from the system of equations

$$(f, e_k) = 0 \text{ for all } k = 1, 2, 3, \ldots$$

it follows that $f = 0$.

It is clear that a complete orthonormal system of elements is closed in H. The converse statement is correct as well. We formulate both statements as

Theorem 1.22.2. Let $g_1, g_2, g_3, \ldots$ be an orthonormal system of elements in a Hilbert space H. This system is complete in H if and only if it is closed in H.

Proof. We need to demonstrate only that a closed orthonormal system in H is complete. Proving the previous theorem, we established that for any element $f \in H$ the sequence of partial Fourier sums $f_n = \sum_{k=1}^{n} \alpha_k g_k$ is a Cauchy sequence. Since H is a complete space, there exists $f^* = \lim_{n\to\infty} f_n$ that belongs to H. To complete the proof we need to show that $f = f^*$. We have

$$(f - f^*, g_m) = \lim_{n\to\infty} \left(f - \sum_{k=1}^{n} \alpha_k g_k, g_m \right) = \alpha_m - \alpha_m = 0.$$

By Definition 1.22.4 it follows that $f = f^*$, hence the system $g_1, g_2, g_3, \ldots$ is complete. □

It is normally simpler to check whether a system is closed than to check whether it is complete. At the beginning of this section we established that any system of linearly independent elements in H can be transformed into an orthonormal system equivalent to the original system in a certain way. So we conclude

Theorem 1.22.3. A complete system $g_1, g_2, g_3, \ldots$ in H is closed in H; conversely, a system closed in H is complete in H.

As we said above, the existence of a countable basis in a Hilbert space provides its separability. The converse statement is also valid. We formulate that as

Theorem 1.22.4. A Hilbert space H has a countable orthonormal basis if and only if H is separable.

The proof follows immediately from the previous theorem. Indeed, in H select a countable set of elements that is dense everywhere in H. Using the Gram–Schmidt procedure, produce an orthonormal system of elements from this set (removing any linearly dependent elements). Since the initial system is dense it is complete and thus, as a result of the Gram–Schmidt procedure, we get an orthonormal basis of the space.

Remember that all of the energy spaces we introduced above are separable. Hence each of them has a countable orthonormal basis (non-unique, of course).

In conclusion, we consider whether the system $\{e^{ikx}/\sqrt{2\pi}\}$ is a basis of the complex space $L^2(0, 2\pi)$. From standard calculus it is known that the system is orthonormal in $L^2(0, 2\pi)$ (the reader, however, can check this). Weierstrass's theorem on the approximation of a function continuous on $[0, 2\pi]$ can be formulated as the statement that the set of trigonometric polynomials, that is, finite sums of the form $\sum_k \alpha_k e^{ikx}$, is dense in the complex space $C(0, 2\pi)$. But the set of functions $C(0, 2\pi)$ is the base for construction of $L^2(0, 2\pi)$, hence the same set of finite sums $\sum_k \alpha_k e^{ikx}$ is dense in $L^2(0, 2\pi)$. This shows that $\{e^{ikx}/\sqrt{2\pi}\}$ is an orthonormal basis of $L^2(0, 2\pi)$.

1.23 Weak Convergence in a Hilbert Space

We know that in $\mathbb{R}^n$, the convergence of a sequence of vectors is equivalent to coordinate-wise convergence.

In a Hilbert space H, the Fourier coefficients (f, g_k) of an element $f \in H$ play the role of the coordinates of f. Suppose $g_1, g_2, g_3, \ldots$ is an orthonormal basis of H. What can we say about convergence of a sequence $\{f_n\}$ if, for every fixed k, the numerical sequence (f_n, g_k) is convergent?

Let us consider $\{g_n\}$, the sequence of elements of the orthonormal basis. It is seen that for every k,

$$\lim_{n\to\infty} (g_n, g_k) = 0,$$

hence we have coordinate-wise convergence. But the sequence $\{g_n\}$ is not convergent, since $\|g_n - g_m\| = \sqrt{2}$ if $n \neq m$.

Therefore, coordinate-wise convergence in a Hilbert space is not equivalent to the usual form of convergence in the space. We define a new type of convergence in a Hilbert space.

Definition 1.23.1. Let H be a Hilbert space. A sequence $\{x_k\}$, $x_k \in H$, is said to be *weakly convergent* to $x_0 \in H$ if for every continuous linear

functional $F(x)$ in H,

$$\lim_{k\to\infty} F(x_k) = F(x_0).$$

If every numerical sequence $\{F(x_k)\}$ is a Cauchy sequence, then $\{x_k\}$ is called a *weak Cauchy sequence*.

To distinguish between weak convergence and convergence as defined on page 19, we shall refer to the latter as *strong convergence*. We shall retain the notation $x_k \to x$ for strong convergence, and adopt $x_k \rightharpoonup x$ for weak convergence.

Definition 1.23.1 is given in a form which (with suitable modifications) is valid in a metric space. But in a Hilbert space any continuous linear functional, by the Riesz representation theorem, takes the form $F(x) = (x, f)$ where f is an element of H. So Definition 1.23.1 may be rewritten as follows:

Definition 1.23.2. Let H be a Hilbert space. A sequence $\{x_n\}$, $x_n \in H$, is weakly convergent to $x_0 \in H$ if for every element $f \in H$ we have

$$\lim_{n\to\infty} (x_n, f) = (x_0, f).$$

If every numerical sequence $\{(x_n, f)\}$ is a Cauchy sequence, then $\{x_k\}$ is a weak Cauchy sequence.

We have seen that some weak Cauchy sequences in H are not strong Cauchy sequences. But a strong Cauchy sequence is always a weak Cauchy sequence, by virtue of the continuity of the linear functionals in the definition.

We formulate a simple sufficient condition for strong convergence of a weakly convergent sequence:

Theorem 1.23.1. Suppose that $x_k \rightharpoonup x_0$, where x_k, x_0 belong to a Hilbert space H. Then $x_k \to x_0$ if $\lim_{k\to\infty} \|x_k\| = \|x_0\|$.

Proof. Consider $\|x_k - x_0\|^2$. We get

$$\|x_k - x_0\|^2 = (x_k - x_0, x_k - x_0) = \|x_k\|^2 - (x_0, x_k) - (x_k, x_0) + \|x_0\|^2.$$

By Definition 1.23.2 we have

$$\lim_{k\to\infty} [(x_0, x_k) + (x_k, x_0)] = 2\|x_0\|^2,$$

hence $\lim_{k\to\infty} \|x_k - x_0\|^2 = 0$. □

We shall see later that for many numerical methods it is easier to first establish weak convergence of approximate solutions and then strong convergence, than to establish strong convergence directly; the last theorem allows us to justify a method successively, beginning with a simple approximate result and then passing to the needed one. That is why weak convergence is a major preoccupation in this presentation.

Theorem 1.23.2. In a Hilbert space, every weak Cauchy sequence $\{x_n\}$ is bounded.

Proof. Suppose to the contrary that there is a weak Cauchy sequence $\{x_n\}$ which is not bounded in H. So let $\|x_n\| \to \infty$ as $n \to \infty$. We get a contradiction. Consider the set of all numbers U that consists of the numbers of the form (x_n, y), where y belongs to a closed ball $B(y_0, \varepsilon)$ with arbitrary (but momentarily fixed) $\varepsilon > 0$ and arbitrary center $y_0 \in H$. We first claim that any U is unbounded from above. Indeed, elements of the form $y_n = y_0 + \varepsilon x_n/(2\|x_n\|)$ belong to $B(y_0, \varepsilon)$ since

$$\|y_n - y_0\| = \left\| \frac{\varepsilon x_n}{2\|x_n\|} \right\| = \frac{\varepsilon}{2}$$

and

$$(x_n, y_n) = (x_n, y_0) + \frac{\varepsilon}{2\|x_n\|}(x_n, x_n) = (x_n, y_0) + \frac{\varepsilon}{2}\|x_n\| \to \infty \text{ as } n \to \infty$$

since, by definition of weak convergence, the numerical sequence $\{(x_n, y_0)\}$ is bounded.

We now want to obtain a contradiction. Take the ball $B(y_0, \varepsilon_1)$, $\varepsilon_1 = 1$, $y_0 = 0$. By the above note, we can find x_{n_1} and then $y_1 \in B(y_0, \varepsilon_1)$ such that

$$(x_{n_1}, y_1) > 1. \tag{1.23.1}$$

By continuity of the inner product in both its variables, we can find a ball $B(y_1, \varepsilon_2)$ such that $B(y_1, \varepsilon_2) \subset B(y_0, \varepsilon_1)$ and such that (1.23.1) holds not only for y_1 but for all $y \in B(y_1, \varepsilon_2)$:

$$(x_{n_1}, y) > 1 \text{ for all } y \in B(y_1, \varepsilon_2).$$

Then in the ball $B(y_1, \varepsilon_2)$ we similarly find x_{n_2}, $n_2 > n_1$, and a corresponding element y_2 such that

$$(x_{n_2}, y_2) > 2,$$

and after this a ball $B(y_2, \varepsilon_3)$ such that $B(y_2, \varepsilon_3) \subset B(y_1, \varepsilon_2)$ and

$$(x_{n_2}, y) > 2 \text{ for all } y \in B(y_2, \varepsilon_3).$$

Repeating this procedure ad infinitum, we produce a sequence of balls $B(y_k, \varepsilon_{k+1})$, $B(y_0, \varepsilon_1) \supset B(y_1, \varepsilon_2) \supset B(y_2, \varepsilon_3) \supset \cdots$ and corresponding members x_{n_k}, $n_{k+1} > n_k$, of the sequence $\{x_n\}$ such that

$$(x_{n_k}, y) > k \text{ for all } y \in B(y_k, \varepsilon_{k+1}).$$

Since H is a Hilbert space there is at least one element y^* which belongs to every $B(y_k, \varepsilon_{k+1})$, so

$$(x_{n_k}, y^*) > k.$$

Thus we find a continuous linear functional $F^*(x) = (x, y^*)$ for which the numerical sequence $\{F^*(x_{n_k})\}$ is not a Cauchy sequence. This contradicts the definition of weak convergence. □

Using this proof we can get an important result:

Lemma 1.23.1. Assume that $\{x_k\}$ is an unbounded sequence in H:

$$\|x_k\| \to \infty.$$

Then there exists $y^* \in H$ and a subsequence $\{x_{n_k}\}$ such that

$$(x_{n_k}, y^*) \to \infty \text{ as } k \to \infty.$$

Proof. Let us introduce the supplementary sequence $z_n = x_n/\|x_n\|$. For any y with unit norm, the numerical sequence (z_n, y) is bounded and thus we can select a convergent subsequence from it. If there exists such a unit element y^* and a subsequence $\{z_{n_k}\}$ for which $(z_{n_k}, y^*) \to a \neq 0$, then the statement of the lemma is valid for the subsequence $\{x_{n_k}\}$ and y^* if $a > 0$; when $a < 0$ then y^* must be changed to $-y^*$. Indeed, if $a > 0$ then $(x_{n_k}, y^*) = (z_{n_k}, y^*)\|x_{n_k}\| \to \infty$.

Now we consider the case when we cannot find such an element y^* and a subsequence $\{z_{n_k}\}$ for which $a \neq 0$, so $(z_n, y) \to 0$ for any $y \in H$. By the Riesz representation theorem, this means that $\{z_n\}$ converges weakly to zero. We will demonstrate the statement of Lemma 1.23.1 to hold for the latter class of sequences as well. For this we repeat two steps of the proof of Theorem 1.23.2.

First we show that for any center y_0 and radius ε, the numerical set (x_n, y) with y running over $B(y_0, \varepsilon)$ is unbounded. Indeed, taking the sequence $y_n = y_0 + \varepsilon/(2\|x_n\|)\, x_n$ we get an element from $B(y_0, \varepsilon)$. Next,

$$(x_n, y_n) = (x_n, y_0) + \frac{\varepsilon}{2\|x_n\|}(x_n, x_n) = \left((z_n, y_0) + \frac{\varepsilon}{2}\right)\|x_n\|.$$

Since ε is finite and $(z_n, y_0) \to 0$ as $n \to \infty$, we have $(x_n, y_n) \to \infty$.

Another step of the proof of Theorem 1.23.2, establishing the existence of a subsequence $\{x_{n_k}\}$ and an element y^* such that $(x_{n_k}, y^*) \to \infty$, requires only that $\|x_n\| \to \infty$ and that for any $\varepsilon > 0$ the set (x_n, y) is unbounded when y runs over $B(y_0, \varepsilon)$, which was just proved. Thus we immediately state the validity of Lemma 1.23.1 for all the unbounded sequences. □

This is used in proving an important theorem called the *principle of uniform boundedness*, which is

Theorem 1.23.3. Let $\{F_k(x)\}$, $k = 1, 2, \ldots$, be a family of continuous linear functionals defined on a Hilbert space H. If $\sup_k |F_k(x)| < \infty$, then $\sup_k \|F_k\| < \infty$.

Proof. By the Riesz representation theorem, each of the functionals $F_k(x)$ has the form

$$F_k(x) = (x, f_k), \quad \text{where } f_k \in H, \quad \|f_k\| = \|F_k\|.$$

So the condition of the theorem can be rewritten as

$$\sup_k |(x, f_k)| < \infty. \tag{1.23.2}$$

By Lemma 1.23.1, the assumption that $\sup_k \|f_k\| = \infty$ implies the existence of $x_0 \in H$ and $\{f_{k_n}\}$ such that

$$|(x_0, f_{k_n})| \to \infty \text{ as } k \to \infty.$$

This contradicts (1.23.2). □

Corollary 1.23.1. Let $\{F_k(x)\}$ be a sequence of continuous linear functionals given on H, such that for every $x \in H$ the numerical sequence $\{F_k(x)\}$ is a Cauchy sequence. Then there is a continuous linear functional $F(x)$ on H such that

$$F(x) = \lim_{k\to\infty} F_k(x) \text{ for all } x \in H \tag{1.23.3}$$

and

$$\|F\| \le \liminf_{k\to\infty} \|F_k\| < \infty. \tag{1.23.4}$$

Proof. The limit on the right-hand side of (1.23.3), existing by the condition, defines a functional $F(x)$ which is clearly linear. Since the condition of Theorem 1.23.3 is met, we have $\sup_k \|F_k\| < \infty$; from

$$|F(x)| = \lim_{k\to\infty} |F_k(x)| \le \sup_k \|F_k\| \, \|x\|$$

it follows that $F(x)$ is continuous. Moreover,

$$|F(x)| = \lim_{k\to\infty} |F_k(x)| \le \liminf_{k\to\infty} \|F_k\| \, \|x\|,$$

i.e., (1.23.4) is proved also. □

The following theorem gives an equivalent, but more convenient, definition of weak convergence:

Theorem 1.23.4. A sequence $\{x_n\}$ is weakly Cauchy in a Hilbert space H if and only if the following pair of conditions holds:

(i) $\{x_n\}$ is bounded in H, i.e., there is a constant M such that $\|x_n\| \le M$;

(ii) for any $f_\alpha \in H$ from a system $\{f_\alpha\}$ which is complete in H, the numerical sequence (x_n, f_α) is a Cauchy sequence.

Proof. Necessity of the conditions follows from the definition of weak convergence and Theorem 1.23.2.

Now we prove sufficiency. Suppose the conditions (i) and (ii) hold. Take an arbitrary continuous linear functional defined, by the Riesz representation theorem, by an element $f \in H$ and consider the numerical sequence

$$d_{nm} = (x_n, f) - (x_m, f).$$

As the system $\{f_\alpha\}$ is complete, there is a linear combination

$$f_\varepsilon = \sum_{k=1}^{N} c_k f_k$$

such that

$$\|f - f_\varepsilon\| < \varepsilon/3M.$$

Then

$$\begin{aligned} |d_{nm}| &= |(x_n - x_m, f)| \\ &= |(x_n - x_m, f_\varepsilon + f - f_\varepsilon)| \\ &\le |(x_n - x_m, f_\varepsilon)| + |(x_n - x_m, f - f_\varepsilon)| \\ &\le \sum_{k=1}^{N} |c_k||(x_n - x_m, f_k)| + (\|x_n\| + \|x_m\|)\|f - f_\varepsilon\|. \end{aligned}$$

Since, by (ii), the sequences $\{(x_n, f_k)\}$, $k = 1, \ldots, N$, are Cauchy sequences, we can find a number R such that

$$\sum_{k=1}^{N} |c_k||(x_n - x_m, f_k)| < \varepsilon/3 \quad \text{when } m, n > R$$

hence

$$|d_{nm}| \le \varepsilon/3 + 2M\varepsilon/(3M) = \varepsilon \quad \text{for } m, n > R.$$

This means that $\{(x_n, f)\}$ is a Cauchy sequence. □

Problem 1.23.1. Show that a sequence $\{x_n\}$ is weakly convergent to x_0 in H if and only if the following pair of conditions holds:

(i) $\{x_n\}$ is bounded in H;

(ii) for any f_α from a system $\{f_\alpha\}$, $f_\alpha \in H$, which is complete in H, we have $\lim_{n\to\infty}(x_n, f_\alpha) = (x_0, f_\alpha)$.

Because weak convergence differs from strong convergence, we are led to consider weak completeness of a Hilbert space.

Theorem 1.23.5. Any weak Cauchy sequence $\{x_n\}$ in a Hilbert space converges weakly to an element of this space.

Proof. For any fixed $y \in H$ we define $F(y) = \lim_{n\to\infty}(y, x_n)$. The functional $F(y)$, whose linearity is evident, is defined on the whole of H. From the inequality

$$|(y, x_n)| \le M\|y\|,$$

M being a constant such that $\|x_n\| \le M$, it follows that

$$|F(y)| \le M\|y\| \qquad \text{and} \qquad \|F\| \le M.$$

Therefore $F(y)$ is a continuous linear functional which, by the Riesz representation theorem, can be written in the form

$$F(y) = (y, f), \qquad f \in H, \quad \|f\| = \|F\| \le M.$$

But this means that f is a weak limit of $\{x_n\}$. □

From this proof also follows

Lemma 1.23.2. If a sequence $\{x_n\} \subset H$ converges weakly to x_0 in H and $\|x_n\| \le M$ for all n, then $\|x_0\| \le M$.

This states that a closed ball about zero is weakly closed. Any closed subspace of a Hilbert space is also weakly closed. Mazur's theorem states that any closed convex set in a Hilbert space is weakly closed. The interested reader can find a proof in Yosida [29].

Theorem 1.23.6. Assume that a sequence $\{x_n\}$ in a Hilbert space H converges weakly to $x_0 \in H$. Then there is a subsequence $\{x_{n_k}\}$ of $\{x_n\}$ such that the sequence of arithmetic means $\frac{1}{N}\sum_{k=1}^{N} x_{n_k}$ converges strongly to x_0.

We now consider the problem of weak compactness of a set in a Hilbert space. We have seen that a ball in an infinite dimensional Hilbert space is not strongly compact. But for weak compactness an analog of the Bolzano–Weierstrass theorem holds as follows:

Theorem 1.23.7. A bounded sequence $\{x_n\}$ in a separable Hilbert space contains a weak Cauchy subsequence.

In other words, a bounded set in a Hilbert space is weakly precompact.

Proof. In a separable Hilbert space there is an orthonormal basis $\{g_n\}$. By Theorem 1.23.4 it suffices to show that there is a subsequence $\{x_{n_k}\}$ such that, for fixed g_m, the numerical sequence $\{(x_{n_k}, g_m)\}$ is a Cauchy sequence.

The bounded numerical sequence $\{(x_n, g_1)\}$ contains a convergent subsequence $\{(x_{n_1}, g_1)\}$. Considering the numerical sequence $\{(x_{n_1}, g_2)\}$, for the same reason we can choose a convergent sequence $\{(x_{n_2}, g_2)\}$. Continuing this process, on the kth step we obtain a convergent numerical subsequence $\{(x_{n_k}, g_k)\}$.

Choosing now the elements x_{n_n}, we obtain a sequence $\{x_{n_n}\}$ such that for any fixed g_m the numerical sequence $\{(x_{n_n}, g_m)\}$ is a Cauchy sequence. That is, $\{x_{n_n}\}$ is a weak Cauchy sequence. □

This theorem has important applications; in the justification of some numerical methods we can sometimes prove boundedness of the set of approximate solutions in a Hilbert (as a rule, energy) space.

Let us demonstrate this procedure on the example of the problem of approximation, namely, we want to find a minimizer of a functional

$$F(x) = \|x - x_0\|^2$$

given on a real Hilbert space when x_0 is a fixed element of H, $x_0 \notin M$, and x is an arbitrary element of a closed subspace $M \subset H$.

In Section 1.18 we established the existence of a minimizer of $F(x)$. We now treat this problem once more, as though this existence were unknown to us.

This very simple problem (at least in theory) exhibits the following typical steps, which are common for the justification of approximate solutions to many boundary value problems:

1. the formulation of an approximation problem and the demonstration of its solvability;

2. a global *a priori* estimate of the approximate solutions that does not depend on the step of approximation;

3. the demonstration of convergence of the approximate solutions to a solution of the initial problem, and a study of the nature of convergence.

Thus we begin to study our problem with the formulation of the approximation problem.

We try to solve the problem approximately, using the so-called Ritz method. Assume that $\{g_k\}$ is a complete system in M such that any of its finite subsystems in linearly independent. Consider M_n spanned by $(g_1, \dots, g_n)$ and find an element which minimizes $F(x)$ on M_n. A solution of this problem, denoted by x_n, is the so-called nth Ritz approximation of the solution.

A real-valued function $f(t) = F(x_n + tg_k)$ of the real variable t takes its minimal value at $t = 0$ and, thanks to differentiability of $f(t)$,

$$\left.\frac{df(t)}{dt}\right|_{t=0} = 0.$$

This gives an equation

$$\begin{aligned} 0 &= \frac{d}{dt}\|x_n - x_0 + tg_k\|^2\big|_{t=0} \\ &= \frac{d}{dt}(x_n - x_0 + tg_k, x_n - x_0 + tg_k)\big|_{t=0} \\ &= 2(x_n - x_0, g_k) \end{aligned}$$

so $x_n - x_0$ is orthogonal to each g_k, $k = 1, \ldots, n$.

Representing $x_n = \sum_{k=1}^{n} c_{kn} g_k$, we get a linear system of algebraic equations which is called the Ritz system of nth approximation:

$$\sum_{k=1}^{n} c_{kn}(g_k, g_m) = (x_0, g_m), \qquad m = 1, \ldots, n. \tag{1.23.5}$$

The determinant of this system is the Gram determinant of a linearly independent system $(g_1, \ldots, g_m)$ (not equal to zero) so the system (1.23.5) has a unique solution.

Now we will find a global estimate of the approximate solutions that does not depend on n. Although we know here that the approximate solution exists, we can get the estimate without this knowledge. That is why such estimates are called *a priori* estimates.

We begin with the definition of x_n:

$$\|x_n - x_0\|^2 \le \|x - x_0\|^2 \text{ for all } x \in M_n.$$

As $x = 0 \in M_n$, it follows that

$$\|x_n - x_0\|^2 \le \|x_0\|^2,$$

from which

$$\|x_n\|^2 \le 2\|x_n\| \, \|x_0\|,$$

hence

$$\|x_n\| \le 2\|x_0\|. \tag{1.23.6}$$

This is the required estimate.

Remark 1.23.1. It is possible to get a sharper estimate than (1.23.7); however, for this problem it is necessary to establish only the existence of a bound.

Our last goal is to demonstrate that the sequence of approximations converges to a solution of the problem. First we demonstrate that this convergence is weak, and then that it is strong.

By (1.23.6), the sequence $\{x_n\}$ is bounded and, thanks to Theorem 1.23.7, contains a weakly convergent subsequence $\{x_{n_k}\}$ whose weak limit x^* belongs to M (a closed subspace is weakly closed).

For any fixed g_m, we can pass to the limit as $k \to \infty$ in the equality

$$(x_{n_k} - x_0, g_m) = 0$$

and get

$$(x^* - x_0, g_m) = 0$$

(this is allowed since (x, g_m) is a continuous linear functional in x).

Now consider $(x^* - x_0, h)$ where h is an arbitrary but fixed element of M. By completeness of the system $g_1, g_2, g_3, \dots$ in M, given $\varepsilon > 0$ we can find a finite linear combination $h_\varepsilon = \sum_{k=1}^{N} c_k g_k$ such that

$$\|h - h_\varepsilon\| \le \varepsilon/(3\|x_0\|).$$

Then

$$\begin{aligned} |(x^* - x_0, h)| &= |(x^* - x_0, h - h_\varepsilon + h_\varepsilon)| \\ &\le |(x^* - x_0, h - h_\varepsilon)| + |(x^* - x_0, h_\varepsilon)| \\ &= |(x^* - x_0, h - h_\varepsilon)| \\ &\le \|x^* - x_0\| \, \|h - h_\varepsilon\| \\ &\le (\|x^*\| + \|x_0\|)\|h - h_\varepsilon\| \\ &\le (2\|x_0\| + \|x_0\|)\varepsilon/(3\|x_0\|) \\ &= \varepsilon. \end{aligned}$$

Therefore, for any $h \in M$ we get

$$(x^* - x_0, h) = 0. \tag{1.23.7}$$

Finally, considering values of $F(x) = \|x - x_0\|^2$ on elements of the form $x = x^* + h$ when $h \in M$, we obtain, thanks to (1.23.7),

$$\begin{aligned} F(x^* + h) &= (x^* - x_0 + h, x^* - x_0 + h) \\ &= \|x^* - x_0\|^2 + 2(x^* - x_0, h) + \|h\|^2 \\ &= \|x^* - x_0\|^2 + \|h\|^2 \\ &\ge \|x^* - x_0\|^2 \\ &= F(x^*). \end{aligned}$$

It follows that x^* is a solution of the problem, and existence of solution has been proved.

Now we can show that the approximation sequence converges strongly to a solution of the problem. By Theorem 1.18.3, a minimizer of $F(x)$ is unique; this gives us weak convergence of the sequence $\{x_n\}$ on the whole. Indeed, suppose to the contrary that $\{x_n\}$ does not converge weakly to x^*. Then there is an element $f \in H$ such that

$$(x_n, f) \nrightarrow (x^*, f). \tag{1.23.8}$$

By boundedness of the numerical set $\{(x_n, f)\}$, the statement (1.23.8) implies that there is a subsequence $\{x_{n_k}\}$ such that there exists

$$\lim_{k\to\infty}(x_{n_k}, f) \neq (x^*, f). \tag{1.23.9}$$

But for the subsequence $\{x_{n_k}\}$ we can repeat Step 3 and find that it contains a subsequence which converges weakly to a solution of the problem. Since the solution is unique, this contradicts (1.23.9). Finally, multiplying both sides of (1.23.5) by the Ritz coefficient c_{mn} and summing over m, we get

$$(x_n, x_n) = (x_0, x_n).$$

We can pass to the limit as $n \to \infty$, obtaining

$$\lim_{n\to\infty}(x_n, x_n) = \lim_{n\to\infty}(x_0, x_n) = (x_0, x^*).$$

By (1.23.7) with $h = x^*$

$$(x_0, x^*) = (x^*, x^*),$$

so

$$\lim_{n\to\infty}\|x_n\|^2 = \|x^*\|^2.$$

Therefore, by Theorem 1.23.1, the sequence $\{x_n\}$ converges to x^* strongly.

So we have demonstrated, via the Ritz method, a general way of justifying the solution of a minimal problem and the Ritz method itself. The method is common to a wide variety of problems, some nonlinear. In the latter case, many difficulties center on Steps 2 or 3, depending on the problem. The problem under discussion can also be interpreted another way, and this is of so much importance that we devote a separate section to it.

1.24 The Ritz and Bubnov–Galerkin Methods in Linear Problems

Consider once more the problem of minimizing the quadratic functional (1.14.8) in a Hilbert space, namely,

$$I(x) = \|x\|^2 + 2\Phi(x) \to \min_{x\in H}. \tag{1.24.1}$$

Assuming that $\Phi(x)$ is a continuous linear functional, by the Riesz representation theorem we have

$$\Phi(x) = (x, -x_0)$$

where x_0 is a unique element of H defined by $\Phi(x)$. Then

$$I(x) = \|x\|^2 - 2(x, x_0) = \|x - x_0\|^2 - \|x_0\|^2.$$

Since $\|x_0\|^2$ is fixed, the problem (1.14.1) is equivalent to

$$F(x) = \|x - x_0\|^2 \to \min_{x \in H}.$$

This problem has the unique (and obvious) solution $x = x_0$. Of much interest is the fact that it coincides with the problem of the previous section if $M = H$. So application of the Ritz method in this problem is justified. Let us recall those results in terms of the new problem.

Let $g_1, g_2, g_3, \ldots$ be a complete system in H, every finite subsystem of which is linearly independent, and let the nth Ritz approximation to a minimizer be $x_n = \sum_{k=1}^{n} c_{kn} g_k$. The system giving the nth approximation of the Ritz method is

$$\sum_{k=1}^{n} c_{kn}(g_k, g_m) = -\Phi(g_m), \qquad m = 1, \ldots, n. \tag{1.24.2}$$

Let us collect the results in

Theorem 1.24.1. (i) For each $n \geq 1$ the system (1.24.2) of nth approximation of the Ritz method has the unique solution $c_{1n}, \ldots, c_{nn}$.

(ii) The sequence $\{x_n\}$ of Ritz approximations defined by (1.24.2) converges strongly to the minimizer of the quadratic functional $\|x\|^2 + 2\Phi(x)$, $\Phi(x)$ being a continuous linear functional on H.

It is interesting to note that if $g_1, g_2, g_3, \ldots$ is an orthonormal basis of H, then (1.24.2) gives the Fourier coefficients of the solution (in energy space).

As to Bubnov's method, we only mention that it appeared when A.S. Bubnov, reviewing an article by S. Timoshenko, noted that the Ritz equations can be obtained by multiplying by g_m, a function of a complete system, the differential equation of equilibrium in which u was replaced by $u_n = \sum_{k=1}^{n} c_{kn} g_k$, integrating the latter over the region, and then integrating by parts. In our terms this is

$$(u_n, g_m) = -\Phi(g_m), \qquad m = 1, \ldots, n.$$

Since this system indeed coincides with (1.24.2), Theorem 1.24.1 also justifies Bubnov's method.

Galerkin was the first to propose multiplying by f_m, a function of another system, for better approximation of the residual. The corresponding system is, in our notation,

$$(u_n, f_m) = -\Phi(f_m), \qquad m = 1, \ldots, n.$$

Discussion of this modification of the method can be found in Mikhlin [20].

Finally, we note that the finite element method for solution of mechanics problems is a particular case of the Bubnov–Galerkin method, hence it is also justified for the problems we consider.

1.25 Curvilinear Coordinates, Non-Homogeneous Boundary Conditions

We have considered some problems of mechanics using the Cartesian coordinate system. Almost all of the textbooks present the theory of the same problems in Cartesian frames; the few exceptions are the textbooks on the theory of shells and curvilinear bars, where it is impossible to consider the problems in Cartesian frames. However, in practice other coordinate systems are used quite frequently. The question arises whether it is necessary to investigate the boundary value problems for other coordinates, or whether it is enough to reformulate the results for Cartesian systems. For the generalized statement of the problems of mechanics in energy spaces, the answer is simple: it is possible to reformulate the results, and a key tool is a simple change of the coordinates. This change allows us to reformulate the imbedding theorems in energy spaces, to establish the requirements for admitting classes of loads, etc. We note that it is a hard problem to obtain similar results independently, without the use of coordinate transformations, if the coordinate frame has singular points.

Let us illustrate the above on a simple example of a circular membrane with fixed edge (Dirichlet problem). In Cartesian coordinates we have the Sobolev imbedding theorem

$$\left(\iint\limits_{\Omega} |u(x)|^p \, dx\, dy\right)^{1/p} \le m \left(\iint\limits_{\Omega} \left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] dx\, dy\right)^{1/2} \tag{1.25.1}$$

for $p \ge 1$, which is valid for any $u \in \dot{W}^{1,2}(\Omega) \equiv E_{MC}$ satisfying the boundary condition

$$u\big|_{\partial\Omega} = 0. \tag{1.25.2}$$

Taking a function $u \in C^{(1)}(\Omega)$ satisfying (1.25.2), in both integrals of (1.25.1) we pass to the polar coordinate system:

$$\left(\int_0^R \int_0^{2\pi} |u|^p r \, d\phi \, dr\right)^{1/p} \le$$
$$m \left(\int_0^R \int_0^{2\pi} \left[\left(\frac{\partial u}{\partial r}\right)^2 + \frac{1}{r^2}\left(\frac{\partial u}{\partial \phi}\right)^2\right] r \, d\phi \, dr\right)^{1/2} \tag{1.25.3}$$

where (r, ϕ) are the polar coordinates in a disk of radius R. Passing to the limit along a Cauchy sequence of E_{MC} in the inequality (1.25.1), which is valid in Cartesian coordinates, shows us that it remains valid in the form (1.25.3) in polar coordinates. Inequality (1.25.3) is an imbedding theorem in the energy space of the circular membrane in terms of polar coordinates.

The expression

$$\|u\| = \left(\int_0^R \int_0^{2\pi} \left[\left(\frac{\partial u}{\partial r} \right)^2 + \frac{1}{r^2} \left(\frac{\partial u}{\partial \phi} \right)^2 \right] r \, d\phi \, dr \right)^{1/2} \tag{1.25.4}$$

is the norm in this coordinate system, whereas

$$(u, v) = \int_0^R \int_0^{2\pi} \left(\frac{\partial u}{\partial r} \frac{\partial v}{\partial r} + \frac{1}{r^2} \frac{\partial u}{\partial \phi} \frac{\partial v}{\partial \phi} \right) r \, d\phi \, dr$$

is the corresponding inner product.

The requirement imposed on forces for existence of a generalized solution to the problem has the form

$$\int_0^R \int_0^{2\pi} |F|^q r \, d\phi \, dr < \infty, \quad q > 1.$$

We have a natural form of the norm in the energy space (which is determined by the energy itself) using curvilinear coordinates, as well as a form of the imbedding theorem (i.e., properties of elements of the energy space and natural requirements on forces for the problem to be uniquely solvable).

Then we note that we can replace formally the Cartesian system by any other system of coordinates which is admissible for smooth functions, and also change formally any variables in any expression which makes sense in the energy space considered in Cartesian coordinates.

Finally, note that a norm like (1.25.4) is usually called a weighted norm because of the presence of weight factors, here connected with powers of r. There is an abstract theory of such weighted Sobolev spaces, not being so elementary as in the space we have considered.

For more complicated problems such as problems of elasticity, we can use the same method of introducing curvilinear coordinates; here we can change not only the independent variables (x_1, x_2, x_3), but also unknown components of vectors of displacements and prescribed forces, to the new coordinate system. We leave it to the reader to write down an equation determining a generalized solution, the forms of norm and scalar product, and restrictions for forces as well as imbedding inequalities, in other curvilinear coordinate systems such as cylindrical and spherical.

Now let us consider two questions connected with non-homogeneous boundary value problems in mechanics. The first is to identify the whole class of admissible external forces for which an energy solution exists. We know that the condition for existence of a solution is that the functional of external forces

$$\int_\Omega F(\mathbf{x}) v(\mathbf{x}) \, d\Omega \tag{1.25.5}$$

(say, in the membrane problem) is continuous and linear with respect to $v(\mathbf{x})$ on an energy space. We shall show how this condition can be expressed in terms of so-called spaces with negative norms, a notion due to P.D. Lax [17].

The functional (1.25.5) can be considered as the scalar product of $F(\mathbf{x})$ by $v(\mathbf{x})$ in $L^2(\Omega)$. But $v(\mathbf{x})$ belongs to an energy space E whose norm, for simplicity, is assumed to be such that $\|v\|_E = 0$ implies $v = 0$. We know that $v \in L^2(\Omega)$ if $v \in E$; moreover, E is dense in $L^2(\Omega)$. For any $F(\mathbf{x}) \in L^2(\Omega)$, we can introduce a new norm

$$\|F\|_E = \sup_{\|v\|_E \le 1} \left| \int_\Omega F(\mathbf{x}) v(\mathbf{x})\, d\Omega \right|.$$

It is clear that $L^2(\Omega)$ with this norm is not complete (since all $v \in L^p(\Omega)$ for any $p > 2$, $p < \infty$). The completion of $L^2(\Omega)$ in the norm $\|\cdot\|_E$ is called the space with negative norm, denoted E^-. In Lax [17] (and in other books, for example, Yosida [29]) it is shown that the set of all continuous linear functionals on E can be identified with E^- since E is dense in $L^2(\Omega)$.

So the condition $F(\mathbf{x}) \in E^-$ is necessary and sufficient for the work functional (1.25.5) to be continuous with respect to $v(\mathbf{x})$ on E.

In Lax [17], such a construction was introduced for a Sobolev space $\dot{W}^{k,2}(\Omega)$; the corresponding space with negative norm was denoted by $W^{-k,2}(\Omega)$. An equivalent approach to the introduction of $W^{-k,2}(\Omega)$ involves use of the Fourier transformation in Sobolev spaces (cf., Yosida [29]).

The notion of the space with negative norm is a useful tool in the study of problems, but it is not too informative when we want to know whether certain forces are of a needed class; here sufficient conditions are more convenient.

The second question we consider is what to do when boundary conditions (of Dirichlet type) are non-homogeneous. Consider, for example, the problem

$$-\Delta v = F, \tag{1.25.6}$$

$$v\big|_{\partial\Omega} = \varphi. \tag{1.25.7}$$

We can try the classical approach to the treatment of this problem, finding a function $v_0(\mathbf{x})$ satisfying (1.25.7), i.e.,

$$v_0\big|_{\partial\Omega} = \varphi.$$

Now we are seeking $v(\mathbf{x})$ in the form $v = u + v_0$, where $u(\mathbf{x})$ satisfies the homogeneous boundary condition

$$u\big|_{\partial\Omega} = 0. \tag{1.25.8}$$

An integro-differential equation of equilibrium of the membrane is

$$\iint\limits_{\Omega} \left(\frac{\partial u}{\partial x}\frac{\partial \psi}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial \psi}{\partial y} \right) d\Omega + \iint\limits_{\Omega} \left(\frac{\partial v_0}{\partial x}\frac{\partial \psi}{\partial x} + \frac{\partial v_0}{\partial y}\frac{\partial \psi}{\partial y} \right) d\Omega = \iint\limits_{\Omega} F\psi \, d\Omega. \tag{1.25.9}$$

wherein virtual displacements must also satisfy (1.25.8):

$$\psi|_{\partial\Omega} = 0.$$

Considering the term

$$\iint\limits_{\Omega} \left(\frac{\partial v_0}{\partial x}\frac{\partial \psi}{\partial x} + \frac{\partial v_0}{\partial y}\frac{\partial \psi}{\partial y} \right) d\Omega$$

we see that it is a continuous linear functional on E_{MC} if $\partial v_0/\partial x$ and $\partial v_0/\partial y$ belong to $L^2(\Omega)$. In such a case there is a generalized solution to the problem, i.e., $u \in E_{MC}$ satisfying (1.25.9) for any $\psi \in E_{MC}$.

We have supposed that there exists an element of $W^{1,2}(\Omega)$ satisfying the boundary condition (1.25.7). In more detailed textbooks on the theory of partial differential equations, one may find the conditions for a function φ given on the boundary that are sufficient for the existence of the function v_0. Corresponding theorems for v_0 from Sobolev spaces are called trace theorems. The trace theorems suppose the boundary to be sufficiently smooth. The case of a piecewise smooth boundary that is frequently encountered in practice has not been completely studied as of yet. The problem of the trace of functions is beyond the scope of this book.

A final remark. In mathematics we normally deal with dimensionless quantities. In this presentation we have also supposed all quantities to be dimensionless. However, variables having dimensional units can be used without difficulty, provided we check carefully for units in all inequalities and equations, and introduce additional factors as may be required. In particular, in imbedding theorems the constants normally carry dimensional units, hence these constants change if the units are changed.

1.26 The Bramble–Hilbert Lemma and Its Applications

This lemma is widely used to establish the rate of convergence of the finite element method (see, for example, Ciarlet [6]). The lemma gives a bound for a functional with special properties in a Sobolev space. We would like to note that sometimes it is useful to read classical books because, for example, the reader (as well as the authors of the lemma) could find that the lemma is a simple consequence of the theorem on equivalent norming of $W^{l,p}(\Omega)$ in Sobolev [22].

Recall the Poincaré inequality (1.10.9)

$$\int_S u^2\,dS \le m\left\{\left(\int_S u\,dS\right)^2 + \int_S\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right]dS\right\}, \quad (1.26.1)$$

which was derived when S was the square $[0,a]\times[0,a]$.

The proof of (1.10.9) is easily extended to the case of an n-dimensional cube. We now discuss how to extend it to a compact set Ω which is star-shaped with respect to a square S; that is, any ray starting in S intersects the boundary of Ω exactly once. We shall establish the following estimate, which is also called the Poincaré inequality:

$$\int_\Omega u^2\,d\Omega \le m_1\left(\int_S u\,dS\right)^2 + m_2\int_\Omega\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right]d\Omega. \quad (1.26.2)$$

Let us rewrite this in a system of polar coordinates (r,ϕ) whose origin is at the center of S. Let $\partial\Omega$ be given by the equation $r = R(\phi) \ge a/2$, $R(\phi) < R_0$. Then (1.26.1) has the form

$$\int_0^{2\pi}\int_0^{R(\phi)} u^2 r\,dr\,d\phi \le m_1\left(\int_S u\,dS\right)^2 + \\ + m_2\int_0^{2\pi}\int_0^{R(\phi)}\left[\left(\frac{\partial u}{\partial r}\right)^2 + \frac{1}{r^2}\left(\frac{\partial u}{\partial\phi}\right)^2\right] r\,dr\,d\phi.$$

Because of (1.26.1), it follows that it is sufficient to get the estimate

$$\int_0^{2\pi}\int_{a/2}^{R(\phi)} u^2 r\,dr\,d\phi \le m_3\int_0^{2\pi}\int_{a/4}^{a/2} u^2 r\,dr\,d\phi + \\ + m_4\int_0^{2\pi}\int_{a/4}^{R(\phi)} r\left(\frac{\partial u}{\partial r}\right)^2 dr\,d\phi \quad (1.26.3)$$

with constants which are independent of $u\in C^{(1)}(\Omega)$ ($C^{(1)}(\Omega)$ is introduced in Cartesian coordinates!). We now proceed to prove this.

The starting point is the representation

$$u(r_2,\phi) = u(r_1,\phi) + \int_{r_1}^{r_2}\frac{\partial u(r,\phi)}{\partial r}\,dr, \qquad a/4 \le r_1 \le a/2,\ a/4 \le r_2 \le R_0,$$

from which, by squaring both sides and applying elementary transformations, we get

$$\begin{aligned} u^2(r_2,\phi) &\le 2u^2(r_1,\phi) + 2\left(\int_{r_1}^{r_2} \frac{1}{\sqrt{r}}\left(\sqrt{r}\frac{\partial u}{\partial r}\right) dr\right)^2 \\ &\le 2u^2(r_1,\phi) + 2\int_{r_1}^{r_2}\frac{dr}{r}\int_{r_1}^{r_2} r\left(\frac{\partial u}{\partial r}\right)^2 dr \\ &\le 2u^2(r_1,\phi) + m_5\int_{a/4}^{R(\phi)} r\left(\frac{\partial u}{\partial r}\right)^2 dr, \quad m_5 = 2\ln\frac{4R_0}{a}. \end{aligned}$$

Multiplying this chain of inequalities by $r_1 r_2$ and then integrating it first with respect to r_2 from $a/2$ to $R(\phi)$ and then with respect to r_1 from $a/4$ to $a/2$ gives

$$\begin{aligned} \int_{a/4}^{a/2} r_1 \int_{a/2}^{R(\phi)} u^2(r_2,\phi) r_2\, dr_2\, dr_1 \le 2\int_{a/4}^{a/2} u^2(r_1,\phi) r_1\, dr_1 \int_{a/2}^{R(\phi)} r_2\, dr_2 + \\ + m_5 \int_{a/4}^{a/2}\int_{a/2}^{R(\phi)} r_1 r_2\, dr_1\, dr_2 \int_{a/4}^{R(\phi)} r\left(\frac{\partial u}{\partial r}\right)^2 dr \end{aligned}$$

or

$$\begin{aligned} \frac{3}{32}a^2\int_{a/2}^{R(\phi)} u^2(r,\phi) r\, dr \le R_0^2\int_{a/4}^{a/2} u^2(r,\phi) r\, dr + \\ + \frac{3}{64}a^2R_0^2 m_5\int_{a/4}^{R(\phi)} r\left(\frac{\partial u}{\partial r}\right)^2 dr. \end{aligned}$$

Finally, integrating this with respect to ϕ over $[0, 2\pi]$ and multiplying it by $32/(3a^2)$ completes the proof of (1.26.3) and hence of (1.26.2).

We can similarly extend the Poincaré inequality to the case of a multiconnected domain Ω which is a union of star-shaped domains, and to the case of an n-dimensional domain Ω with $n > 2$. The latter extension is

$$\int_\Omega u^2\, d\Omega \le m_1\left(\int_C u\, d\Omega\right)^2 + m_2\sum_{i=1}^n \int_\Omega \left(\frac{\partial u}{\partial x_i}\right)^2 d\Omega, \tag{1.26.4}$$

where $C \subset \Omega$ is a hypercube in $\mathbb{R}^n$.

We can apply the inequality (1.26.4) to any derivative $D^\alpha u$, $|\alpha| < k$. Combining these estimates successively, we derive the inequality needed for the proof of the Bramble–Hilbert lemma

$$\|u\|^2_{W^{k,2}(\Omega)} \le m_3 \sum_{0\le|\alpha|<k}\left(\int_C D^\alpha u\, d\Omega\right)^2 + m_4\sum_{|\alpha|=k}\int_\Omega |D^\alpha u|^2\, d\Omega. \tag{1.26.5}$$

This estimate permits us to introduce another form of equivalent norm in $W^{k,2}(\Omega)$. (Question to the reader: Which one?) Note that the estimate

was obtained for functions of $C^{(k)}(\Omega)$, but the now standard procedure of completion provides that it is valid for any $u \in W^{k,2}(\Omega)$.

Lemma 1.26.1 (Bramble–Hilbert [5]). Assume that $F(u)$ is a continuous linear functional on $W^{k,2}(\Omega)$ such that for any polynomial $P_r(\mathbf{x})$ of order less than k,

$$F(P_r(\mathbf{x})) = 0. \tag{1.26.6}$$

Then there is a constant m^* depending only on Ω such that

$$|F(u)| \le m^* \|F\|_{W^{k,2}(\Omega)} \left(\sum_{|\alpha|=k} \int_\Omega |D^\alpha u|^2 \, d\Omega \right)^{1/2}. \tag{1.26.7}$$

Proof. From (1.26.5) and continuity of $F(u)$ on $W^{k,2}(\Omega)$, it follows that

$$|F(u)| \le m \|F\|_{W^{k,2}(\Omega)} \left[\sum_{0 \le |\alpha| < k} \left(\int_C D^\alpha u \, d\Omega \right)^2 + \sum_{|\alpha|=k} \int_\Omega |D^\alpha u|^2 \, d\Omega \right]^{1/2}. \tag{1.26.8}$$

By (1.26.6),

$$F(u(\mathbf{x}) + P_{k-1}(\mathbf{x})) = F(u(\mathbf{x}))$$

where $P_{k-1}(\mathbf{x})$ is an arbitrary polynomial of order $k-1$. Fixing $u(\mathbf{x}) \in W^{k,2}(\Omega)$, we can always choose a polynomial $P^*_{k-1}(\mathbf{x})$ such that

$$\int_C D^\alpha (u(\mathbf{x}) + P^*_{k-1}(\mathbf{x})) \, d\Omega = 0 \text{ for all } 0 \le |\alpha| \le k-1.$$

Substituting $u(\mathbf{x}) + P^*_{k-1}(\mathbf{x})$ into (1.26.8), we get the needed (1.26.7) since

$$D^\alpha P_{k-1}(\mathbf{x}) = 0 \quad \text{for } |\alpha| = k.$$

□

Let us consider some simple applications of this lemma. Assume that we find numerically, by Simpson's rule,

$$\int_0^1 u(x) \, dx$$

when $u(x) \in W^{2,2}(0,1)$. What is a bound on the error? First we find the error in one step of the trapezoidal rule:

$$F_k(u) = \int_{x_k}^{x_k+h} u(x) \, dx - \frac{h}{2}[u(x_k + h) + u(x_k)].$$

It is clear that $F_k(u)$ is a linear and continuous functional in $W^{2,2}(0,1)$. Making the change of variable $x = x_k + hz$ in the integral, we get

$$|F_k(u)| = h\left|\int_0^1 u(x_k + hz)\,dz - \frac{1}{2}[u(x_k) + u(x_k + h)]\right| \le 2h \max_{z\in[0,1]} |u(x_k + zh)|. \tag{1.26.9}$$

By the elementary inequality

$$\max_{x\in[0,1]} |f(x)| \le \sqrt{2}\left(\int_0^1 \left(f^2(x) + [f'(x)]^2\right)\,dx\right)^{1/2} \le \sqrt{2}\|f\|_{W^{2,2}(0,1)}$$

(whose proof we leave to the reader), (1.26.9) gives

$$|F_k(u)| \le 2\sqrt{2}h\|u(x_k + hz)\|_{W^{2,2}(0,1)}.$$

Since $F_k(a + bx) = 0$ for any constants a, b we can apply the Bramble–Hilbert lemma and obtain

$$\begin{aligned} |F_k(u)| &\le 2\sqrt{2}hm\left(\int_0^1 [u''(x_k + hz)]^2\,dz\right)^{1/2} \\ &= m_1 h^{5/2}\left(\int_{x_k}^{x_k+h} [u''(x)]^2\,dx\right)^{1/2}. \end{aligned}$$

This is the needed error bound for one step of integration.

Consider now the bound on total error when $[0,1]$ is subdivided into N equal parts

$$F(u) = \int_0^1 u(x)\,dx - \frac{h}{2}\sum_{k=0}^{N-1}[u(x_k) + u(x_{k+1})], \qquad x_k = kh.$$

This is linear and continuous in $W^{2,2}(0,1)$, and $f(u) = \sum_{k=0}^{N-1} F_k(u)$. We get

$$\begin{aligned} |F(u)| = \left|\sum_{k=0}^{N-1} F_k(u)\right| &\le \sum_{k=0}^{N-1} |F_k(u)| \\ &\le m_1 h^{5/2}\sum_{k=0}^{N-1}\left(\int_{x_k}^{x_k+h} [u''(x)]^2\,dx\right)^{1/2} \\ &\le m_1 h^{5/2}\sqrt{N}\left(\sum_{k=0}^{N-1}\int_{x_k}^{x_k+h} [u''(x)]^2\,dx\right)^{1/2}. \end{aligned}$$

Thus the needed bound on the error of the trapezoidal rule is

$$|F(u)| \le m_1 h^2 \left(\int_0^1 [u''(x)]^2 \, dx \right)^{1/2}.$$

No improvements in the order of the error result if we take functions smoother than those from $W^{2,2}(0,1)$. But if $v \in W^{1,2}(0,1)$ the bound is worse:

$$|F(v)| \le m_2 h \left(\int_0^1 [v'(x)]^2 \, dx \right)^{1/2}.$$

Another example of the application of Lemma 1.26.1 is given by

Problem 1.26.1. Show that the local error of approximation of the first derivatives of a function $u(x_1, x_2) \in W^{3,2}(\Omega)$, $\Omega \subset \mathbb{R}^2$, by symmetric differences, is

$$\begin{aligned} l(u) &= \left| \frac{\partial u(0,0)}{\partial x_1} - \frac{u(h_1,0) - u(-h_1,0)}{2h_1} \right| + \\ &\quad + \left| \frac{\partial u(0,0)}{\partial x_2} - \frac{u(0,h_2) - u(0,-h_2)}{2h_2} \right| \\ &\le \frac{M(h_1^2 + h_2^2)}{\sqrt{h_1 h_2}} \|u\|_{W^{3,2}(\Omega)} \end{aligned}$$

if $0 < c_1 < h_1/h_2 < c_2 < \infty$. (Take into account that $l(P_2(x_1, x_2)) = 0$ when $P_2(x_1, x_2)$ is a polynomial of second order.)

This page intentionally left blank

2
Elements of the Theory of Operators

2.1 Spaces of Linear Operators

This chapter aims to present in more detail some results of the theory of linear operators. We cannot pretend to give a full treatment of this vast area, and shall select only those parts which are useful in the applications under consideration. Of course, we are forced to give some general theoretical background.

Let A be an operator from a normed space X to a normed space Y. The domain of A is denoted by $D(A)$, and its range by $R(A)$.

In Section 1.12 we saw that a linear operator A defined on X is continuous if and only if there is constant c such that whenever $x \in X$,

$$\|Ax\| \leq c\|x\|.$$

The infimum of all such constants was called $\|A\|$, the norm of A.

Consider the set $L(X, Y)$ of continuous linear operators from X to Y; it is clearly a linear space.

Lemma 2.1.1. $L(X, Y)$ is a normed space.

Proof. We need to check only that the norm axioms are fulfilled for the norm introduced above. $\|A\|$ is clearly non-negative. If $\|A\| = 0$, then $\|Ax\| = 0$ for all $x \in X$, i.e., $A = 0$. Conversely, if $A = 0$ then $\|A\| = 0$. Hence N1 is satisfied. It is obvious that N2 is satisfied. The chain of inequalities

$$\|(A + B)x\| = \|Ax + Bx\| \leq \|Ax\| + \|Bx\| \leq \|A\|\,\|x\| + \|B\|\,\|x\|$$

shows that $\|A+B\| \le \|A\| + \|B\|$. Hence $\|A\|$ also satisfies norm axiom N3. □

As in any normed space, the notion of convergence acts in $L(X,Y)$. A sequence of continuous linear operators $\{A_n\}$ is said to be convergent to A if $\|A_n - A\| \to 0$ as $n \to \infty$; in such a case we say that $\{A_n\}$ *converges uniformly* to A.

Theorem 2.1.1. If X is a normed space and Y is a Banach space, then $L(X,Y)$ is a Banach space.

Proof. Let $\{A_n\}$ be a Cauchy sequence in $L(X,Y)$, i.e.,

$$\|A_{n+m} - A_n\| \to 0 \text{ as } n \to \infty, \quad m > 0.$$

We must show that there is a continuous operator A such that $A = \lim_{n\to\infty} A_n$. For any $x \in X$, $\{A_n x\}$ is also a Cauchy sequence because

$$\|A_{n+m}x - A_n x\| \le \|A_{n+m} - A_n\|\,\|x\|;$$

hence there exists $y \in Y$ such that $y = \lim_{n\to\infty} A_n x$ since Y is a Banach space. Thus, for every $x \in X$ this defines a unique $y \in Y$, i.e., defines an operator A such that $y = Ax$. By properties of the limit and the linearity of A_n, the operator A is linear. Since $\{A_n\}$ is a Cauchy sequence, the sequence of norms $\{\|A_n\|\}$ is bounded, say $\|A_n\| \le c$, and so

$$\|Ax\| = \lim_{n\to\infty} \|A_n x\| \le \limsup_{n\to\infty} \|A_n\|\,\|x\| \le c\|x\|.$$

That is, A is continuous and the proof is finished. □

In a Banach space $L(X,Y)$ we can introduce series

$$\sum_{n=1}^{\infty} A_n$$

and define their sums by

$$S = \lim_{k\to\infty} \sum_{n=1}^{k} A_n.$$

A series $\sum_{n=1}^{\infty} A_n$ is said to be *absolutely convergent* if the numerical series $\sum_{n=1}^{\infty} \|A_n\|$ is convergent.

Problem 2.1.1. Suppose $A_n \in L(X,Y)$ where Y is a Banach space. Show that if $\sum_{n=1}^{\infty} A_n$ is absolutely convergent, then it is convergent.

We denote by $L(X)$ the space $L(X, X)$, X being a Banach space. In $L(X)$ we can introduce the product of operators A, B by

$$ABx = A(Bx).$$

The product possesses the usual properties of a numerical product except commutativity; we have

$$\begin{aligned}(AB)C &= A(BC),\\ (A+B)C &= AC + BC,\\ A(B+C) &= AB + AC,\\ IA &= AI = A,\end{aligned}$$

where I is the identity operator. The product is also a continuous operator, because

$$\|ABx\| \le \|A\|\,\|Bx\| \le \|A\|\,\|B\|\,\|x\|$$

gives

$$\|AB\| \le \|A\|\,\|B\|.$$

Problem 2.1.2. Show that if $A = \lim_{n\to\infty} A_n$ and $B = \lim_{n\to\infty} B_n$ where A_n, B_n belong to $L(X)$, then $AB = \lim_{n\to\infty} A_n B_n$.

So $L(X)$ is a non-commutative ring. Denoting by A^k the product

$$\underbrace{A \cdots A}_{k},$$

we can introduce some functions of operators in $L(X)$, for example,

$$e^A = I + \sum_{k=1}^{\infty} \frac{1}{k!} A^k.$$

Now we introduce another type of convergence of linear operators. Let $g_1, g_2, g_3, \ldots$ be an orthonormal basis of a Hilbert space H. By the Fourier representation of an element

$$x = \sum_{k=1}^{\infty} c_k g_k, \qquad c_k = (x, g_k),$$

we can define an operator P_n called the *operator of orthogonal projection* onto a subspace H_n of H spanned by $g_n, \ldots, g_n$:

$$P_n x = \sum_{k=1}^{n} (x, g_k) g_k.$$

By Bessel's inequality

$$\|P_n x\| \le \|x\|,$$

hence $\|P_n\| \leq 1$ and since $\|P_n g_1\| = \|g_1\|$ it follows that $\|P_n\| = 1$. By definition,

$$P_{n+m}g_{n+1} - P_n g_{n+1} = g_{n+1}, \qquad m > 0,$$

so

$$\|(P_{n+m} - P_n)g_{n+1}\| = \|g_{n+1}\|$$

and thus $\|P_{n+m} - P_n\| \geq 1$ for $m > 0$. This means that the sequence $\{P_n\}$ is not uniformly convergent; however, for any $x \in H$ we get

$$x = \lim_{n\to\infty} P_n x.$$

This forces us to introduce

Definition 2.1.1. A sequence $\{A_n\}$, $A_n \in L(X, Y)$, is said to be *strongly convergent* to $A \in L(X, Y)$ if, whenever $x \in X$,

$$\|A_n x - Ax\| \to 0 \text{ as } n \to \infty.$$

If $\{A_n\}$ is uniformly convergent to A, then it is strongly convergent to A; indeed,

$$\|A_n x - Ax\| \leq \|A_n - A\|\,\|x\| \to 0 \text{ as } n \to \infty.$$

As we have seen, strong convergence of a sequence of operators does not imply its uniform convergence. Strong convergence is sometimes referred to as pointwise convergence.

2.2 Banach–Steinhaus Principle

Let A be a linear operator whose domain is dense in a normed space X. The operator acts from $D(A)$ to a Banach space Y and is bounded on $D(A)$, i.e.,

$$\|Ax\| \leq M\|x\| \quad \text{for all } x \in D(A)$$

(such a situation was met implicitly in proving the imbedding theorems). The infimum of the constants M we can also call $\|A\|$, the norm of A. First we prove

Theorem 2.2.1. Under the above conditions there is a continuation of A, denoted A_c, such that

(i) $A_c \in L(X, Y)$;

(ii) $A_c x = Ax$ for every $x \in D(A)$;

(iii) $\|A_c\| = \|A\|$.

Proof. If $x \in D(A)$ then $A_c x = Ax$. Take $x_0 \notin D(A)$. We construct the value Ax_0 as follows. There is a sequence $\{x_n\} \subset D(A)$ such that $\|x_n - x_0\| \to 0$. The sequence $\{Ax_n\}$ is a Cauchy sequence because

$$\|Ax_n - Ax_m\| \le \|A\| \, \|x_n - x_m\| \to 0 \text{ as } n, m \to \infty;$$

hence, thanks to the completeness of Y, there exists $\lim_{n\to\infty} Ax_n$ which does not depend on the choice of sequence $\{x_n\}$ (verify). We define $A_c x_0$ as $\lim_{n\to\infty} Ax_n$. Since

$$\|Ax_n\| \le \|A\| \, \|x_n\|,$$

passage to the limit gives

$$\|A_c x_0\| \le \|A\| \, \|x_0\|.$$

This means that A_c is continuous and $\|A_c\| \le \|A\|$. But on $D(A)$ we have $\|A_c\| = \|A\|$, so (iii) is valid. □

We now prove the *Banach–Steinhaus principle*, which is given by

Theorem 2.2.2. Let $\{A_n\}$ be a sequence of continuous linear operators in $L(X, Y)$, Y being a Banach space, such that

(i) $\|A_n\| \le M$ for all n;

(ii) there exists $\lim_{n\to\infty} A_n x$ for all $x \in X^*$, X^* being a subspace of X such that X^* is dense in X.

Then the sequence $\{A_n\}$ converges strongly to a continuous linear operator A, i.e., for every $x \in X$

$$\|A_n x - Ax\| \to 0 \text{ as } n \to \infty.$$

Proof. The linear operator A, defined on X^* by the relation

$$Ax = \lim_{n\to\infty} A_n x, \qquad x \in X^*$$

is bounded on X^* by (i); indeed,

$$\|Ax\| = \lim_{n\to\infty} \|A_n x\| \le M\|x\|, \qquad x \in X^*.$$

Using the construction of Theorem 2.2.1 we can extend this operator to X with preservation of norm. Denoting this continuation again by A, we shall show that $\lim_{n\to\infty} A_n x_0 = Ax_0$ for any $x_0 \in X$. Let $\{x_n\} \subset X^*$ be such that

$$\|x_n - x_0\| \to 0 \text{ as } n \to \infty.$$

Then $Ax_0 = \lim_{n\to\infty} Ax_n$. On the other hand,

$$\begin{aligned} \|Ax_0 - A_n x_0\| &= \|Ax_0 - Ax_m + Ax_m - A_n x_m + A_n x_m - A_n x_0\| \\ &\le \|Ax_0 - Ax_m\| + \|Ax_m - A_n x_m\| + \|A_n x_m - A_n x_0\| \\ &\le \|A\| \, \|x_0 - x_m\| + \|Ax_m - A_n x_m\| + M\|x_m - x_0\|. \end{aligned}$$

Given $\varepsilon > 0$, we can choose m_1 such that

$$\|x_0 - x_{m_1}\| < \varepsilon/(3M);$$

fixing this m_1, thanks to (ii), we can find n_1 such that for $n > n_1$ we have

$$\|Ax_{m_1} - A_n x_{m_1}\| < \varepsilon/3.$$

Hence we get

$$\|Ax_0 - A_n x_0\| < \varepsilon \quad \text{when } n > n_1,$$

and this completes the proof. □

The next theorem is the *principle of uniform boundedness*:

Theorem 2.2.3. Assume $A_n \in L(X, Y)$. If $\{A_n x\}$ is bounded for every $x \in X$, then the set $\{\|A_n\|\}$ is also bounded.

The proof can be found in any textbook on functional analysis (see, for example, Yosida [29]).

2.3 The Inverse Operator

We are now interested in solving an equation

$$Ax = y$$

where A is a linear operator, y is a given element of a normed space Y, and x is an unknown element of a normed space X.

We dealt with some problems of this type, reducing them to a trivial equation $x = y$ in an energy space. Now we consider the general case.

If for any $y \in Y$ there is no more than one solution $x \in X$ of the equation, then the correspondence from Y to X defined by the equation is an operator; this operator is called the *inverse* to A and is denoted A^{-1}. It is clear that $D(A^{-1}) = R(A)$ and $R(A^{-1}) = D(A)$.

As an easy exercise, the reader can prove

Theorem 2.3.1. The operator A^{-1} exists if and only if the equation $Ax = 0$ has the unique solution $x = 0$. The operator A^{-1}, if it exists, is also linear.

We are interested not only in solvability of the equation, but in continuity of dependence of its solution on external data. A simple result of this kind is

Theorem 2.3.2. The operator A^{-1} is bounded on $R(A)$ if and only if there is a positive constant c such that

$$\|Ax\| \geq c\|x\| \text{ for all } x \in D(A). \tag{2.3.1}$$

Proof. Necessity. Let A^{-1} exist and be bounded on $R(A)$. It follows that there is a constant $m > 0$ such that $\|A^{-1}y\| \le m\|y\|$. Denoting $y = Ax$, $c = 1/m$, we get (2.3.1).

Sufficiency. From (2.3.1), it follows that the equation $Ax = 0$ has the unique solution $x = 0$, i.e., A^{-1} exists. Putting $x = A^{-1}y$ in (2.3.1), we get $\|A^{-1}y\| \le (1/c)\|y\|$ for all $y \in R(A)$. This completes the proof. □

An important case is when $A^{-1} \in L(Y, X)$; then we shall say that A is *continuously invertible*.

From Theorem 2.3.2 there follows

Theorem 2.3.3. An operator A is continuously invertible if and only if (i) $R(A) = Y$ and (ii) there is a positive constant c such that (2.3.1) holds.

Let us consider some examples.

We begin with the Fredholm equation with degenerate kernel:

$$u(t) - \lambda \int_a^b \sum_{k=1}^n \varphi_k(t)\psi_k(s)u(s)\,ds = f(t), \tag{2.3.2}$$

where λ is a parameter. Assume that $\varphi_k(t)$, $\psi_k(t)$, and $f(t)$ are of class $C(a, b)$. What can we say about the inverse of the operator A_F given by

$$(A_F u)(t) = u(t) - \lambda \int_a^b \sum_{k=1}^n \varphi_k(t)\psi_k(s)u(s)\,ds,$$

acting in $C(a, b)$? If equation (2.3.2) is solvable, then its solution has the form

$$u(t) = f(t) + \sum_{k=1}^n c_k\varphi_k(t).$$

Putting this into (2.3.2), we get

$$\sum_{k=1}^n c_k\varphi_k(t) + f(t) - \lambda \sum_{k=1}^n \varphi_k(t) \int_a^b \psi_k(s) \left(\sum_{i=1}^n c_i\varphi_i(s) + f(s) \right) ds = f(t).$$

Supposing that the system $\varphi_1(t), \ldots, \varphi_n(t)$ is linearly independent, we obtain the linear algebraic system

$$c_k - \lambda \sum_{i=1}^n c_i \int_a^b \varphi_i(s)\psi_k(s)\,ds = \lambda \int_a^b f(s)\psi_k(s)\,ds, \qquad k = 1, \ldots, n,$$

whose solution by Cramer's rule is

$$c_k = \frac{D_k(\lambda, f)}{D(\lambda)}, \qquad k = 1, \ldots, n.$$

Hence

$$u(t) = f(t) + \sum_{k=1}^{n} \frac{D_k(\lambda, f)}{D(\lambda)} \varphi_k(t). \tag{2.3.3}$$

This solution is valid if $D(\lambda) \neq 0$; then from (2.3.3) we see that

$$\max_{t\in[a,b]} |u(t)| \leq m(\lambda) \max_{t\in[a,b]} |f(t)|,$$

or, in terms of norms,

$$\|u\| \leq m(\lambda)\|f\|.$$

This means that $A_F^{-1} \in L(C(a,b))$ if $D(\lambda) \neq 0$ and $\|A_F^{-1}\| \leq m(\lambda)$.

Suppose that $D(\lambda) = 0$. Since $D(\lambda)$ is a polynomial in λ of order n, it has no more than n distinct zeros λ_i. For $\lambda = \lambda_i$, there is a nontrivial solution $c_1^{(i)}, \ldots, c_n^{(i)}$ to the system

$$c_k - \lambda_i \sum_{j=1}^{n} c_j \int_a^b \varphi_j(s)\psi_k(s)\, ds = 0, \qquad k = 1, \ldots, n.$$

This means that the equation $A_F u = 0$ has a nonzero solution and thus A_F^{-1} does not exist. These $\{\lambda_i\}$ comprise the spectrum of the integral operator.

Now we consider a simple boundary value problem

$$u''(t) = f(t), \qquad u(0) = u_0, \qquad u(1) = u_1,$$

when $f(t) \in C(0,1)$. Its solution is

$$u(t) = \int_0^t \int_0^s f(s_1)\, ds_1\, ds + u_0 + \left(u_1 - u_0 - \int_0^1 \int_0^s f(s_1)\, ds_1\, ds\right) t. \tag{2.3.4}$$

In terms of operators, we get an operator B whose domain is $C^{(2)}(0,1)$ and range is the space whose elements are pairs consisting of a function $f(t) \in C(0,1)$ and a vector (u_0, u_1). From (2.3.4) it follows that B^{-1} exists and is bounded.

Finally, we formulate

Theorem 2.3.4. Assume that X and Y are Banach spaces, $A \in L(X,Y)$ is continuously invertible, and $B \in L(X,Y)$ is such that $\|B\| < \|A^{-1}\|^{-1}$. Then $A + B$ has an inverse $(A+B)^{-1} \in L(Y,X)$ and

$$\|(A+B)^{-1}\| \leq (\|A^{-1}\|^{-1} - \|B\|)^{-1}. \tag{2.3.5}$$

Proof. The equation

$$(A+B)x = y \tag{2.3.6}$$

can be reduced to

$$x - Cx = x_0, \qquad C = -A^{-1}B, \qquad x_0 = A^{-1}y.$$

By the condition of the theorem, $\|C\| \le \|A^{-1}\| \, \|B\| < 1$. So we can apply the contraction mapping principle to this equation and find that it has a unique solution x^* for any $y \in Y$ which can be found by iteration:

$$x_{k+1} = x_0 + Cx_k$$

or

$$x_k = (I + C + \cdots + C^k)x_0.$$

Existence of the unique solution to (2.3.6) means that the inverse to $A+B$ exists and its domain is Y.

We now obtain (2.3.5). From $x = A^{-1}Ax$ it follows that

$$\|x\| \le \|A^{-1}\| \, \|Ax\|,$$

and so

$$\|Ax\| \ge \|A^{-1}\|^{-1}\|x\|.$$

For any $y \in Y$, (2.3.6) shows that

$$\begin{aligned} \|y\| &= \|(A+B)x\| \\ &\ge \|Ax\| - \|Bx\| \\ &\ge \|A^{-1}\|^{-1}\|x\| - \|B\| \, \|x\| \\ &= (\|A^{-1}\|^{-1} - \|B\|)\|x\|. \end{aligned}$$

From this (2.3.5) follows, and the proof is complete. □

2.4 Closed Operators

Broader than the class of continuous linear operators is the class of closed linear operators.

Definition 2.4.1. A linear operator A acting from a Banach space X to a Banach space Y is called *closed* if for any sequence $\{x_n\} \subset D(A)$ such that $\lim_{n\to\infty} x_n = x$ and $\lim_{n\to\infty} Ax_n = y$, it follows that $x \in D(A)$ and $y = Ax$.

By definition, a continuous linear operator whose domain is X is closed. There are closed linear operators which are not continuous. We now give an example of such an operator.

The differentiation operator d/dt acting from $C(0,1)$ to $C(0,1)$ is closed. Indeed, let $x_n(t) \to x(t)$ and $x_n'(t) \to y(t)$, both uniformly (i.e., in $C(0,1)$). By a well known theorem of calculus, it follows that $x'(t) = y(t)$, i.e., Definition 2.4.1 is fulfilled for d/dt. Its unboundedness is seen if we consider d/dt acting on the sequence $\{t^n\}$.

In a similar way, it may be shown that a general differential operator A given by $Af(\mathbf{x}) = \sum_{|\alpha|\le n} c_\alpha(\mathbf{x}) D^\alpha f(\mathbf{x})$ with smooth coefficients $c_\alpha(\mathbf{x})$ acting in $C(\Omega)$ is closed.

Now we consider a closed operator from another point of view. We begin with the definition of the product $X \times Y$ of Banach spaces X, Y over the same scalar field. The elements of $X \times Y$ are ordered pairs (x, y), $x \in X$, $y \in Y$. This is a linear space with operations defined as follows:

$$(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1 + y_2), \qquad \alpha(x, y) = (\alpha x, \alpha y).$$

Moreover, $X \times Y$ is a Banach space under the norm

$$\|(x, y)\| = (\|x\|^2 + \|y\|^2)^{1/2}.$$

Definition 2.4.2. The subset $\{(x, Ax) \mid x \in D(A)\}$ of $X \times Y$ is called the *graph* of an operator A acting from $D(A) \subset X$ to Y.

The following is equivalent to Definition 2.4.1 (prove this):

Definition 2.4.3. A linear operator A acting from $D(A) \subset X$ to Y is called closed if its graph is a closed linear subspace of $X \times Y$.

We shall say that a linear operator A acting from $D(A) \subset X$ to Y has a *closed extension* if the closure of the graph $G(A)$ in $X \times Y$ is the graph of a linear operator, say B, acting from $D(B) \subset X$ to Y.

Lemma 2.4.1. An operator A acting from a Banach space X to a Banach space Y has a closed extension if and only if from the condition

(*) let $\{x_n\} \subset D(A)$ be an arbitrary sequence such that $\lim_{n\to\infty} x_n = 0$ and $\lim_{n\to\infty} Ax_n = y$

it follows that $y = 0$.

Proof. Necessity follows from the definition of a closed operator. For sufficiency, we construct directly an operator B, called the *least closed extension* of A, as follows. An element x belongs to $D(B)$ if and only if there is a sequence $\{x_n\} \subset D(A)$ such that $\lim_{n\to\infty} \|x_n - x\| = 0$ and $Ax_n \to y$ (strong limit); for this x, the value Bx is defined by $Bx = y$. By the condition (*), y is uniquely defined by x so B is an operator whose linearity is evident.

Then let $\{u_n\} \subset D(B)$ be a sequence such that $u_n \to u$ and $Bu_n \to v$. To complete the proof, we must show that $Bu = v$. By definition of B, there is a sequence $\{x_n\} \subset D(A)$ such that $\|x_n - u_n\| < \varepsilon_n$ and $\|Ax_n - Bu_n\| < \varepsilon_n$ where $\varepsilon_n \to 0$ as $n \to \infty$. Hence $x_n \to u$ and $Ax_n \to v$ as $n \to \infty$. Thus, by definition of B, we have $u \in D(B)$ and $Bu = v$. □

As an example of the application of this lemma we consider an extension of the operator

$$A = \sum_{|\alpha|\le k} c_\alpha(\mathbf{x}) D^\alpha$$

with coefficients $c_\alpha(\mathbf{x}) \in C^{(k)}(\Omega)$, Ω being compact in $\mathbb{R}^n$, acting in $L^2(\Omega)$. We define the domain of A as the set $L^2(\Omega) \cap C^{(k)}(\Omega)$. The range of A lies in $L^2(\Omega)$. We now try the condition of the lemma to show that A has a closed extension. Let $\{u_n(\mathbf{x})\} \subset D(A)$ be such that $\|u_n\|_{L^2(\Omega)} \to 0$ and $\|Au_n(\mathbf{x}) - v(\mathbf{x})\|_{L^2(\Omega)} \to 0$ as $n \to \infty$. Let $\varphi(\mathbf{x}) \in C_0^{(k)}(\Omega)$, where $C_0^{(k)}(\Omega)$ is the subspace of $C^{(k)}(\Omega)$ consisting of functions that together with all their derivatives up to the order k are zero on the boundary $\partial\Omega$. Integration by parts gives

$$\int_\Omega \varphi(\mathbf{x}) A u_n(\mathbf{x})\, d\Omega = \int_\Omega u_n(\mathbf{x}) \sum_{|\alpha| \le k} (-1)^{|\alpha|} D^\alpha [c_\alpha(\mathbf{x})\varphi(\mathbf{x})]\, d\Omega$$

(the boundary terms vanish because $\varphi(\mathbf{x})$ and $D^\alpha\varphi(\mathbf{x})$ equal zero on $\partial\Omega$). By definition of $L^2(\Omega)$, passage to the limit gives

$$\int_\Omega \varphi(\mathbf{x}) v(\mathbf{x})\, d\Omega = \int_\Omega 0 \cdot \sum_{|\alpha| \le k} (-1)^{|\alpha|} D^\alpha [c_\alpha(\mathbf{x})\varphi(\mathbf{x})]\, d\Omega = 0.$$

Since $C_0^k(\Omega)$ is dense in $L^2(\Omega)$, it follows that $v(\mathbf{x}) = 0$ as an element of $L^2(\Omega)$. Hence there is a closed extension of the differential operator A. (This is another approach which can lead us to generalized derivatives, and is equivalent to the approach due to S.L. Sobolev.)

Theorem 2.4.1. If A is a closed linear operator and its inverse A^{-1} exists, then A^{-1} is also closed.

Proof. The graph of A^{-1} can be obtained from the graph of A by rearrangement: $(x, Ax) \mapsto (Ax, x)$. This means that $G(A^{-1})$ is a closed set in $Y \times X$. □

A useful result is

Theorem 2.4.2. Let A be a closed linear operator whose domain is a Banach space X and whose range lies in a Banach space Y. Assume there is a set M which is dense in X and a positive constant c such that

$$\|Ax\| \le c\|x\| \text{ for all } x \in M.$$

Then the operator A is continuous.

Proof. Take $x_0 \in X$. Since M is dense in X, there is a sequence $\{x_n\} \subset M$ such that

$$\|x_n - x_0\| < 1/n.$$

By the condition of the theorem we get

$$\begin{aligned} \|Ax_{k+m} - Ax_k\| &\le c\|x_{k+m} - x_k\| \\ &\le c(\|x_{k+m} - x_0\| + \|x_k - x_0\|) \\ &\le 2c/k \to 0 \text{ as } k \to \infty, \end{aligned}$$

so $\{Ax_k\}$ is a Cauchy sequence which, thanks to completeness of Y, has limit y. By closedness of A we get $Ax_0 = y$. On the other hand

$$\|Ax_0\| = \lim_{k\to\infty} \|Ax_k\| \le \lim_{k\to\infty} c\|x_k\| = c\|x_0\|,$$

which completes the proof. □

We can now formulate Banach's *closed graph theorem*.

Theorem 2.4.3. Let A be a closed linear operator acting from a Banach space X to a Banach space Y, and let $D(A) = X$. Then A is continuous on X.

The proof can be found in any textbook on functional analysis (e.g., Yosida [29]). A consequence of this theorem is

Theorem 2.4.4. If A is a closed linear operator mapping a Banach space X onto a Banach space Y (i.e., $R(A) = Y$) one-to-one, then A^{-1} is continuous on Y.

Proof. By Theorem 2.4.1, A^{-1} is closed; by Theorem 2.4.3, it is continuous. □

Problem 2.4.1. Let a linear space X be a Banach space with respect to each of two norms $\|x\|_1$ and $\|x\|_2$. Show that if for every $x \in X$ there is a positive constant c_1 not depending on x such that $\|x\|_1 \le c_1\|x\|_2$, then there is a positive constant c_2 such that $\|x\|_2 \le c_2\|x\|_1$. That is, show that the norms $\|x\|_1$ and $\|x\|_2$ are equivalent.

2.5 The Notion of Adjoint Operator

This notion will be introduced for operators acting in a Hilbert space, although it can also be applied in other settings. So we let H be a Hilbert space and A be a continuous linear operator acting from H to H.

Consider the inner product (Ax, y) as a functional with respect to the variable $x \in H$ when $y \in H$ is arbitrary but fixed. This functional, thanks to the linearity of A, is linear and bounded because

$$|(Ax, y)| \le \|Ax\|\,\|y\| \le (\|A\|\,\|y\|)\|x\|.$$

By the Riesz representation theorem, it can be represented in the form

$$(Ax, y) = (x, z)$$

where the element z is uniquely defined by y and A. So the correspondence $y \mapsto z$ can be viewed as an operator A^*, $z = A^*y$, and we call A^* the *adjoint* of A.

Let us consider some properties of A^*.

Lemma 2.5.1. A^* is a linear operator.

Proof. By definition we get

$$(Ax, y_1) = (x, A^*y_1), \qquad (Ax, y_2) = (x, A^*y_2),$$

and

$$(Ax, \alpha_1 y_1 + \alpha_2 y_2) = (x, A^*(\alpha_1 y_1 + \alpha_2 y_2)).$$

But

$$(Ax, \alpha_1 y_1 + \alpha_2 y_2) = \overline{\alpha_1}(Ax, y_1) + \overline{\alpha_2}(Ax, y_2),$$

so

$$\begin{aligned}(x, A^*(\alpha_1 y_1 + \alpha_2 y_2)) &= \overline{\alpha_1}(x, A^*y_1) + \overline{\alpha_2}(x, A^*y_2) \\ &= (x, \alpha_1 A^*y_1) + (x, \alpha_2 A^*y_2).\end{aligned}$$

Since x is an arbitrary element of H, we have

$$A^*(\alpha_1 y_1 + \alpha_2 y_2) = \alpha_1 A^*y_1 + \alpha_2 A^*y_2.$$

This completes the proof. □

Lemma 2.5.2. We have

(i) $(A + B)^* = A^* + B^*$,

(ii) $(AB)^* = B^*A^*$.

Proof. Property (i) is evident. Comparing the equalities

$$((AB)x, y) = (x, (AB)^*y)$$

and

$$(A(Bx), y) = (Bx, A^*y) = (x, B^*(A^*y)),$$

we prove (ii). □

Lemma 2.5.3. If A is a continuous linear operator, then A^* is continuous; moreover, $\|A^*\| = \|A\|$.

Proof. Using the Schwarz inequality, we get

$$M = \sup_{x,y \in H} \frac{|(Ax, y)|}{\|x\| \, \|y\|} \leq \sup_{x,y \in H} \frac{\|A\| \, \|x\| \, \|y\|}{\|x\| \, \|y\|} = \|A\|.$$

By definition of A^*, $(Ax, y) = (x, A^*y)$ and

$$M = \sup_{x,y \in H} \frac{|(x, A^*y)|}{\|x\| \, \|y\|}.$$

Given $x = A^*y$, we have

$$M_1 = \sup_{y \in H} \frac{|(A^*y, A^*y)|}{\|A^*y\| \, \|y\|} = \sup_{y \in H} \frac{\|A^*y\|}{\|y\|}.$$

But $M_1 \leq M$, so A^* is bounded and

$$M_1 = \|A^*\| \leq M \leq \|A\|.$$

Thus we obtain that A^* is continuous and $\|A^*\| \leq \|A\|$. The reverse inequality $\|A\| \leq \|A^*\|$, which completes the proof, follows from the next lemma. □

Lemma 2.5.4. $(A^*)^* = A$.

Proof. Since A^* is continuous then, by definition,

$$(A^*x, y) = (x, (A^*)^*y).$$

On the other hand

$$(A^*x, y) = \overline{(y, A^*x)} = \overline{(Ay, x)} = (x, Ay),$$

so we get

$$(x, (A^*)^*y) = (x, Ay).$$

Since x and y are arbitrary elements of H, we conclude that $(A^*)^* = A$. □

We have introduced the adjoint operator for a continuous linear operator in a Hilbert space. If A is unbounded then we can try to introduce A^* by the same equality

$$(Ax, y) = (x, A^*y), \qquad x \in D(A)$$

which defines a value of A^*y uniquely if $D(A)$ is dense in H. What are the properties of A^* now? The reader should study this problem. The notion of adjoint can also be introduced for operators acting in Banach spaces (see, for example, Yosida [29]). In what follows we will not use the notion of the adjoint to an unbounded operator.

We consider some examples.

1. *A matrix operator in ℓ^2.* This was considered in Section 1.13:

$$(Ax)_i = \sum_{j=1}^{\infty} a_{ij} x_j.$$

By (1.13.9), its norm in ℓ^2 is bounded by

$$\|A\| \leq \left(\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} |a_{ij}|^2 \right)^{1/2}.$$

The adjoint of A is defined as follows:

$$(Ax, y) = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{ij}x_j\overline{y_i} = \sum_{j=1}^{\infty} x_j \overline{\left(\sum_{i=1}^{\infty} \overline{a_{ij}}y_i\right)} = (x, A^*y),$$

so

$$(A^*y)_j = \sum_{i=1}^{\infty} \overline{a_{ij}}y_i.$$

Here the subscript j denotes the jth component of the vector A^*y. Sometimes it happens that $a_{ij} = \overline{a_{ji}}$; in such cases $A = A^*$, and A is called *self-adjoint*.

Definition 2.5.1. A linear operator A is called *self-adjoint* if $A^* = A$.

2. *An integral operator.* We consider an integral operator of the form

$$(Bf)(x) = \int_0^1 K(x, y)f(y)\,dy$$

in $L^2(0, 1)$. If we suppose that $K(x, y) \in L^2([0, 1]\times[0, 1])$ then it is bounded in $L^2(0, 1)$. Indeed

$$\|Bf\|_{L^2(0,1)} = \left(\int_0^1 \left|\int_0^1 K(x, y)f(y)\,dy\right|^2 dx\right)^{1/2}.$$

Using the Schwarz inequality, we get

$$\begin{aligned}\|Bf\|_{L^2(0,1)} &\le \left(\int_0^1 \left(\int_0^1 |K(x, y)|^2\,dy \int_0^1 |f(y)|^2\,dy\right) dx\right)^{1/2}\\ &= \left(\int_0^1\int_0^1 |K(x, y)|^2\,dy\,dx\right)^{1/2} \|f\|_{L^2(0,1)}.\end{aligned}$$

Thus

$$\|B\| \le \left(\int_0^1\int_0^1 |K(x, y)|^2\,dy\,dx\right)^{1/2}.$$

Let us introduce B^* when $K(x, y) \in L^2([0, 1]\times[0, 1])$. For this we consider

$$\begin{aligned}(Bf, g) &= \int_0^1\int_0^1 K(x, y)f(y)\,dy\,\overline{g(x)}\,dx\\ &= \int_0^1 f(y)\overline{\int_0^1 \overline{K(x, y)}g(x)\,dx}\,dy\\ &= (f, B^*g),\end{aligned}$$

so

$$(B^*g)(y) = \int_0^1 \overline{K(x,y)}g(x)\,dx.$$

Therefore B^* is also an integral operator. If $K(x,y) = \overline{K(y,x)}$, then $B = B^*$ and so B is self-adjoint.

3. *Stability of a thin plate.* The linearized integro-differential equation in the theory of stability of a plate under tension can be written in the form

$$(w,\varphi)_{E_P} - \mu C(w,\varphi) = 0, \tag{2.5.1}$$

$$C(w,\varphi) = \int_\Omega \left[T_x \frac{\partial w}{\partial x}\frac{\partial \varphi}{\partial x} + T_{xy}\left(\frac{\partial w}{\partial y}\frac{\partial \varphi}{\partial x} + \frac{\partial w}{\partial x}\frac{\partial \varphi}{\partial y}\right) + T_y \frac{\partial w}{\partial y}\frac{\partial \varphi}{\partial y}\right] dx\,dy$$

where $(w,\varphi)_{E_P}$ is introduced by (1.10.14). For definiteness, we consider the problem in E_{PC}; T_x, T_{xy}, and T_y here are considered as given functions from $L^2(\Omega)$. Condition (2.15.3) that the plate is under tension provides that all the eigenvalues of the problem are positive. The results of this section do not depend on (2.15.3).

The problem is to find the minimal μ such that there is a nontrivial function $w \in E_{PC}$ that satisfies (2.5.1) for every $\varphi \in E_{PC}$. So this is the problem of finding the least eigenvalue.

Let us transform this problem into operator form. For this, consider $C(w,\varphi)$. Applying the Hölder inequality to a term

$$R(w,\varphi) = \int_\Omega T_x \frac{\partial w}{\partial x}\frac{\partial \varphi}{\partial x}\,dx\,dy,$$

we get

$$|R(w,\varphi)| \le \left(\int_\Omega T_x^2\,dx\,dy\right)^{1/2}\left(\int_\Omega \left(\frac{\partial w}{\partial x}\right)^4 dx\,dy\right)^{1/4} \cdot \left(\int_\Omega \left(\frac{\partial \varphi}{\partial x}\right)^4 dx\,dy\right)^{1/4}.$$

Remembering the imbedding theorem in E_{PC}, we then obtain

$$|R(w,\varphi)| \le m\|w\|_{E_P}\|\varphi\|_{E_P}.$$

In a similar way, we can bound other terms in $C(w,\varphi)$ and so

$$|C(w,\varphi)| \le m_1\|w\|_{E_P}\|\varphi\|_{E_P}. \tag{2.5.2}$$

The linearity of $C(w,\varphi)$ with respect to both variables w and φ is evident.

Let $w \in E_{PC}$ be fixed. Then, thanks to (2.5.2), $C(w,\varphi)$ is a continuous linear functional with respect to $\varphi \in E_{PC}$ and thus, by the Riesz representation theorem, can be represented in the form

$$C(w,\varphi) = (\varphi, v)_{E_P}.$$

Since for every $w \in E_{PC}$ there is a unique element $v \in E_{PC}$ we have obtained a correspondence $w \mapsto v$ that is an operator G, $v = Gw$, whose linearity is evident. By (2.5.2) we have

$$|C(w, \varphi)| = |(\varphi, Gw)_{E_P}| \leq m_1 \|w\|_{E_P} \|\varphi\|_{E_P}.$$

Putting $\varphi = Gw$, we get

$$(Gw, Gw)_{E_P} \leq m_1 \|w\|_{E_P} \|Gw\|_{E_P}$$

or

$$\|Gw\|_{E_P} \leq m_1 \|w\|_{E_P}.$$

So the operator G is continuous.

Since the form $C(w, \varphi)$ is symmetrical in the variables w and φ, we get

$$(\varphi, Gw)_{E_P} = C(w, \varphi) = C(\varphi, w) = (w, G\varphi) \text{ for all } w, \varphi \in E_{PC}.$$

This means that the operator G is self-adjoint.

4. *The operator of differentiation.* Let us consider an example of an unbounded operator. This is the operator $D_t = i\, d/dt$ acting in $L^2(0,1)$ whose domain consists of functions $\dot{W}^{1,2}(0,1)$, i.e., functions $f(x)$ which, along with their derivatives, belong to $L^2(0,1)$, and which satisfy $f(0) = f(1) = 0$.

As above, we shall find $(D_t)^*$:

$$(D_t f, g) = \int_0^1 i \frac{df(t)}{dt} \overline{g(t)}\, dt = \int_0^1 f(t) \overline{i \frac{dg(t)}{dt}}\, dt = (f, D_t^* g).$$

This formula is valid if $g \in C^{(1)}(0,1)$. Passage to the limit here shows that it remains valid if $g \in W^{1,2}(0,1)$.

Thus $D_t^* = i\, d/dt$; i.e., D_t^* has the same form as D_t but its domain includes $W^{1,2}(0,1)$. So $D(D_t^*)$ is wider than $D(D_t)$, and thus $D_t^* \neq D_t$. In this case, the operator is called symmetrical (not self-adjoint).

We now obtain a pair of simple but useful lemmas.

Lemma 2.5.5. If a linear operator A is strongly continuous on a Hilbert space then it is also weakly continuous; that is, it takes any weakly convergent sequence into a weakly convergent sequence.

Proof. Let $x_n \rightharpoonup x_0$ in H. An arbitrary continuous linear functional $F(x)$ takes the form $F(x) = (x, f)$, $f \in H$, and hence we must show that

$$(Ax_n - Ax_0, f) \to 0 \text{ as } n \to \infty.$$

But $(Ax, f) = (x, A^* f)$, and so

$$(Ax_n - Ax_0, f) = (x_n - x_0, A^* f) \to 0 \text{ as } n \to \infty$$

since $A^* f \in H$ and $\{x_n\}$ converges weakly to x_0. □

In fact this is a more general result which we formulate as

Lemma 2.5.6. A continuous linear operator A acting from a normed space X into a normed space Y is also weakly continuous.

The proof follows from the fact that for any continuous linear functional $F(y)$ given on Y the functional $\Phi(x) = F(Ax)$ is also continuous and linear, but on X.

Lemma 2.5.7. Assume that A is a continuous linear operator acting in a Hilbert space H. Let $x_n \rightharpoonup x_0$ and $y_n \to y_0$ in H. Then

$$(Ax_n, y_n) \to (Ax_0, y_0) \text{ as } n \to \infty.$$

Proof. Because $(Ax, y) = (x, A^*y)$, where A^* is continuous, we get

$$R_n = (Ax_n, y_n) - (Ax_0, y_0) = (x_n, A^*y_n) - (x_0, A^*y_0).$$

Transforming this, we have

$$\begin{aligned} R_n &= (x_n, A^*y_n) - (x_0, A^*y_0) + (x_n, A^*y_0) - (x_n, A^*y_0) \\ &= (x_n, A^*(y_n - y_0)) + (x_n - x_0, A^*y_0); \end{aligned}$$

but $(x_n - x_0, A^*y_0) \to 0$ because $x_n \rightharpoonup x_0$ in H, and

$$|(x_n, A^*(y_n - y_0))| \le \|x_n\| \, \|A^*\| \, \|y_n - y_0\| \to 0$$

since $\{x_n\}$ is bounded as a weakly convergent sequence. □

Finally, we propose a simple but important

Problem 2.5.1. Let an operator K be defined by the Riesz representation theorem from the equality

$$(Ku, \varphi)_E = \int_\Omega \rho(\mathbf{x}) u(\mathbf{x}) \varphi(\mathbf{x}) \, d\Omega$$

in an energy space E when $\rho(\mathbf{x})$ is a bounded piecewise continuous function (density) on a compact set Ω. Show that K is a self-adjoint continuous linear operator in all of the energy spaces E_M, E_P, and E_E introduced earlier. (For the body with free boundary, only the spaces of balanced functions should be considered!)

For a self-adjoint operator, the norm can be defined in another way.

Theorem 2.5.1. If A is a self-adjoint continuous linear operator given on a Hilbert space H, then

$$\|A\| = \sup_{\|x\| \le 1} |(Ax, x)|. \tag{2.5.3}$$

Proof. We denote $\sup_{\|x\|\le 1} |(Ax, x)| = \gamma$. Using the Schwarz inequality, we get

$$\gamma \le \sup_{\|x\|\le 1} \{\|Ax\|\,\|x\|\} \le \sup_{\|x\|\le 1} \{\|A\|\,\|x\|^2\} = \|A\|.$$

We now show the reverse inequality. By definition of γ we have

$$|(Ax, x)| \le \gamma \|x\|^2. \tag{2.5.4}$$

Setting $x_1 = y + \lambda z$ and $x_2 = y - \lambda z$, λ being a real number and $y, z \in H$, we have

$$\begin{aligned} C &\equiv |(Ax_1, x_1) - (Ax_2, x_2)| \\ &= |2\lambda|\,|(Ay, z) + (Az, y)| \\ &= |2\lambda|\,|(Ay, z) + (z, Ay)|. \end{aligned}$$

On the other hand

$$\begin{aligned} C &\le |(Ax_1, x_1)| + |(Ax_2, x_2)| \\ &\le \gamma(\|x_1\|^2 + \|x_2\|^2) \\ &= 2\gamma(\|y\|^2 + \lambda^2\|z\|^2) \end{aligned}$$

so

$$|2\lambda|\,|(Ay, z) + (z, Ay)| \le 2\gamma(\|y\|^2 + \lambda^2\|z\|^2) \text{ for all } y, z \in H.$$

Putting $z = Ay$, we obtain

$$|4\lambda|\,\|Ay\|^2 \le 2\gamma(\|y\|^2 + \lambda^2\|Ay\|^2).$$

Take $\lambda = \|y\|/\|Ay\|$; then

$$4\|y\|\,\|Ay\| \le 2\gamma(\|y\|^2 + \|y\|^2)$$

or

$$\|Ay\| \le \gamma\|y\| \quad \text{for all } y \in H.$$

So $\|A\| \le \gamma$, and this completes the proof. □

2.6 Compact Operators

The study of linear operators in infinite dimensional spaces is complicated in comparison with the theory in finite dimensional spaces. However, some classes of these operators can be fully described; the first to note this was D. Hilbert. Among these, the class of compact operators is one of the most important: on the one hand, compact operators are close to finite dimensional operators in terms of properties and, on the other hand, they play important roles in applications.

In this section we are dealing with a linear operator A acting from a normed space X to a Banach space Y.

Definition 2.6.1. A linear operator A is *compact* if it takes bounded sets of X into precompact subsets of Y.

Alternatively, compact linear operators are sometimes called *completely continuous operators*.

Theorem 2.6.1. A compact linear operator A is bounded.

Proof. Suppose to the contrary that it is not bounded. This means there is a bounded sequence $\{x_n\} \subset X$ such that $\|Ax_n\| \to \infty$ as $n \to \infty$. But the sequence $\{Ax_n\}$ does not contain a convergent subsequence, and this contradicts the definition of a compact operator. □

So a compact operator is continuous. The converse is, in general, false. For example, the identity operator $Ix = x$ is continuous but not compact if X is not a finite dimensional space, since a ball is compact only in a finite dimensional space.

We consider an example of a compact operator. We show that the operator

$$(Bf)(t) = \int_0^1 K(t,s)f(s)\,ds$$

acting in $C(0,1)$ is compact when $K(t,s) \in C([0,1] \times [0,1])$. It suffices to show that B takes the unit ball of $C(0,1)$ into a precompact subset S of $C(0,1)$. By boundedness of B, the set S is uniformly bounded. By Arzelà's theorem on compactness in $C(0,1)$, it remains to establish that S is equicontinuous; this follows from the inequality chain

$$\begin{aligned} |(Bf)(t+\delta) - (Bf)(t)| &= \left| \int_0^1 K(t+\delta,s)f(s)\,ds - \int_0^1 K(t,s)f(s)\,ds \right| \\ &\le \max_{s\in[0,1]} |K(t+\delta,s) - K(t,s)| \le \varepsilon \end{aligned}$$

since $K(t,s)$ is uniformly continuous on $[0,1+\delta_0] \times [0,1]$, $\delta_0 > 0$. (It is assumed that $K(t,s)$ is extended by continuity outside of $[0,1] \times [0,1]$.)

Problem 2.6.1. Show that if $K(t,s) \in C([0,1] \times [0,1])$, then the integral operator B is also a compact operator in $L^2(0,1)$.

We shall see below that this restriction on $K(t,s)$ can be weakened.

The set of all compact linear operators is closed in $L(X,Y)$; this follows from

Theorem 2.6.2. If a sequence $\{A_n\} \subset L(X,Y)$ of compact linear operators converges uniformly to A (in the norm of $L(X,Y)$), then A is compact.

Proof. Let $\|A_n - A\| \to 0$ as $n \to \infty$. Take a bounded sequence $\{x_n\} \subset X$. We need only show that there is a subsequence $\{x_{n_k}\}$ such that $\{Ax_{n_k}\}$ is a Cauchy sequence in Y. For this, from $\{x_n\}$, thanks to the compactness of A_k, we first select a subsequence $\{x_{n_1}\}$ such that $\{A_1 x_{n_1}\}$ is a Cauchy

sequence in Y. From $\{x_{n_1}\}$ we select similarly a subsequence $\{x_{n_2}\}$ such that the sequence $\{A_2 x_{n_2}\}$ is also a Cauchy sequence. This can be repeated with $\{x_{n_2}\}$, $\{x_{n_3}\}$, and so on, producing a Cauchy sequence $\{A_k x_{n_k}\}$ each time. The subsequence $\{x_{n_n}\}$ is needed; indeed re-denoting x_{n_n} by z_n, we find that the $\{A_k z_n\}$ are Cauchy sequences for every fixed $k = 1, 2, 3, \ldots$ and

$$\begin{aligned}\|Az_{n+m} - Az_n\| &= \|(Az_{n+m} - A_k z_{n+m}) + (A_k z_{n+m} - A_k z_n) + \\ &\quad + (A_k z_n - Az_n)\| \\ &\le \|A - A_k\|\,\|z_{n+m}\| + \|A_k(z_{n+m} - z_n)\| + \\ &\quad + \|A_k - A\|\,\|z_n\| \\ &\to 0 \text{ as } k, n \to \infty.\end{aligned}$$

This completes the proof. □

Now we can establish the promised result for an integral operator. We assume that $K(t,s) \in L^2([0,1] \times [0,1])$ and show that the operator is a compact operator in $C(0,1)$ and in $L^2(0,1)$. We begin with a note that there is a sequence of functions $\{K_n(t,s)\} \subset C([0,1] \times [0,1])$ such that

$$\int_0^1 \int_0^1 |K(t,s) - K_n(t,s)|^2 \, ds\, dt \to 0 \text{ as } n \to \infty$$

(this is by definition of $L^2(\Omega)$). Each of these kernels defines an operator A_n,

$$A_n f(t) = \int_0^1 K_n(t,s) f(s)\, ds,$$

which is a compact operator in $L^2(0,1)$. From the inequality

$$\begin{aligned}\|Af\|_{L^2(0,1)} &= \left(\int_0^1 \left| \int_0^1 K(t,s) f(s)\, ds \right|^2 dt \right)^{1/2} \\ &\le \left(\int_0^1 \int_0^1 |K(t,s)|^2 \, ds\, dt \right)^{1/2} \left(\int_0^1 |f(s)|^2 \, ds \right)^{1/2}\end{aligned}$$

it follows that

$$\|A\| \le \left(\int_0^1 \int_0^1 |K(t,s)|^2 \, ds\, dt \right)^{1/2} \quad \text{in } L^2(0,1)$$

and so $\|A - A_n\| \to 0$ as $n \to \infty$. By Theorem 2.6.2 this means that A is a compact linear operator in $L^2(0,1)$.

Theorem 2.6.3. A compact operator $A \in L(X,Y)$ takes a weakly Cauchy sequence $\{x_n\} \subset X$ into a strongly Cauchy sequence $\{Ax_n\} \subset Y$.

Proof. A weakly Cauchy sequence $\{x_n\}$ is bounded in X so, by definition of complete continuity of A, we can take a strongly Cauchy subsequence $\{Ax_{n_1}\}$ which converges to an element y of Y. By Lemma 2.5.6, $\{Ax_n\}$ is a weak Cauchy sequence in Y; because its subsequence converges strongly to y, the whole sequence $\{Ax_n\}$ converges weakly to $y \in Y$.

We now show that the whole sequence $\{Ax_n\}$ converges strongly to y. Suppose to the contrary that there is a subsequence $\{Ax_{n_2}\}$ which does not converge to y, i.e., there exists $\varepsilon > 0$ such that

$$\|Ax_{n_2} - y\| > \varepsilon. \tag{2.6.1}$$

But from the sequence $\{Ax_{n_2}\}$ we can select a subsequence $\{Ax_{n_3}\}$ which is strongly Cauchy in Y and thus has a limit point $y_1 \in Y$. This subsequence converges weakly to the same element y_1. By the above, it converges weakly to y, too. By uniqueness of the weak limit, we get $y_1 = y$ and this contradicts (2.6.1). □

In Section 1.11 we formulated the imbedding theorems in Sobolev spaces (Theorems 1.11.1–4). Now we can give a clear meaning to the term "compact operator": such an operator takes every sequence which is weakly convergent in $W^{k,p}(\Omega)$ into a sequence which is strongly convergent in a corresponding space shown by the condition of a theorem. In particular, in any $W^{k,2}(\Omega)$, $k \geq 1$, the imbedding operator into $L^2(\Omega)$ is compact. Since all energy spaces introduced earlier can be considered as closed subspaces of $W^{k,2}(\Omega)$, $k = 1$ or 2, we can say more about the operator K defined according to the Riesz representation theorem by the equality

$$(Ku, \varphi)_E = \int_\Omega \rho(\mathbf{x}) u(\mathbf{x}) \varphi(\mathbf{x})\, d\Omega$$

in Section 2.5. We combine its properties into Lemma 2.6.1 below.

Remark 2.6.1. In the case of a free boundary of a body, we should consider only the variants of energy spaces where the elements are "balanced" functions to avoid "rigid" motions.

Lemma 2.6.1. If Ω is a closed and bounded domain (in $\mathbb{R}^2$ or $\mathbb{R}^3$) and $\rho(\mathbf{x})$ is a bounded piecewise continuous function, then K is a compact self-adjoint operator in any of the spaces E_M, E_P, or E_E.

Proof. We begin with the inequalities

$$\begin{aligned} |(Ku, \varphi)_E| &= \left| \int_\Omega \rho(\mathbf{x}) u(\mathbf{x}) \varphi(\mathbf{x})\, d\Omega \right| \\ &\leq \sup_{\mathbf{x} \in \Omega} |\rho(\mathbf{x})|\, \|u\|_{L^2(\Omega)} \|\varphi\|_{L^2(\Omega)} \\ &\leq m \|u\|_{L^2(\Omega)} \|\varphi\|_E \end{aligned} \tag{2.6.2}$$

which follow from the Schwarz inequality and the imbedding theorems in an energy space.

Let $\{u_n\}$ be a bounded sequence in an energy space so that it contains a weakly convergent subsequence which we denote by $\{u_n\}$ again. As was said above, the latter is a Cauchy sequence in $L^2(\Omega)$. Setting $u = u_{n+m} - u_n$ and $\varphi = K(u_{n+m} - u_n)$ in (2.6.2) we get

$$(K(u_{n+m} - u_n), K(u_{n+m} - u_n))_E \le m\|u_{n+m} - u_n\|_{L^2(\Omega)}\|K(u_{n+m} - u_n)\|_E$$

or

$$\|Ku_{n+m} - Ku_n\|_E \le m\|u_{n+m} - u_n\|_{L^2(\Omega)} \to 0 \text{ as } n \to \infty,$$

which means that K is compact. Its self-adjointness was shown earlier. □

Definition 2.6.2. An operator A_n acting from X to Y is called *finite dimensional* if it takes the form

$$A_n x = \sum_{k=1}^{n} \Phi_k(x) y_k$$

where the $\Phi_k(x)$ are functionals on X and $y_k \in Y$.

If the Φ_k are continuous linear functionals, then A_n is a continuous linear operator. Moreover,

Theorem 2.6.4. A continuous finite dimensional linear operator A_n is compact.

Proof. Let S be a bounded set in X. By boundedness of the functionals Φ_k, the numerical set $\{\Phi_k(x) \mid x \in S\}$ is also bounded. Take a sequence $\{x_m\} \subset S$. By boundedness of the numerical sequence $\{\Phi_1(x_m)\}$, we can select a convergent numerical subsequence $\{\Phi_1(x_{m_1})\}$. Considering the numerical sequence $\{\Phi_2(x_{m_1})\}$, we can choose a convergent subsequence $\{\Phi_2(x_{m_2})\}$. Continuing this process, we get a subsequence $\{x_{m_m}\}$ for which each of the sequences $\{\Phi_k(x_{m_m})\}$, $k = 1, \dots, n$, are convergent and thus $\{A_n x_{m_m}\}$ is a Cauchy sequence in Y. □

We apply this theorem to a matrix operator

$$\mathbf{y} = A\mathbf{x}, \qquad \mathbf{y} = (y_1, y_2, \dots), \qquad y_k = \sum_{l=1}^{\infty} a_{kl} x_l, \qquad k = 1, 2, 3, \dots,$$

in ℓ^2. We have shown that

$$\|A\| \le \left(\sum_{k=1}^{\infty} \sum_{l=1}^{\infty} |a_{kl}|^2 \right)^{1/2}.$$

Consider an operator A_n defined as follows:

$$\mathbf{y} = A_n\mathbf{x}, \qquad y_k = \begin{cases} \sum_{l=1}^{\infty} a_{kl}x_l, & k \le n, \\ 0, & k > n. \end{cases}$$

A_n is a finite-dimensional operator and so is compact. Since

$$\|A - A_n\| \le \left(\sum_{k=n+1}^{\infty} \sum_{l=1}^{\infty} |a_{kl}|^2 \right)^{1/2} \to 0 \text{ as } n \to \infty$$

then, thanks to Theorem 2.6.2, A is compact if

$$\sum_{k=1}^{\infty} \sum_{l=1}^{\infty} |a_{kl}|^2 < \infty.$$

Problem 2.6.2. Assume $K(\mathbf{x}, \mathbf{y}) \in L^2(\Omega \times \Omega)$ where Ω is a closed bounded domain in $\mathbb{R}^n$. Show that an integral operator

$$(Af)(\mathbf{x}) = \int_{\Omega} K(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) d\Omega_{\mathbf{y}}$$

is a compact operator in $L^2(\Omega)$.

Problem 2.6.3. Using the Riesz representation theorem, introduce a nonlinear operator K_1 in E_{MC} from the equality

$$(K_1(u), \varphi)_{E_M} = \int_{\Omega} \rho(\mathbf{x}) u^n(\mathbf{x}) \varphi(\mathbf{x})\, d\Omega$$

where $\rho(\mathbf{x})$ is a given bounded piecewise continuous function on a compact set Ω and n is a positive integer. Show that K_1 takes every weakly Cauchy sequence $\{u_m(\mathbf{x})\}$ into the strongly Cauchy sequence $\{K_1 u_m\}$ in E_{MC}.

2.7 Compact Operators in Hilbert Space

In a Hilbert space, the statements of Theorems 2.6.3 and 2.6.4 can be sharpened.

Theorem 2.7.1. An operator A acting in a Hilbert space is compact if and only if it takes every weakly Cauchy sequence $\{x_n\}$ into the strong Cauchy sequence $\{Ax_n\}$ in H.

Proof. Necessity was proved in Theorem 2.6.3. Let us prove sufficiency. Let M be a bounded set in H and AM its image under A. We need to demonstrate that AM is precompact. Take a sequence $\{y_n\}$ belonging to AM and consider the sequence $\{x_n\}$ lying in M such that $Ax_n = y_n$. Since

M is bounded, $\{x_n\}$ is bounded. Thus $\{x_n\}$ contains a subsequence $\{x_{n_k}\}$ that is a weak Cauchy sequence in H. By the condition of the theorem, its image, the sequence $\{Ax_{n_k}\}$, is a (strong) Cauchy sequence in H so AM is precompact. □

Theorem 2.7.2. Assume that A is a compact operator acting in a separable Hilbert space H. Then there is a sequence of finite dimensional continuous linear operators $\{A_n\}$ such that

$$\|A - A_n\| \to 0 \text{ as } n \to \infty.$$

Proof. Let $\{g_n\}$ be an orthonormal basis of H (which exists since H is separable!). Any element f of H can be represented in the form

$$f = \sum_{k=1}^{\infty} (f, g_k) g_k.$$

Then

$$Af = \sum_{k=1}^{\infty} (f, g_k) Ag_k.$$

We denote by A_n a finite dimensional linear operator

$$A_n f = \sum_{k=1}^{n} (f, g_k) Ag_k$$

and by R_n the operator $A - A_n$, and show that $\alpha_n = \|R_n\| \to 0$ as $n \to \infty$.

By definition $\alpha_n = \sup_{\|f\| \le 1} \|R_n f\|$. First we show that there is an element f_n^* such that $\|f_n^*\| \le 1$ and $\alpha_n = \|R_n f_n^*\|$. Indeed, let $\{f_k\}$ be a maximizing sequence such that $\|f_k\| \le 1$ and $\|R_n f_k\| \to \alpha_n$ as $k \to \infty$. Choosing a weakly convergent subsequence $\{f_{k_1}\}$ whose weak limit is f_n^*, thanks to Lemma 1.23.2, we get $\|f_n^*\| \le 1$. As R_n is a compact operator, the sequence $\{R_n f_{k_1}\}$ converges strongly to $R_n f_n^*$. So $\alpha_n = \|R_n f_n^*\|$. (Question for the reader: What is the value of $\|f_n^*\|$?)

But

$$R_n f_n^* = A \left(\sum_{k=n+1}^{\infty} (f_n^*, g_k) g_k \right)$$

so

$$\alpha_n = \|A\varphi_n\|, \qquad \varphi_n = \sum_{k=n+1}^{\infty} (f_n^*, g_k) g_k.$$

We show that the sequence $\{\varphi_k\} \subset H$ converges weakly to zero. Indeed, for an arbitrary continuous linear functional defined by an element f, we

get

$$
\begin{aligned}
|(\varphi_n, f)| &= \left| \left(\sum_{k=n+1}^{\infty} (f_n^*, g_k) g_k, \sum_{m=1}^{\infty} (f, g_m) g_m \right) \right| \\
&= \left| \left(\sum_{k=n+1}^{\infty} (f_n^*, g_k) g_k, \sum_{m=n+1}^{\infty} (f, g_m) g_m \right) \right| \\
&\le \left(\sum_{m=n+1}^{\infty} |(f, g_m)|^2 \right)^{1/2} \|f_n^*\| \to 0 \text{ as } n \to \infty
\end{aligned}
$$

since $\|f_n^*\| \le 1$ and $\sum_{m=1}^{\infty} |(f, g_m)|^2 = \|f\|^2 < \infty$. Since φ_n is weakly convergent to zero and A is compact we have $\|A\varphi_n\| = \alpha_n \to 0$ as $n \to \infty$. This completes the proof. □

Theorems 2.7.1 and 2.7.2 state, in particular, that in energy spaces (which are separable Hilbert spaces) we can use other equivalent definitions of a compact linear operator: such an operator (1) takes every weak Cauchy sequence into a strong Cauchy sequence, or (2) can be approximated with any prescribed accuracy in the operator norm by a finite-dimensional continuous linear operator.

Theorem 2.7.3. If A is a compact linear operator acting in a Hilbert space, then A^* is compact.

Proof. Take a sequence $\{f_n\}$ converging weakly to f_0. It suffices to show that $\{A^* f_n\}$ converges strongly to $A^* f_0$. Indeed,

$$
\begin{aligned}
\|A^* f_n - A^* f_0\|^2 &= (A^* f_n - A^* f_0, A^* f_n - A^* f_0) \\
&= (f_n - f_0, AA^*(f_n - f_0)) \\
&\le \|f_n - f_0\| \, \|AA^*(f_n - f_0)\| \to 0 \text{ as } n \to \infty
\end{aligned}
$$

(here we used the fact that a product AB is compact if the operator B is continuous, and also $\|f_n\| < M = \text{const}$). □

2.8 Functions Taking Values in a Banach Space

We reserve the name "function" for a single-valued correspondence from a subset of $\mathbb{R}^n$ to a Banach space Y. This is a useful convention in many problems of mechanics.

We begin with definitions. A rule that assigns to each point of a domain of $\mathbb{R}^n$ a unique element of a Banach space Y, written $y = f(\mathbf{x})$, $y \in Y$, is called a *function* with values in Y. All notions relative to the functions of classical calculus which are based only on the properties of the metric

are easily transferred to the present setting. For example, $y = f(\mathbf{x})$ is continuous at $\mathbf{x}_0$ if for every positive ε there exists $\delta > 0$ dependent on ϵ such that $\|f(\mathbf{x}) - f(\mathbf{x}_0)\| < \varepsilon$ whenever $|\mathbf{x} - \mathbf{x}_0| < \delta$. If $f(\mathbf{x})$ is continuous at every point of an open domain Ω in $\mathbb{R}^n$ then it is said to be continuous in Ω. Continuity on a closed domain is introduced in a manner similar to the parallel notion in calculus.

The set of all continuous functions given on a closed and bounded domain Ω whose values lie in a Banach space Y is also a Banach space. We denote this space $C(\Omega; Y)$; the norm on $C(\Omega; Y)$ is defined by

$$\|f(\mathbf{x})\|_{C(\Omega;Y)} = \max_{\mathbf{x}\in\Omega} \|f(\mathbf{x})\|_Y.$$

For a function $y = f(t)$, $y \in Y$, $t \in (a, b)$, the derivative df/dt at $t = t_0$ is defined by

$$\frac{df(t_0)}{dt} = \lim_{t\to t_0} \frac{f(t) - f(t_0)}{t - t_0}.$$

Higher order derivatives are introduced similarly. We may also introduce spaces $C^{(k)}(a, b; Y)$ analogous to the corresponding spaces of calculus.

Finally, we can construct the Riemann integral

$$\int_a^b f(t)\, dt$$

for a function with values in a Banach space; the method parallels that of ordinary calculus. There is nothing analogous to the mean value theorem, but we do have, for example,

$$\left\| \int_a^b f(t)\, dt \right\| \le \int_a^b \|f(t)\|\, dt \le \max_{t\in[a,b]} \|f(t)\|(b - a), \quad a < b.$$

These are consequences of passages to the limit in corresponding inequalities for Riemann sums.

The construction of the Lebesgue integral for functions whose values lie in a Banach space is introduced using the completion theorem in a manner similar to that used for scalar-valued functions in Section 1.8.

If functions $y = f(x)$ take their values in a Hilbert space H, we can introduce a Hilbert space $L^2(a, b; H)$ with inner product

$$(f(t), g(t))_{L^2(a,b;H)} = \int_a^b (f(t), g(t))_H\, dt,$$

as the completion of $C(a, b; H)$ in the corresponding norm

$$\|f(t)\|_{L^2(a,b;H)} = \left(\int_a^b \|f(t)\|_H^2\, dt \right)^{1/2}.$$

The particular case $L^2(a,b;L^2(\Omega))$ coincides with the space $L^2([a,b]\times\Omega)$.

In some problems we meet a situation where, in addition to the main energy inner product there is another inner product not depending on time, the product in $L^2(\Omega)$ for example; we denote such an additional inner product by $\langle\cdot,\cdot\rangle$. Assume, as above, that

$$\langle f,f\rangle \le m(f,f)_H \equiv m(f,f) \tag{2.8.1}$$

where the constant m does not depend on $f\in H$.

Analogous to the Sobolev space $W^{1,2}(a,b)$ is a Hilbert space $W^1(a,b)$ in which the inner product is

$$(f(t),g(t))_{W^1(a,b)} = \int_a^b \left[\left\langle \frac{df}{dt},\frac{dg}{dt}\right\rangle + (f,g)\right] dt. \tag{2.8.2}$$

The space $W^1(a,b)$ is the completion of $C^{(1)}(a,b;H)$ in the norm corresponding to (2.8.2).

Thanks to (2.8.1), we get

$$\int_a^b \langle f(t),f(t)\rangle\,dt \le m\int_a^b (f(t),f(t))\,dt \le m\|f\|^2_{W^1(a,b)}.$$

We can obtain some properties of the elements of $W^1(0,T)$ if we take into account the identity

$$f(s+\Delta)-f(s) = \int_s^{s+\Delta} \frac{df(t)}{dt}\,dt$$

which holds for f continuously differentiable. We have

$$\begin{aligned}\int_s^{s+\Delta}\left\langle \frac{df(t)}{dt},f(t)\right\rangle dt &= \frac{1}{2}\int_s^{s+\Delta}\frac{d}{dt}\langle f,f\rangle\,dt\\ &= \frac{1}{2}\langle f(s+\Delta),f(s+\Delta)\rangle - \frac{1}{2}\langle f(s),f(s)\rangle\end{aligned} \tag{2.8.3}$$

and

$$\begin{aligned}\|f(s+\Delta)-f(s)\|_0^2 &= \left\|\int_s^{s+\Delta}\frac{df}{dt}\,dt\right\|_0^2 \le \left(\int_s^{s+\Delta} 1\cdot\left\|\frac{df}{dt}\right\|_0 dt\right)^2\\ &\le \int_s^{s+\Delta} 1^2\,dt\int_s^{s+\Delta}\left\|\frac{df}{dt}\right\|_0^2 dt = \Delta\int_0^T\left\|\frac{df}{dt}\right\|_0^2 dt\end{aligned} \tag{2.8.4}$$

where $\|f\|_0^2 = \langle f,f\rangle$. Passage to the limit shows that they remain valid for elements of $W^1(0,T)$. This means that the elements of $W^1(0,T)$ are

continuous in t with respect to the norm $\|\cdot\|_0$. This is an imbedding theorem for $W^1(0,T)$.

We now turn to the problem of holomorphic functions.

A function $f(\lambda)$ given on an open domain G of the complex plane $\mathbb{C}$ to a Banach space X is called *holomorphic* in G if for each point $\lambda_0 \in G$ there is a neighborhood $D(\lambda_0)$ of λ_0 in which there is a power series expansion

$$f(\lambda) = f(\lambda_0) + \sum_{k=1}^{\infty} c_k(\lambda - \lambda_0)^k, \qquad \lambda \in D(\lambda_0)$$

converging uniformly by the norm of X in $D(\lambda_0)$.

A holomorphic vector valued function has properties similar to those of a scalar valued function. For example, if $f(\lambda)$ is holomorphic in $|\lambda - \lambda_0| < R$ and $\|f(\lambda)\| \le M$, then it is infinitely differentiable in this disk, the Taylor expansion

$$f(\lambda) = \sum_{n=0}^{\infty} \frac{f^{(n)}(\lambda_0)}{n!}(\lambda - \lambda_0)^n, \qquad |\lambda - \lambda_0| < R,$$

exists, and

$$\|f^{(n)}(\lambda_0)\| \le MR^{-n}n!.$$

Then if $f(\lambda)$ is a holomorphic function, $f(\lambda) \in X$, on the domain G, then

$$\oint_C f(\lambda)\, d\lambda = 0$$

for every simple closed rectifiable contour C in G such that the interior of C belongs to G. The Cauchy representation

$$f(\lambda) = \frac{1}{2\pi i} \oint_C \frac{f(z)}{z - \lambda}\, dz$$

is also valid for λ lying in the interior of C.

These and similar results can be found in Yosida [29].

2.9 Spectrum of Linear Operators

In problems of mechanics of continuous media there arises an operator equation

$$x - A(\mu)x = f \tag{2.9.1}$$

in a Banach space X, where $A(\mu)$ is a linear operator depending on a real or complex parameter. A typical example is an equation describing steady vibrations of elastic bodies, which has the form

$$x - \mu Ax = f.$$

In particular, eigen-oscillations of a string with fixed ends are governed by the boundary value problem

$$\lambda x + x'' = 0, \qquad x(0) = x(1) = 0,$$

$\mu = 1/\lambda$. Another instance of equation (2.9.1) is the equation

$$\left(I + \sum_{k=1}^{n} \mu^k A_k \right) x = f$$

which appears, for example, in the theory of an elastic band.

Let us introduce some notation. A value μ_0 is called a *regular point* of the operator $A(\mu)$ if there is a bounded inverse $(I - A(\mu))^{-1}$ whose domain is dense in X; otherwise, μ_0 belongs to the *spectrum* of $A(\mu)$.

The same terms will be used for an operator A: λ is a regular point of A if there is a bounded inverse $R(\lambda, A) = (\lambda I - A)^{-1}$ with domain dense in X; otherwise, λ is a point of the spectrum of A.

The set of all regular points of A is called the *resolvent set* of A. It is denoted by $\rho(A)$.

A point λ may fail to be a regular point of A for several reasons, and the set of spectrum points of A can be thereby classified into three types:

1. *Point spectrum.* This is the set of all complex λ such that $(\lambda I - A)$ does not have an inverse. The equation $(\lambda I - A)x = 0$ then has a nontrivial solution called an *eigensolution*, and λ is an *eigenvalue* of A.

2. *Continuous spectrum.* The set of all $\lambda \in \mathbb{C}$ such that there exists $(\lambda I - A)^{-1}$ whose domain $D(R(\lambda, A))$ is dense in X, but such that the operator $(\lambda I - A)^{-1}$ is not bounded.

3. *Residual spectrum.* The set of $\lambda \in \mathbb{C}$ such that $R(\lambda, A) = (\lambda I - A)^{-1}$ exists but with domain not dense in X.

We consider some examples.

1. A matrix operator acting in the n-dimensional Euclidean space has only a point spectrum consisting of no more than n points called the eigenvalues of the matrix. Other points of the complex plane are regular.

2. Any point of the complex plane belongs to the point spectrum of the differentiation operator d/dt acting in $C(a, b)$, since for every λ the equation

$$\frac{df}{dt} - \lambda f = 0$$

has a solution $f(t) = ce^{\lambda t}$ where c is a constant. So the operator d/dt in $C(a, b)$ has no regular points. (Question for the reader: What happens to the spectrum if the domain of d/dt is the subspace of $C(a, b)$ consisting of functions $f(x)$ that satisfy $f(a) = 0$?)

3. On the square $[0,\pi]\times[0,\pi]$ we consider the boundary value problem

$$\frac{\partial^2 u}{\partial x^2}+\lambda\frac{\partial^2 u}{\partial y^2}=f(x,y),\qquad u\big|_{\partial\Omega}=0. \tag{2.9.2}$$

The operator $B(\lambda)$ on the left-hand side of this equation is considered in $L^2(\Omega)$, $\Omega=[0,\pi]\times[0,\pi]$. Note that we keep the terminology for the spectrum although this operator does not have the form (2.9.1).

The Fourier expansion of $f(x,y)$ is

$$f(x,y)=\sum_{m,n=1}^{\infty} f_{mn}\sin mx\sin ny,\qquad \sum_{m,n=1}^{\infty}|f_{mn}|^2<\infty.$$

Let λ be a point of $\mathbb{C}$ — not on the negative real axis, but otherwise arbitrary. Then the solution of (2.9.2) is

$$u(x,y)=\sum_{m,n=1}^{\infty}-\frac{f_{mn}}{m^2+\lambda n^2}\sin mx\sin ny.$$

If $\lambda=\lambda_1+i\lambda_2$ is such that $\lambda_2\neq 0$, or if $\lambda_2=0$ but $\lambda_1\geq 0$, then $|m^2+\lambda n^2|>\delta>0$ for all integers $m,n\geq 1$. Therefore

$$\|u(x,y)\|^2_{L^2(\Omega)}=\frac{\pi^2}{4}\sum_{m,n=1}^{\infty}|f_{mn}|^2\frac{1}{|m^2+\lambda n^2|^2}\leq\frac{\pi^2}{4\delta^2}\sum_{m,n=1}^{\infty}|f_{mn}|^2$$

and it follows that

$$\|u\|_{L^2(\Omega)}\leq\frac{1}{\delta}\|f\|_{L^2(\Omega)}. \tag{2.9.3}$$

There are no other solutions to (2.9.2) so the inequality (2.9.3) means that the inverse is bounded and thus these λ belong to the resolvent set of the operator $B(\lambda)$ acting in $L^2(\Omega)$. What can we say about $\lambda\in\mathbb{C}$ such that $\lambda=\operatorname{Re}\lambda<0$?

First we consider λ of the form $\lambda=-p^2/q^2$ where p and q are integers. For these λ the corresponding boundary value problem is not solvable for some $f(x,y)$. To show this, take $f(x,y)=\sin px\sin qy$. As is easily seen, if there is a solution to (2.9.2) then it has the form $u(x,y)=c\sin px\sin qy$ and for $\lambda_0=-p^2/q^2$ must satisfy the equation $c(p^2+\lambda_0 q^2)=-1$, which is impossible. Moreover, $u=\sin px\sin qy$ is a solution to the homogeneous equation (2.9.2) at $\lambda=\lambda_0$. So all $\lambda=-p^2/q^2$, where p,q are integers, belong to the point spectrum of $B(\lambda)$.

We consider the remaining part M of the negative real axis, i.e., the set of λ such that $\lambda=\operatorname{Re}\lambda<0$ and λ cannot be represented in the form $-p^2/q^2$ for some integers p,q. For $\lambda\in M$ we can seek a solution in the form of a Fourier series

$$u(x,y)=\sum_{m,n=1}^{\infty}\frac{f_{mn}}{m^2+\lambda n^2}\sin mx\sin ny.$$

The set S of functions $f(x,y)$ of the form $\sum_{m=1}^{N_1}\sum_{n=1}^{N_2} f_{mn}\sin mx\sin ny$ is dense in $L^2(\Omega)$; since solutions corresponding to these $f(x,y)\in S$ are also in $L^2(\Omega)$ and defined uniquely, the inverse of $B(\lambda)$ is determined on a dense subset of $L^2(\Omega)$. We show that it is unbounded for $\lambda\in M$. Indeed, the set of all points of the form $\lambda_{pq}=-p^2/q^2$ is dense in M. Take $\lambda\in M$; there is a sequence $\lambda_n=-p_n^2/q_n^2\to\lambda$ as $n\to\infty$. Take the function $f_n(x,y)=\sin p_n x\sin q_n y$, which is the right-hand side of (2.9.2). To this, there corresponds the solution

$$u_n(x,y)=-\frac{1}{p_n^2+\lambda q_n^2}\sin p_n x\sin q_n y.$$

Their norms are related by

$$\|u_n\|_{L^2(\Omega)}=\frac{1}{|p_n^2+\lambda q_n^2|}\|f_n\|_{L^2(\Omega)}$$

where $|p_n^2+\lambda q_n^2|\to 0$ as $n\to\infty$. So the inverse to the operator $B(\lambda)$ is unbounded when $\lambda\in M$, and thus M is a subset of the continuous spectrum.

4. Now we consider the so-called coordinate operator in $C(a,b)$, defined as

$$(Qu)(t)=tu(t).$$

This operator has no eigenvalues. If $\lambda\notin[a,b]$, then it belongs to the resolvent set of Q since the equation

$$\lambda u(t)-tu(t)=f(t)$$

has the unique solution

$$u(t)=\frac{f(t)}{\lambda-t}$$

in $C(a,b)$.

But if $\lambda\in[a,b]$, then there exists the inverse defined by $u(t)=f(t)/(\lambda-t)$ whose domain consists of functions which can be represented in the form $f(t)=(\lambda-t)z(t)$ with $z(t)\in C(a,b)$. This domain is not dense in $C(a,b)$, hence the points of $[a,b]$ belong to the residual spectrum.

Problem 2.9.1. Show that for the coordinate operator acting in $L^2(a,b)$, the points of $[a,b]$ belong to the continuous spectrum.

2.10 Resolvent Set of a Closed Linear Operator

Theorem 2.10.1. Assume that A is a closed linear operator acting in a complex Banach space X. For any λ_0 belonging to the resolvent set of A, the resolvent $R(\lambda_0,A)=(\lambda_0 I-A)^{-1}$ is a continuous linear operator defined on the whole of X.

Proof. By definition of resolvent set, the domain of $R(\lambda_0, A)$ is dense in X and there is a constant $m > 0$ such that

$$\|(\lambda_0 I - A)x\| \geq m\|x\|. \tag{2.10.1}$$

Take $y \in X$; by definition, there is a sequence $\{x_n\}$ such that $\lim_{n\to\infty}(\lambda_0 I - A)x_n = y$ (strong limit). By (2.10.1), we have $\lim_{n\to\infty} x_n = x$ and $(\lambda_0 I - A)x = y$ since A is closed. So the range of $(\lambda_0 I - A)^{-1}$ is X. □

Theorem 2.10.2. Assume that A is a closed linear operator acting in a complex Banach space X. Then the resolvent set $\rho(A)$ is an open domain of $\mathbb{C}$ and $R(\lambda, A)$ is holomorphic with respect to λ in $\rho(A)$.

Proof. For any $\lambda \in \rho(A)$, as was shown above, $R(\lambda, A)$ is a continuous linear operator on X. So the series

$$R(\lambda_0, A)\left\{I + \sum_{n=1}^{\infty}(\lambda_0 - \lambda)^n R^n(\lambda_0, A)\right\}$$

is convergent in the disk $|\lambda - \lambda_0| < 1/\|R(\lambda_0, A)\|$ of $\mathbb{C}$ and thus is a holomorphic function in this disk. Multiplying this series by $(\lambda I - A) = (\lambda - \lambda_0)I + (\lambda_0 I - A)$, we get I, i.e., it is an inverse to $\lambda I - A$. □

Theorem 2.10.3. Under the condition of Theorem 2.10.2, the Hilbert identity

$$R(\lambda, A) - R(\mu, A) = (\mu - \lambda)R(\lambda, A)R(\mu, A)$$

holds for any $\lambda, \mu \in \rho(A)$.

Proof. The identity follows from

$$\begin{aligned} R(\lambda, A) &= R(\lambda A)(\mu I - A)R(\mu, A) \\ &= R(\lambda, A)\{(\mu - \lambda)I + (\lambda I - A)\}R(\mu, A) \\ &= (\mu - \lambda)R(\lambda, A)R(\mu, A) + R(\mu, A) \end{aligned}$$

since $R(\lambda, A)(\lambda I - A) = I$. □

Let B be a bounded linear operator in X. Then the series

$$\frac{1}{\lambda}\left(I + \sum_{n=1}^{\infty}\lambda^{-n}B^n\right)$$

is convergent if $|\lambda| > \|B\|$. Multiplying it by $\lambda I - B$ we get I, i.e.,

$$R(\lambda, B) = \frac{1}{\lambda}\left(I + \sum_{n=1}^{\infty}\lambda^{-n}B^n\right)$$

for $|\lambda| \geq \|B\|$.

Lemma 2.10.1. The expansion

$$R(\lambda, B) = \frac{1}{\lambda}\left(I + \sum_{n=1}^{\infty} \lambda^{-n} B^n\right)$$

is valid in the domain $|\lambda| > r_\sigma(B)$, where $r_\sigma(B)$ is the spectral radius of B defined by

$$r_\sigma(B) = \lim_{n\to\infty} \|B^n\|^{1/n}.$$

Proof. We show that $r_\sigma(B)$ exists. Denote $r_0 = \inf_n \|B^n\|^{1/n}$. We establish that $r_\sigma(B) = r_0$. By definition of infimum, for any positive ε we can find an integer N such that

$$\|B^N\|^{1/N} \le r_0 + \varepsilon.$$

For large n represented as $n = kN + l$, $0 \le l < N$, we get

$$\begin{aligned}\|B^n\|^{1/n} &\le \|B^{kN}\|^{1/n}\|B^l\|^{1/n}\\ &\le \|B^N\|^{k/n}\|B^l\|^{1/n}\\ &\le (r_0+\varepsilon)^{kN/n}\|B^l\|^{1/n}\\ &\le r_0 + \varepsilon + \varepsilon_1(n)\end{aligned}$$

where $\varepsilon_1(n) \to 0$ as $n \to \infty$. Together with the inequality $\|B^n\|^{1/n} \ge r_0$, this proves that $r_\sigma(B)$ exists and is equal to r_0. The rest of the proof is trivial. □

Problem 2.10.1. Let $A(\mu)$ be a continuous operator-function in X which is holomorphic with respect to μ in $\mathbb{C}$. Show that the resolvent set $\rho(A(\mu))$ of $A(\mu)$ is open and $(I - A(\mu))^{-1}$ is a holomorphic operator-function in $\rho(A(\mu))$.

2.11 Spectrum of Compact Operators in Hilbert Space

An important class of operators for which there is a full description of the spectrum is the class of compact linear operators. The first results in this direction were due to I. Fredholm; studying the integral operator, he established properties of its spectrum similar to those of the matrix operator. The theory was then extended to the class of compact operators (F. Riesz, J. Schauder) which we now consider in a Hilbert space. The theory is of great interest as it describes eigen-oscillations of bounded elastic bodies.

So let A be a compact linear operator acting in a Hilbert space H. We are seeking eigenvectors and eigenvalues of A, i.e., nontrivial solutions to the equation

$$(I - \mu A)x = 0. \tag{2.11.1}$$

In the previous section we used traditional spectrum terminology from functional analysis. From now on we use the term "eigenvalue" for μ from (2.11.1) but not for λ from the equation $(\lambda I - A)x = 0$. The reason is that in later applications of the theory we consider a non-traditional introduction of the operator equations for oscillations that are composed in energy spaces; in this way, the term "eigenvalue" of problems of mechanics refers to μ.

The Fredholm–Riesz–Schauder theory will be presented as a number of lemmas and theorems. We want to underline that the results on the properties of eigenvectors corresponding to a fixed eigenvalue are the same for compact linear operators $A(\mu)$ with general dependence on μ.

Let μ_0 be an eigenvalue of A. By continuity and linearity of A, the set of all eigenvectors corresponding to μ_0, when adjoined to the zero vector 0, is a closed subspace of H denoted by $H(\mu_0)$.

Lemma 2.11.1. $H(\mu_0)$ is finite dimensional.

Proof. By definition of compact operator and the equality $x = \mu_0 Ax$, $x \in H(\mu_0)$, from any bounded sequence $\{x_k\} \subset H(\mu_0)$ we can choose a Cauchy subsequence. This means that every bounded subset of $H(\mu_0)$ is precompact and, by Theorem 1.16.3, $H(\mu_0)$ is finite dimensional. □

Lemma 2.11.2. Assume $\{x_1^{(i)}, x_2^{(i)}, \dots, x_{n_i}^{(i)}\}$ is a linearly independent system of elements in $H(\mu_i)$. Then the union

$$\{x_1^{(1)}, x_2^{(1)}, \dots, x_{n_1}^{(1)}\}, \dots, \{x_1^{(k)}, x_2^{(k)}, \dots, x_{n_k}^{(k)}\}$$

is a linearly independent system in the space $H(\mu_1)\dot{+}\cdots\dot{+}H(\mu_k)$ whose elements are linear combinations of elements of the spaces $H(\mu_i)$, $i = 1, \dots, k$. If each $\{x_1^{(i)}, x_2^{(i)}, \dots, x_{n_i}^{(i)}\}$ is a basis of $H(\mu_i)$, $i = 1, \dots, k$, then their union is a basis in $H(\mu_1)\dot{+}\cdots\dot{+}H(\mu_k)$.

Proof. It suffices to show that the union of the $x_{n_i}^{(j)}$ is a linearly independent system. Let us renumber successively the eigenvectors $x_{n_i}^{(j)}$ and eigenvalues μ_i in such a way that μ_k corresponds to x_k. The proof of independence is carried out by induction. Assume a system $x_1, \dots, x_n$ is linearly independent. Let us add the next eigenvector x_{n+1} and consider the equation

$$\sum_{k=1}^{n+1} c_k x_k = 0$$

with respect to the coefficients c_k. Applying A to both sides we get

$$\sum_{k=1}^{n+1} c_k A x_k = 0$$

or, since x_k is an eigenvector of A, we have

$$\mu_{n+1} \sum_{k=1}^{n+1} \frac{c_k}{\mu_k} x_k = 0.$$

Subtracting this from $\sum_{k=1}^{n+1} c_k x_k = 0$, we have

$$\sum_{k=1}^{n} c_k \left(1 - \frac{\mu_{n+1}}{\mu_k}\right) x_k = 0.$$

By assumption, $c_k = 0$ for those $k = 1, \ldots, n-s$ for which $\mu_k \neq \mu_{n+1}$. For the rest of the eigenvectors, that is for $x_{n-s+1}, \ldots, x_{n+1}$, we have

$$\sum_{k=n-s+1}^{n+1} c_k x_k = 0.$$

In this all x_k correspond to μ_{n+1} and so are linearly independent. Thus $c_k = 0$ for $k = 1, \ldots, n+1$. □

Lemma 2.11.3. The set of eigenvalues of a compact linear operator A has no finite limit points in $\mathbb{C}$.

Proof. Suppose there is a sequence of distinct eigenvalues $\mu_n \to \mu_0$, $|\mu_0| < \infty$. For each μ_n take an eigenvector x_n. Denote by H_n the subspace spanned by $x_1, x_2, \ldots, x_n$. By Lemma 2.11.2, $H_n \subset H_{n+1}$ and $H_n \neq H_{n+1}$ so there is an element $y_{n+1} \in H_{n+1}$ such that $\|y_{n+1}\| = 1$ and y_{n+1} is orthogonal to H_n.

The sequence $\{\mu_n y_n\}$ is bounded in H so the sequence $\{A(\mu_n y_n)\}$ must contain a Cauchy subsequence. But this is impossible as is shown below. Indeed

$$A(\mu_{n+m} y_{n+m}) - A(\mu_n y_n) = y_{n+m} - (y_{n+m} - \mu_{n+m} A y_{n+m} + \mu_n A y_n). \tag{2.11.2}$$

Now $y_{n+m} \in H_{n+m}$. We show that the term in parentheses on the right-hand side belongs to H_{n+m-1} ($m \geq 1$); indeed $y_{n+m} = \sum_{k=1}^{n+m} c_k x_k$ and so

$$\begin{aligned} y_{n+m} - \mu_{n+m} A y_{n+m} &= \sum_{k=1}^{n+m} c_k x_k - \mu_{n+m} A \left(\sum_{k=1}^{n+m} c_k x_k \right) \\ &= \sum_{k=1}^{n+m-1} c_k \left(1 - \frac{\mu_{n+m}}{\mu_k}\right) x_k \in H_{n+m-1} \end{aligned}$$

since $\mu_n A y_n \in H_n \subset H_{n+m-1}$ and thus we proved the needed property for the term in parentheses.

Thus y_{n+m} and $(y_{n+m} - \mu_{n+m} A y_{n+m} + \mu_n A y_n)$ are mutually orthogonal. From (2.11.2) it follows that

$$\begin{aligned} \|A(\mu_{n+m} y_{n+m}) - A(\mu_n y_n)\|^2 &= \|y_{n+m}\|^2 + \\ &\quad + \|y_{n+m} - \mu_{n+m} A y_{n+m} + \mu_n A y_n\|^2 \\ &\geq 1, \end{aligned}$$

which contradicts the fact that $\{A(\mu_n y_n)\}$ contains a Cauchy sequence. □

Combining these three lemmas, we formulate

Theorem 2.11.1. There are no more than a countable number of eigenvalues of a compact linear operator acting in a Hilbert space; the set of eigenvalues has no finite limit point in $\mathbb{C}$. A subspace $H(\mu_k)$ of all eigenvectors of A corresponding to a μ_k is finite dimensional and $H(\mu_k) \cap H(\mu_n) = 0$ if $\mu_k \neq \mu_n$.

It is time to formulate

Theorem 2.11.2. A compact linear operator A in a Hilbert space has a point spectrum only.

The proof follows immediately from Theorem 2.11.3 and Lemma 2.11.4.

Let us denote by $M(\mu_0)$ the orthogonal complement to $H(\mu_0)$ in H (which is possible by the theorem on the orthogonal decomposition of a Hilbert space).

Lemma 2.11.4. There are constants $m_1 > 0$ and $m_2 > 0$ such that

$$m_1 \|x\| \leq \|x - \mu_0 A x\| \leq m_2 \|x\| \tag{2.11.3}$$

for all $x \in M(\mu_0)$.

Proof. The right-most inequality is evident; let us prove the left-most inequality. Suppose there is no $m_1 > 0$ such that the inequality holds for all $x \in M(\mu_0)$. This means there is a sequence $\{x_n\} \subset M(\mu_0)$ such that $\|x_n\| = 1$ and $\|x_n - \mu_0 A x_n\| \to 0$ as $n \to \infty$. Because A is compact, the sequence $\{A x_n\}$ contains a Cauchy subsequence. By the identity

$$x_n = \mu_0 A x_n + (x_n - \mu_0 A x_n)$$

the sequence $\{x_n\}$ also contains a Cauchy subsequence. Let us denote this Cauchy subsequence by $\{x_n\}$ again and let it converge to an element $x_0 \in M(\mu_0)$. Since $A x_n \to A x_0$ as $n \to \infty$, we have $x_0 = \mu_0 A x_0$ and so $x_0 \in H(\mu_0)$. However, this contradicts the fact that $x_0 \in M(\mu_0)$. □

The inequalities (2.11.3) state that on $M(\mu_0)$ we can introduce the norm $\|x\|_1 = \|x - \mu_0 Ax\|$ (and the scalar product) which is equivalent to the norm of H.

Now we begin to treat the problem of solvability of the equation

$$x - \mu Ax = f$$

in detail. So as not to complicate the presentation and to include into consideration the case of general dependence on μ of $A(\mu)$ at once, we denote $\mu A = B$ (or $A(\mu) = B$) and study the equation

$$x - Bx = f \tag{2.11.4}$$

with a compact linear operator B acting in a Hilbert space.

We denote by N a subspace of eigenvectors of B corresponding to $\mu = 1$, i.e., all solutions to

$$x = Bx,$$

and by M the orthogonal complement to N in H. For B^*, the adjoint of B, which is also a compact operator, we denote by N^* the space of all eigenvectors $x = B^*x$ and by M^* the orthogonal complement to N^* in H.

Lemma 2.11.5. The equation

$$x - B^*x = f \tag{2.11.5}$$

has a solution if and only if $f \in M$.

Proof. Necessity. Let (2.11.5) have a solution x_0. Then for an arbitrary element y of N we get

$$(f, y) = (x_0 - B^*x_0, y) = (x_0, y - By) = (x_0, 0) = 0,$$

i.e., f does belong to M.

Sufficiency. Let $f \in M$. We mentioned that $\|x\|_1 = \|x - Bx\|$ is a norm in M which is equivalent to the norm of H in M; we can say the same about the inner product $(x, y)_1 = (x - Bx, y - By)$ in M.

The functional (x, f) is linear and continuous on H (and so on M) and, by the Riesz representation theorem, it can be represented on M using $(\cdot, \cdot)_1$ as

$$(x, f) = (x, f^*)_1 = (x - Bx, f^* - Bf^*).$$

This equality, being valid for $x \in M$, holds for all $x \in H$ too; indeed bearing $x = x_1 + x_2$, $x_1 \in N$, $x_2 \in M$, we have

$$x - Bx = x_1 - Bx_1 + x_2 - Bx_2 = x_2 - Bx_2$$

and so, for all $x \in H$,

$$(x - Bx, f^* - Bf^*) = (x_2 - Bx_2, f^* - Bf^*) = (x_2, f^*)_1 = (x_2, f) = (x, f)$$

since $(x_1, f) = 0$. We denote $f^* - Bf^*$ by g; then

$$(x - Bx, g) = (x, f) \text{ for all } x \in H.$$

It follows that

$$(x, g - B^*g) = (x, f) \text{ for all } x \in H.$$

Therefore g is a solution to (2.11.5). □

Since B and B^* are mutually adjoint, we get

Corollary 2.11.1. The equation

$$x - Bx = f \tag{2.11.6}$$

has a solution if and only if $f \in M^*$.

From the inequality (2.11.3) there follows

Corollary 2.11.2. For all $f \in M^*$ there is a positive constant m not depending on f such that

$$\|x\| \leq m\|f\|$$

where x is a solution to (2.11.6).

Lemma 2.11.5 and Corollary 2.11.1 can be reformulated in other terms as

$$R(I - B) = M^*, \qquad R(I - B^*) = M,$$

where $R(S)$ is the range of operator S.

Lemma 2.11.6. Let N_n be the space of all solutions of the equation

$$(I - B)^n x = 0$$

(which is also called the kernel of $(I - B)^n$). Then

(i) N_n is a finite dimensional subspace of H;

(ii) for all $n = 1, 2, \ldots$, $N_n \subseteq N_{n+1}$;

(iii) there is an integer k such that $N_n = N_k$ if $n > k$.

Proof. Since $(I-B)^n = I - nB + \cdots$, then $(I-B)^n$ has a structure $I - B_1$ where B_1 is a compact linear operator. Hence (i) is fulfilled. Property (ii) is evident.

To check (iii) we first mention that if for some k we have $N_{k+1} = N_k$ then $N_{k+m} = N_k$ for all $m = 1, 2, 3, \ldots$; indeed, in this case take $x_0 \in N_{k+2}$ so that

$$0 = (I - B)^{k+2}x_0 = (I - B)^{k+1}((I - B)x_0),$$

i.e., $(I-B)x_0 \in N_{k+1}$. But this means $(I-B)x_0 \in N_k$ or $(I-B)^{k+1}x_0 = 0$, and so $x_0 \in N_{k+1} = N_k$. Thus $N_{k+2} = N_k$. This can be repeated for any $m > 2$: $N_{k+m} = N_k$.

Now suppose to the contrary that there is no k such that $N_k = N_{k+1}$. Then there is a sequence of elements $\{x_n\}$ such that $x_n \in N_n$, $\|x_n\| = 1$, and x_n is orthogonal to N_{n-1}. Let us consider the sequence $\{Bx_n\}$. By compactness of B it must contain a Cauchy subsequence but this leads to a contradiction. Indeed, we have

$$Bx_{n+m} - Bx_n = x_{n+m} - (x_{n+m} - Bx_{n+m} + Bx_n),$$

where $x_{n+m} \in N_{n+m}$ and $(x_{n+m} - Bx_{n+m} + Bx_n) \in N_{n+m-1}$ since $Bx_n \in N_n$:

$$(I-B)^n Bx_n = B(I-B)^n x_n = 0$$

and

$$(I-B)^{n+m-1}(x_{n+m} - Bx_{n+m}) = (I-B)^{n+m}x_{n+m} = 0.$$

Therefore x_{n+m} is orthogonal to $(x_{n+m} - Bx_{n+m} + Bx_n)$ and so

$$\|Bx_{n+m} - Bx_n\|^2 = \|x_{n+m}\|^2 + \|x_{n+m} - Bx_{n+m} + Bx_n\|^2 \geq 1.$$

This means that $\{Bx_n\}$ does not contain a Cauchy subsequence. □

Theorem 2.11.3. *$R(I-B) = H$ if and only if $N = 0$.*

Proof. Necessity. Let $R(I-B) = H$ and suppose that $N \neq 0$. Take $x_0 \in N$, $x_0 \neq 0$. Since $R(I-B) = H$ we can solve successively the following infinite system of equations:

$$(I-B)x_1 = x_0; \quad (I-B)x_2 = x_1; \quad \cdots \quad (I-B)x_{n+1} = x_n; \quad \cdots$$

The sequence of solutions has the following property:

$$(I-B)^n x_n = x_0 \neq 0 \qquad \text{but} \qquad (I-B)^{n+1}x_n = (I-B)x_0 = 0,$$

i.e., there is no finite k such that $N_{k+1} = N_k$. This contradicts statement (iii) of Lemma 2.11.6.

Sufficiency. This proof is a chain of direct implications of the type that lends itself to proof by a computer. Let $N = 0$. Then $M = H$ and so, by Lemma 2.11.5, $R(I - B^*) = M = H$. By the necessity part of the proof given above, $N^* = 0$ and thus $M^* = H$. By Corollary 2.11.1, we get $R(I-B) = M^* = H$. □

Corollary 2.11.3. *If $R(I-B) = H$, then the inverse $(I-B)^{-1}$ is continuous.*

This follows from (2.11.3) written in terms of B.

Theorem 2.11.4. *The spaces N and N^* have the same dimension.*

Proof. Let the dimensions of N and N^* be n and m, respectively, and suppose that $m > n$. Choose orthonormal bases $x_1, \dots, x_n$ and $y_1, \dots, y_m$ of N and N^*, respectively. Let us introduce an auxiliary operator Q by

$$Qx = (I - B)x + \sum_{k=1}^{n} (x, x_k) y_k \equiv (I - C)x,$$

where C is a compact linear operator as the sum of the compact operator $-B$ and a finite dimensional operator.

First we show that the kernel of Q does not contain nonzero elements. Indeed if $Qx_0 = 0$ then

$$(I - B)x_0 + \sum_{k=1}^{n} (x_0, x_k) y_k = 0.$$

Since $R(I - B) = M^* \perp N^*$, all terms of the sum on the left-hand side of the equality are mutually orthogonal and so each of them equals zero:

$$(I - B)x_0 = 0, \qquad (x_0, x_k) y_k = 0, \quad k = 1, \dots, n.$$

From $(I - B)x_0 = 0$ it follows that $x_0 \in N$, the remainder means x_0 is orthogonal to all basis elements of N, thus $x_0 = 0$.

By Theorem 2.11.3, the range of Q is H and thus the equation $Qx = y_{n+1}$ has a solution x_0. But we get

$$\begin{aligned} 1 &= (y_{n+1}, y_{n+1}) \\ &= (y_{n+1}, Qx_0) \\ &= (y_{n+1}, (I - B)x_0) + \left(y_{n+1}, \sum_{k=1}^{n} (x_0, x_k) y_k \right) \\ &= ((I - B^*) y_{n+1}, x_0) \\ &= 0. \end{aligned}$$

This contradiction shows that $m \le n$. On the other hand, B is adjoint to B^* and, by the above, $n \le m$. Thus $n = m$. □

Remark 2.11.1. In the last proof we used the operator Q, which was continuously invertible on H. The same property holds for an operator Q_ε defined by

$$Q_\varepsilon = (I - B)x + \varepsilon \sum_{k=1}^{n} (x, x_k) y_k$$

with any small $\varepsilon \neq 0$. This operator has a continuous inverse and

$$\|I - B - Q_\varepsilon\| = O(\varepsilon).$$

So Q_ε close to $I - B$ allows us to solve the equation $Q_\varepsilon x = f$ for any $f \in H$, whereas the original equation $Bx = f$ has no solution for some f. Such operators are called *regularizers* and are widely used in applications.

Remark 2.11.2. Collectively, the results of this section are known as the *Fredholm alternative*.

2.12 Analytic Nature of the Resolvent of a Compact Linear Operator

We know that the resolvent $(I-\mu A)^{-1}$ is a holomorphic operator-function in μ in non-spectral points. But what is its behavior near the spectrum? We can answer this question for a compact operator.

We begin the study with the case of a continuous finite-dimensional operator acting in a Hilbert space. The general form of such an operator is

$$A_n x = \sum_{k=1}^{n} (x, a_k) x_k$$

where the system $x_1, \dots, x_n$ is assumed to be linearly independent.

We consider the equation

$$x - \mu \sum_{k=1}^{n} (x, a_k) x_k = f. \tag{2.12.1}$$

Its solution has the form

$$x = f + \sum_{k=1}^{n} c_k x_k.$$

Placing this into (2.12.1), we get

$$f + \sum_{k=1}^{n} c_k x_k - \mu \sum_{k=1}^{n} \left(f + \sum_{j=1}^{n} c_j x_j, a_k \right) x_k = f$$

or

$$\sum_{k=1}^{n} x_k \left(c_k - \mu \sum_{j=1}^{n} c_j (x_j, a_k) \right) = \mu \sum_{k=1}^{n} (f, a_k) x_k.$$

Since $x_1, \dots, x_n$ is a linearly independent system we get an algebraic system

$$c_k - \mu \sum_{j=1}^{n} (x_j, a_k) c_j = \mu (f, a_k), \qquad k = 1, \dots, n, \tag{2.12.2}$$

which can be solved using Cramer's rule:

$$c_k = \frac{D_k(\mu, f)}{D(\mu)}, \qquad k = 1, \dots, n.$$

Thus a solution to (2.12.1) is

$$x = \frac{D(\mu)f + \sum_{k=1}^{n} D_k(\mu, f)x_k}{D(\mu)}.$$

In this case the solution to (2.12.1) is a ratio of two polynomials in μ of degree no more than n. All μ which are not eigenvalues of A_n are points where the resolvent is holomorphic, hence they cannot be zeros of $D(\mu)$. But if μ_0 is an eigenvalue of A_n then $D(\mu_0) = 0$. If it is not true then for any $f \in H$ there is a solution to (2.12.1) and this means μ_0 is not an eigenvalue. So the set of all zeros of $D(\mu)$ coincides with the set of all eigenvalues of A_n, and so each eigenvalue of A_n is a pole of finite multiplicity of the resolvent $(I - \mu A_n)^{-1}$.

Now we consider a general case:

Theorem 2.12.1. Every eigenvalue of a compact linear operator A acting in a Hilbert space is a pole of finite multiplicity of the resolvent $(I-\mu A)^{-1}$.

Proof. It was shown (Theorem 2.7.2) that for any small $\varepsilon > 0$, the operator A can be represented as

$$A = A_n + A_\varepsilon, \qquad \|A_\varepsilon\| \le \varepsilon,$$

where A_n is a finite dimensional operator. Fix $\varepsilon > 0$. The equation under consideration takes the form

$$x - \mu(A_n + A_\varepsilon)x = f. \tag{2.12.3}$$

Consider the operator $(I - \mu A_\varepsilon)^{-1}$. In the disk $|\mu| < 1/\varepsilon$ it can be represented in the form

$$(I - \mu A_\varepsilon)^{-1} = I + \sum_{k=1}^{\infty} \mu^k A_\varepsilon^k.$$

(The series is majorized by the series $1 + \sum_{k=1}^{\infty} |\mu^k| \|A_\varepsilon\|^k$. The fact that it is inverse to $(I - \mu A_\varepsilon)$ is checked directly.) So in the disk $|\mu| < 1/\varepsilon$ the operator $(I - \mu A_\varepsilon)^{-1}$ is a holomorphic operator-function in μ.

We apply this operator to equation (2.12.3):

$$x - \mu(I - \mu A_\varepsilon)^{-1} A_n x = (I - \mu A_\varepsilon)^{-1} f$$

wherein, as above,

$$A_n x = \sum_{k=1}^{n} (x, a_k) x_k.$$

Let us denote

$$f^* = (I - \mu A_\varepsilon)^{-1} f, \qquad x_k^* = (I - \mu A_\varepsilon)^{-1} x_k;$$

then the equation takes the form

$$x - \mu \sum_{k=1}^{n} (x, a_k) x_k^* = f^*$$

which looks like (2.12.1) — the difference is that f^* and x_k^* are holomorphic functions in μ in the disk $|\mu| < 1/\varepsilon$. When $|\mu| < 1/\varepsilon$, the system $x_1^*, \dots, x_n^*$ is linearly independent since $x_1, \dots, x_n$ is linearly independent and $(I - \mu A_\varepsilon)$ is continuously invertible. So, by analogy with (2.12.1), we can point out that for $|\mu| < 1/\varepsilon$ the solution to (2.12.3) is

$$x = \frac{D(\mu) f^* + \sum_{k=1}^{n} D_k(\mu, f^*) x_k^*}{D(\mu)} \tag{2.12.4}$$

wherein all zeros of $D(\mu)$ have multiplicity no more than n.

If μ_0 is not an eigenvalue of A then the solution (2.12.4) is holomorphic in μ in a neighborhood of μ_0 and so $D(\mu_0) \neq 0$. But if μ_0 is an eigenvalue then $D(\mu_0) = 0$ since otherwise the equation would be solvable for all f^*, which is impossible.

So the set of eigenvalues of A belonging to the disk $|\mu| < 1/\varepsilon$ coincides with the set of zeros of $D(\mu)$ lying in this disk. □

2.13 Spectrum of Holomorphic Compact Operator Function

Let $A(\mu)$ be an operator-function whose value, for any $\mu \in G$, an open domain in $\mathbb{C}$, is a compact linear operator in a Hilbert space and $A(\mu)$ be holomorphic in G. We know that the spectrum of such operator-functions is a point spectrum. Following I.Ts. Gokhberg and M.G. Krein [11] we study the distribution of eigenvalues.

Lemma 2.13.1. For $\mu_0 \in G$ there is a positive ε such that for all μ in a domain $0 < |\mu - \mu_0| < \varepsilon$ the equation

$$(I - A(\mu))x = 0 \tag{2.13.1}$$

has the same number of linearly independent solutions.

Proof. Let $x_1, \dots, x_n$ be an orthonormal basis of the space of solutions of (2.13.1) when $\mu = \mu_0$. By Theorem 2.11.4 there is an orthonormal basis $y_1, \dots, y_n$ of the space of solutions of the equation $(I - A^*(\mu_0))x = 0$ and, by the proof of Theorem 2.11.4, the operator

$$Q(\mu_0)x = (I - A(\mu_0))x + \sum_{k=1}^{n} (x, x_k) y_k$$

has a continuous inverse. As $A(\mu)$ depends on μ continuously, there is some neighborhood $|\mu - \mu_0| < \rho$ of μ_0 in which $Q(\mu)$ has a continuous inverse.

Equation (2.13.1) is equivalent to

$$Q(\mu)x = \sum_{k=1}^{n} (x, x_k) y_k$$

and, in the above-mentioned neighborhood, to a system

$$x = \sum_{k=1}^{n} \xi_k Q^{-1}(\mu) y_k, \quad |\mu - \mu_0| < \rho,$$
$$\xi_k = (x, x_k), \quad k = 1, \ldots, n,$$

which, in turn, can be reduced to an equivalent system of algebraic equations (substituting x from the first equation into the others)

$$\xi_k - \sum_{j=1}^{n} (Q^{-1}(\mu) y_j, x_k) \xi_j = 0, \qquad k = 1, \ldots, n \tag{2.13.2}$$

whose number of linearly independent solutions coincides with the number for (2.13.1) when $|\mu - \mu_0| < \rho$. In this domain, all terms in (2.13.2) are holomorphic as are all elements of its determinant.

If all elements of the main determinant of the system (2.13.2) equal zero identically, then the system (2.13.2) has n linearly independent solutions in the disk $|\mu - \mu_0| < \rho$, and the lemma is proven.

Otherwise, let $\Delta_p(\mu)$ be a minor of highest order p which is nonzero at some point of the disk $|\mu - \mu_0| < \rho$. Being holomorphic, $\Delta_p(\mu)$ is nonzero in this disk except, perhaps, at a finite number of points. This means that in this disk, except at those points, the number of linearly independent solutions of (2.13.2) is $n - p$. Therefore, we can exhibit a disk $|\mu - \mu_0| < \varepsilon$ such that for all its points μ, except perhaps $\mu = \mu_0$, the system (2.13.2) and thus (2.13.1) has the same number $n - p$ of linearly independent solutions. □

Theorem 2.13.1. Assume $A(\mu)$ is an operator-function, being holomorphic on a connected open domain $G \subset \mathbb{C}$, whose values are compact linear operators in a Hilbert space. Then $\alpha(\mu)$, the number of linearly independent solutions of (2.13.1), is the same, $\alpha(\mu) = n$, for all points of G, except some isolated points of G at which $\alpha(\mu) > n$. In particular, if there exists $\mu_0 \in G$ such that $\alpha(\mu_0) = 0$, then the spectrum of $A(\mu)$ consists of isolated points of G. (This happens if, for example, there exists $\mu_0 \in G$ such that $A(\mu_0) = 0$.)

Proof. Consider $\alpha(\mu)$. Assume its minimal value is n and that it is taken at $\mu = \mu_0$. Let μ_1 be a point at which $\alpha(\mu_1) > n$. We shall show that

this point is isolated and, moreover, that there is an $\varepsilon > 0$ such that for all $\mu \neq \mu_1$, $|\mu - \mu_1| < \varepsilon$, the number $\alpha(\mu) = n$. Draw a curve lying in G from μ_0 to μ_1. By Lemma 2.13.1, for any $\mu^* \in G$ there is a positive number $\varepsilon(\mu^*)$ such that for any $\mu \in G$, $0 < |\mu - \mu^*| < \varepsilon(\mu^*)$, the number of linearly independent solutions of (2.13.1) is constant. These disks make up a covering of G from which we can choose a finite covering of G. As neighboring disks of the covering are mutually intersecting, then for all points of the disks of the finite covering, except perhaps for their centers, the number of linearly independent solutions of (2.13.1) is constant and equals n. □

2.14 Spectrum of Self-Adjoint Compact Linear Operator in Hilbert Space

We did not touch on the problem of existence of the spectrum. The example of the zero operator shows that there are compact operators having no finite spectral points. But there is a class of operators having eigenvalues; it is the class shown in the title of this section.

Lemma 2.14.1. All eigenvalues of a self-adjoint continuous linear operator A acting in a Hilbert space are real, as are all values of the form (Ax, x). Eigenvectors x_1, x_2 of A corresponding to μ_1, μ_2, respectively, $\mu_1 \neq \mu_2$, are mutually orthogonal; moreover, $(Ax_1, x_2) = 0$.

Proof. (Ax, x) is a real-valued functional since

$$(Ax, x) = (x, Ax) = \overline{(Ax, x)}.$$

So if $x_0 = \mu_0 A x_0$, then $(x_0, x_0) = \mu_0 (Ax_0, x_0)$ and μ_0 is real too.

Let $x_1 = \mu_1 A x_1$ and $x_2 = \mu_2 A x_2$. Multiply the terms of the first equation by x_2 from the right, and the terms of the second by x_1 from the left:

$$(x_1, x_2) = \mu_1 (Ax_1, x_2), \qquad (x_1, x_2) = (x_1, \mu_2 A x_2) = \mu_2 (Ax_1, x_2).$$

It follows that

$$(\mu_2 - \mu_1)(x_1, x_2) = 0$$

so $x_1 \perp x_2$ if $\mu_1 \neq \mu_2$. Returning to $(x_1, x_2) = \mu_1 (Ax_1, x_2)$, we get $(Ax_1, x_2) = 0$. Note that in the theory of elasticity the last equality is called the relation of generalized orthogonality of eigenvectors. □

Definition 2.14.1. A functional $F(x)$ given on a Hilbert space is called *weakly continuous* if for every weakly convergent sequence $\{x_n\}$, $x_n \rightharpoonup x_0$, we have $F(x_n) \to F(x_0)$.

Note that a continuous linear functional is weakly continuous by the definition of weak convergence of a sequence of elements.

Lemma 2.14.2. A real-valued weakly continuous functional $F(x)$ given on a Hilbert space takes its maximal and minimal values on a ball $\|x\| \le a$.

Proof. Let $\sup_{\|x\|\le a} F(x) = M$. Then there is a sequence $\{x_n\}$, $\|x_n\| \le a$, such that $F(x_n) \to M$ as $n \to \infty$. From $\{x_n\}$ we can choose a subsequence $\{x_{n_k}\}$ which is a weak Cauchy sequence and its weak limit x_0 satisfies $\|x_0\| \le a$. By Definition 2.14.1, $\lim_{n_k\to\infty} F(x_{n_k}) = F(x_0) = M$. The proof for the minimal point is similar. □

Lemma 2.14.3. Assume A is a self-adjoint compact linear operator in a Hilbert space. Then (Ax, x) is a real-valued weakly continuous functional on this space.

Proof. By Lemma 2.14.1, (Ax, x) is real-valued. Let $\{x_k\}$ be weakly convergent to x_0. Then

$$\begin{aligned}(Ax_k, x_k) - (Ax_0, x_0) &= (Ax_k, x_k) - (Ax_0, x_k) + (Ax_0, x_k) - (Ax_0, x_0)\\ &= (Ax_k - Ax_0, x_k) + (Ax_0, x_k - x_0)\\ &\to 0 \text{ as } k \to \infty\end{aligned}$$

since $\|Ax_k - Ax_0\| \to 0$ by compactness of A and $(Ax_0, x_k - x_0) \to 0$ as (x, Ax_0) is a continuous linear functional in H. □

Denote

$$\lambda_+ = \sup_{\|x\|\le 1} (Ax, x), \qquad \lambda_- = \inf_{\|x\|\le 1} (Ax, x).$$

Theorem 2.14.1. Assume $A \ne 0$ is a self-adjoint compact linear operator acting in a Hilbert space. Then there is at least one eigenvalue μ of A. If both of λ_+ and λ_- are not zero, then there are at least two eigenvalues of A which are $\mu_1 = 1/\lambda_+$ and $\mu_2 = 1/\lambda_-$.

Proof. Since $\|A\| = \sup_{\|x\|\le 1} |(Ax, x)|$, at least one of λ_+, λ_- is nonzero. Without loss of generality, assume that $\lambda_+ \ne 0$. By Lemmas 2.14.2 and 2.14.3, (Ax, x) takes this maximal value in the unit ball at an element x_0: $(Ax_0, x_0) = \lambda_+$. By homogeneity of (Ax, x) in x, it is evident that $\|x_0\| = 1$.

Now consider a functional

$$\Phi(x) = \frac{(Ax, x)}{\|x\|^2} = (A\xi, \xi), \qquad \xi = \frac{x}{\|x\|},$$

whose range of values coincides with the set of values of (Ax, x) when x runs over the sphere $\|x\| = 1$. Thus

$$\sup \Phi(x) = \sup_{\|x\|=1} (Ax, x) = (Ax_0, x_0) = \Phi(x_0).$$

We shall show that x_0 is an eigenvector of A. Consider $\Phi(x_0 + \alpha y)$, where y is an arbitrary but fixed element of H, as a function of a real variable α; it is differentiable in some neighborhood of $\alpha = 0$ and takes its maximal value at $\alpha = 0$ so

$$\frac{d\Phi(x_0 + \alpha y)}{d\alpha}\Big|_{\alpha=0} = 0.$$

This gives

$$\operatorname{Re}(Ax_0, y) - \frac{(Ax_0, x_0)}{\|x_0\|^2} \operatorname{Re}(x_0, y) = 0$$

or

$$\operatorname{Re}[(Ax_0, y) - \lambda_+(x_0, y)] = 0.$$

Replacing y by iy, we get

$$\operatorname{Im}[(Ax_0, y) - \lambda_+(x_0, y)] = 0$$

so

$$(Ax_0, y) - \lambda_+(x_0, y) = 0. \tag{2.14.1}$$

Since y is an arbitrary element of H, we have

$$Ax_0 - \lambda_+ x_0 = 0 \qquad \text{or} \qquad x_0 - \frac{1}{\lambda_+} Ax_0 = 0,$$

i.e., x_0 is an eigenvector of A. If $\lambda_- \neq 0$, then we can show similarly that $\mu = 1/\lambda_-$ is also an eigenvalue of A. □

Definition 2.14.2. A self-adjoint continuous linear operator A is called *strictly positive* if (1) $(Ax, x) \geq 0$ for all $x \in H$, and (2) $(Ax, x) = 0$ implies $x = 0$.

By Lemma 2.14.1, any two eigenvectors corresponding to different eigenvalues of a self-adjoint operator are mutually orthogonal. But we can orthonormalize a linearly independent system of eigenvectors corresponding to the same eigenvalue and so we can consider a basis of the set of all eigenvectors of a self-adjoint operator to be orthonormal. This and the method used in the proof of Theorem 2.14.1 allow us to prove

Theorem 2.14.2. Assume A is a strictly positive self-adjoint compact linear operator acting in a separable Hilbert space. Then

(i) A possesses infinitely many eigenvalues $\mu_1, \mu_2, \mu_3, \ldots$; the sequence of eigenvalues does not contain subsequences having finite limit points;

(ii) there is a system of eigenvectors $x_1, x_2, x_3, \ldots$ of A which is an orthonormal basis of H;

(iii) A has a representation

$$Ax = \sum_{k=1}^{\infty} \frac{(x, x_k)}{\mu_k} x_k, \qquad x_k - \mu_k A x_k = 0.$$

Proof. (i) Assume $x_1, x_2, \ldots, x_n$ is an orthonormal system of eigenvectors of A corresponding to eigenvalues $\mu_1, \mu_2, \ldots, \mu_n$ respectively. Some of μ_i and μ_j can coincide. We show how to construct the next eigen-pair (x_{n+1}, μ_{n+1}). Denote by H_n the orthogonal complement in H to a subspace spanned by $x_1, x_2, \ldots, x_n$. Considering A on H_n, we can repeat the proof of Theorem 2.14.1 and, denoting $\mu_{n+1} = 1/\lambda_{n+1}$, $\lambda_{n+1} = \sup_{\substack{\|x\| \le 1 \\ x \in H_n}} (Ax, x)$, get a vector x_{n+1} such that $x_{n+1} \in H_n$, $\|x_{n+1}\| = 1$, $\lambda_{n+1} = (Ax_{n+1}, x_{n+1})$; this vector x_{n+1} satisfies the equation (an analogy to (2.14.1))

$$(Ax_{n+1}, y) - \lambda_{n+1}(x_{n+1}, y) = 0 \text{ for all } y \in H_n.$$

In fact, this equality holds for all $y \in H$; if $y = x_k$, $k = 1, \ldots, n$, then

$$\begin{aligned}(Ax_{n+1} - \lambda_{n+1}x_{n+1}, x_k) &= (x_{n+1}, Ax_k) - \lambda_{n+1}(x_{n+1}, x_k) \\ &= (x_{n+1}, \lambda_k x_k) - \lambda_{n+1}(x_{n+1}, x_k) \\ &= 0\end{aligned}$$

because $(x_{n+1}, x_k) = 0$, so

$$Ax_{n+1} - \lambda_{n+1}x_{n+1} = 0.$$

Thus x_{n+1} is an eigenvector and $\mu_{n+1} = 1/\lambda_{n+1}$ an eigenvalue of A.

Now we can realize the process of successive construction of eigenvalues and eigenvectors of A, which could be disrupted only if $\sup_{\substack{\|x\| \le 1 \\ x \in H_n}} (Ax, x) = 0$ for some n. But this is possible only if H is a finite dimensional space. The remainder of statement (i) is evident.

(ii) Let y be an arbitrary element of H. Consider

$$y_n = y - \sum_{k=1}^{n} (y, x_k)x_k$$

where $\{x_k\}$ is a sequence of eigenvectors defined in (i) above. We recall that y_n is the nth Fourier remainder of y. It is clear that $y_n \in H_n$. In the theory of Fourier expansions (Section 1.22) it was shown that $\{y_n\}$ is a Cauchy sequence. Assume its strong limit is $y_0 \neq 0$. For y_n, thanks to $y_n \in H_n$, we get

$$\frac{(Ay_n, y_n)}{\|y_n\|^2} \le \lambda_{n+1}.$$

But $\lambda_n \to 0$ as $n \to \infty$ since the set $\{\mu_n\}$ has no finite limit points. Passage to the limit gives

$$\frac{(Ay_0, y_0)}{\|y_0\|^2} = 0$$

so $y_0 = 0$. This completes the proof of (ii).

(iii) In the proof of Theorem 2.7.2, we showed that the operator A is a uniform limit of a sequence of finite-dimensional operators A_n,

$$A_n x = \sum_{k=1}^{n} (x, g_k) A g_k,$$

where $g_1, g_2, \ldots, g_n, \ldots$ is an orthonormal basis of H. We take as a basis the set $x_1, x_2, \ldots$ constructed in part (i). Then

$$A_n x = \sum_{k=1}^{n} (x, x_k) A x_k = \sum_{k=1}^{n} \frac{(x, x_k)}{\mu_k} x_k$$

and so

$$Ax = \sum_{k=1}^{\infty} \frac{(x, x_k)}{\mu_k} x_k.$$

The proof is thereby completed. □

Let us note that under the conditions of Theorem 2.14.2, thanks to the Parseval equality, we get

$$\|x\|^2 = \sum_{k=1}^{\infty} |(x, x_k)|^2 = \sum_{k=1}^{\infty} |(x, \mu_k A x_k)|^2 = \sum_{k=1}^{\infty} \mu_k^2 |(x, A x_k)|^2$$

for any $x \in H$.

Since A is strictly positive, we can introduce a new norm

$$\|x\|_A = (Ax, x)^{1/2}$$

and the corresponding scalar product $(x, y)_A = (Ax, y)$. When H with this norm is incomplete, we introduce its completion H_A with respect to this norm.

Let $y_k = \sqrt{\mu_k} x_k$, x_k being an eigenvector of part (i).

Lemma 2.14.4. Under the conditions of Theorem 2.14.2, the set

$$\sqrt{\mu_1} x_1, \sqrt{\mu_2} x_2, \sqrt{\mu_3} x_3, \ldots$$

is an orthonormal basis of H_A.

Proof. The system $y_1, y_2, y_3, \ldots$ is orthonormal in H_A; indeed

$$(y_k, y_n)_A = (Ay_k, y_n) = \sqrt{\mu_k}\sqrt{\mu_n}(Ax_k, x_n) = \frac{\sqrt{\mu_k \mu_n}}{\mu_k}(x_k, x_n)$$

$$= \delta_{kn} = \begin{cases} 1, & k = n, \\ 0, & k \neq n. \end{cases}$$

For any $x \in H$ there holds the Parseval equality in H_A:

$$(x,x)_A = (Ax,x) = \left(A\sum_{k=1}^{\infty}(x,x_k)x_k, x\right) = \sum_{k=1}^{\infty}(x,x_k)(Ax_k,x)$$
$$= \sum_{k=1}^{\infty}(x,\mu_k Ax_k)(Ax_k,x) = \sum_{k=1}^{\infty}(x,Ay_k)(Ay_k,x)$$
$$= \sum_{k=1}^{\infty}|(Ax,y_k)|^2 = \sum_{k=1}^{\infty}|(x,y_k)_A|^2.$$

This means that the system $y_1, y_2, y_3, \ldots$ is an orthonormal basis of the set of elements $x \in H$ in H_A. But this set is dense in H_A, and that completes the proof. □

As an example, consider the eigenvalue problem

$$y'' + \mu^2 y = 0, \qquad y(0) = y(\pi) = 0,$$

with the well-known set of eigenfunctions $\left\{\frac{\sqrt{2}}{\pi}\sin kx\right\}$, $k = 1, 2, \ldots$. We are interested in its properties. If E is a Hilbert space of functions $y \in \dot{W}^{1,2}(0,\pi)$ with scalar product

$$(y,z)_E = \int_0^{\pi} y'(x)\overline{z'(x)}\,dx$$

then the problem can be posed as the problem of finding nontrivial y satisfying the equation

$$(y,z)_E - \mu^2(Ay,z)_E = 0$$

for any $z \in E$, where

$$(Ay,z)_E = \int_0^{\pi} y(x)\overline{z(x)}\,dx.$$

It was shown that A satisfies all conditions of Theorem 2.14.2, so the system $\left\{\frac{\sqrt{2}}{\pi}\sin kx\right\}$ is an orthonormal basis in $L^2(0,\pi)$, which is the space H_A from Lemma 2.14.4. Simultaneously, the same system is an orthogonal basis in $\dot{W}^{1,2}(0,\pi)$.

2.15 Some Applications of Spectral Theory

First we recall the spectrum for the different elastic bodies we considered. These are the problems of membranes, plates, and bodies in the framework of two- and three-dimensional linear elasticity. In generalized form, we have

$$(u,v)_E = \mu\int_{\Omega}\rho(\mathbf{x})u(\mathbf{x})\overline{v(\mathbf{x})}\,d\Omega. \tag{2.15.1}$$

Here E is an energy space for the corresponding elastic object occupying a bounded domain Ω, and $u(\mathbf{x})$ is a function for a membrane or a plate and a vector-function of displacements for an elastic body. The spaces E were introduced as real spaces — here we use their complex versions (i.e., the complex conjugate is applied to v in the integrand of $(u,v)_E$). Now we introduce an operator K using the Riesz representation theorem (Sections 1.20 and 2.6)

$$(Ku,v)_E = \int_\Omega \rho(\mathbf{x})u(\mathbf{x})\overline{v(\mathbf{x})}\,d\Omega.$$

For free-boundary problems E consists of "balanced" elements; for a membrane, say, they satisfy $\int_\Omega u(\mathbf{x},\mathbf{y})\,d\Omega = 0$. It was shown (Lemma 2.6.1) that K is a self-adjoint compact linear operator in an energy space. Moreover, if $\rho_0 \le \rho(\mathbf{x}) \le \rho_1$, with ρ_0, ρ_1 positive constants, then K is a strictly positive operator; indeed,

$$(Ku,u)_E = \int_\Omega \rho(\mathbf{x})|u(\mathbf{x})|^2\,d\Omega \ge \rho_0 \int_\Omega |u(\mathbf{x})|^2\,d\Omega$$

and if $(Ku,u)_E = 0$ then $u = 0$ in the sense of $L^2(\Omega)$ (almost everywhere).

So for any of the models of bounded elastic bodies that we considered, we get

Theorem 2.15.1. In the framework of all main (Dirichlet, Neumann, and mixed) spectral boundary value problems in the generalized statement for bounded membranes, plates, or linear elastic bodies, the spectrum of each problem contains only eigenvalue points μ_k, and:

(i) All μ_k are positive, $\mu_k \ge \mu_0 > 0$.

(ii) The set $\{\mu_k\}$ is infinite and does not contain a finite limit point.

(iii) To each μ_k there corresponds no more than a finite number of linearly independent eigenvectors which are assumed to be orthonormalized; the set of all these eigenvectors $\{u_k(\mathbf{x})\}$ is an orthonormal basis in the corresponding energy space and the set $\{\sqrt{\mu_k}u_k(\mathbf{x})\}$ is an orthonormal basis in $L^2(\Omega)$ with scalar product

$$(u,v)_{L^2(\Omega)} = \int_\Omega \rho(\mathbf{x})u(\mathbf{x})\overline{v(\mathbf{x})}\,d\Omega.$$

In Section 2.5 we considered a stability problem for a thin plate. In generalized terms this problem can be stated as

$$(w,\varphi)_{E_P} = \mu(Cw,\varphi)_{E_P} \tag{2.15.2}$$

where

$$(Cw,\varphi)_{E_P} = \int_\Omega \left[T_x \frac{\partial w}{\partial x}\overline{\frac{\partial \varphi}{\partial x}} + T_{xy}\left(\frac{\partial w}{\partial y}\overline{\frac{\partial \varphi}{\partial x}} + \frac{\partial w}{\partial x}\overline{\frac{\partial \varphi}{\partial y}}\right) + T_y \frac{\partial w}{\partial y}\overline{\frac{\partial \varphi}{\partial y}}\right] d\Omega$$

with given functions $T_x, T_{xy}, T_y \in L^2(\Omega)$. Then we get a spectral problem

$$w = \mu C w.$$

For a clamped plate, it was shown that C is a self-adjoint continuous operator. Since the imbedding operator from $\dot{W}^{2,2}(\Omega)$ to $W^{1,4}(\Omega)$ is compact, the inequality

$$\begin{aligned} |(Cw, \varphi)_{E_P}| \le m &\left[\int_\Omega (T_x^2 + T_{xy}^2 + T_y^2)\, d\Omega\right]^{1/2} \cdot \\ &\cdot \left[\int_\Omega \left(\left|\frac{\partial w}{\partial x}\right|^4 + \left|\frac{\partial w}{\partial y}\right|^4\right) d\Omega\right]^{1/4} \cdot \\ &\cdot \left[\int_\Omega \left(\left|\frac{\partial \varphi}{\partial x}\right|^4 + \left|\frac{\partial \varphi}{\partial y}\right|^4\right) d\Omega\right]^{1/4} \end{aligned}$$

shows that C is also compact.

Finally we assume the external tangential load to be compressible in total, which is expressed by the inequality

$$T_x w_1^2 + 2T_{xy} w_1 w_2 + T_y w_2^2 \ge c_0 (w_1^2 + w_2^2) \tag{2.15.3}$$

which is valid with a positive constant c_0 for all real w_1, w_2 and $\mathbf{x} \in \Omega$. Under the condition (2.15.3), C is a strictly positive operator on E_{PC}, and so we can say that this spectral problem is similar to one considered in Theorem 2.15.1. Thus its spectrum has properties as stated in Theorem 2.15.1.

We mentioned that the spectral theory of compact operators began with the study of integral equations originated by I. Fredholm. We found an integral operator to be compact in $L^2(\Omega)$ if its kernel belongs to $L^2(\Omega \times \Omega)$ and so we can reformulate all general results for these equations (try to do this yourself). Now we wish to consider another important class of integral operators, the so-called operators with kernels having weak singularities. These are kernels of the form

$$\mathcal{K}(\mathbf{x}, \mathbf{y}) = \frac{R(\mathbf{x}, \mathbf{y})}{r^\alpha}, \qquad r = |\mathbf{x} - \mathbf{y}|, \qquad \mathbf{x}, \mathbf{y} \in \Omega \subset \mathbb{R}^n$$

where $\alpha < n$ and $R(\mathbf{x}, \mathbf{y}) \in C(\Omega \times \Omega)$.

Lemma 2.15.1. A linear integral operator whose kernel has a weak singularity is compact in $L^2(\Omega)$.

Proof. First we show that this operator is bounded in $L^2(\Omega)$:

$$\begin{aligned} \|Au\|^2_{L^2(\Omega)} &= \int_\Omega \left| \int_\Omega \frac{R(\mathbf{x}, \mathbf{y})}{r^\alpha} u(\mathbf{y})\, d\Omega_{\mathbf{y}} \right|^2 d\Omega_{\mathbf{x}} \\ &\le m \int_\Omega \left(\int_\Omega \frac{1}{r^\alpha} d\Omega_{\mathbf{y}} \int_\Omega \frac{|u(\mathbf{y})|^2}{r^\alpha}\, d\Omega_{\mathbf{y}} \right) d\Omega_{\mathbf{x}}; \end{aligned}$$

since $\left|\int_\Omega \frac{d\Omega_{\mathbf{y}}}{r^\alpha}\right| \le M$ when $\mathbf{x} \in \Omega$, we further get

$$\|Au\|^2_{L^2(\Omega)} \le mM \int_\Omega \int_\Omega \frac{|u(\mathbf{y})|^2}{r^\alpha}\, d\Omega_{\mathbf{x}}\, d\Omega_{\mathbf{y}} \le mM^2 \int_\Omega |u(\mathbf{y})|^2\, d\Omega_{\mathbf{y}},$$

which demonstrates the continuity of A.

To show that A is compact, we introduce an auxiliary operator with kernel

$$\mathcal{K}_\varepsilon(\mathbf{x},\mathbf{y}) = \frac{R(\mathbf{x},\mathbf{y})}{r_\varepsilon^\alpha}, \qquad r_\varepsilon = \begin{cases} r, & r = |\mathbf{x}-\mathbf{y}| \ge \varepsilon, \\ \varepsilon, & r = |\mathbf{x}-\mathbf{y}| < \varepsilon. \end{cases}$$

The corresponding integral operator A_ε is compact because its kernel is continuous on $\Omega \times \Omega$. Now it suffices to prove that $\|A - A_\varepsilon\| \to 0$ as $\varepsilon \to 0$. Denote by $B(\mathbf{x})$ a ball about $\mathbf{x}$ of radius $\varepsilon > 0$. We have

$$\begin{aligned}
\|Au - A_\varepsilon u\|^2_{L^2(\Omega)} &= \int_\Omega \left| \int_\Omega R(\mathbf{x},\mathbf{y}) \left(\frac{1}{r^\alpha} - \frac{1}{r_\varepsilon^\alpha}\right) u(\mathbf{y})\, d\Omega_{\mathbf{y}} \right|^2 d\Omega_{\mathbf{x}} \\
&= \int_\Omega \left| \int_{\Omega \cap B(\mathbf{x})} R(\mathbf{x},\mathbf{y}) \left(\frac{1}{r^\alpha} - \frac{1}{r_\varepsilon^\alpha}\right) u(\mathbf{y})\, d\Omega_{\mathbf{y}} \right|^2 d\Omega_{\mathbf{x}} \\
&\le 4 \int_\Omega \left(\int_{\Omega \cap B(\mathbf{x})} \frac{|R(\mathbf{x},\mathbf{y})|}{r^\alpha} |u(\mathbf{y})|\, d\Omega_{\mathbf{y}} \right)^2 d\Omega_{\mathbf{x}} \\
&\le 4m \int_\Omega \left(\int_{\Omega \cap B(\mathbf{x})} \frac{d\Omega_{\mathbf{y}}}{r^\alpha} \int_{\Omega \cap B(\mathbf{x})} \frac{|u(\mathbf{y})|^2}{r^\alpha}\, d\Omega_{\mathbf{y}} \right) d\Omega_{\mathbf{x}}.
\end{aligned}$$

Since

$$\int_{\Omega \cap B(\mathbf{x})} \frac{d\Omega_{\mathbf{y}}}{r^\alpha} \le m_1 \varepsilon^{n-\alpha},$$

we have

$$\|Au - A_\varepsilon u\|^2_{L^2(\Omega)} \le m_2 \varepsilon^{n-\alpha} \|u\|^2_{L^2(\Omega)}$$

and thus $\|A - A_\varepsilon\| \to 0$ as $\varepsilon \to 0$. □

Note that integral operators with weakly singular kernels appear in the theory of Sobolev spaces: they participate in the integral representation of functions, and their properties led Sobolev to his famous imbedding theorems.

We leave it to the reader to formulate properties of the spectrum of a linear integral operator with kernel having weak singularity.

2.16 Courant's Minimax Principle

R. Courant proposed a way to determine the nth eigenvalue of a strictly positive self-adjoint compact linear operator A, by which this eigenvalue could be found independently of the other eigenvalues.

In Section 2.14 we have shown that μ_n given by

$$\mu_n = \frac{1}{\lambda_n}, \qquad \lambda_n = \sup_{\substack{\|u\|\le 1 \\ u\in H_n}} (Ax, x),$$

is the nth eigenvalue of A determined successively: $0 < \mu_1 \le \mu_2 \le \mu_3 \le \cdots$. As the corresponding orthonormal system of eigenvectors $x_1, x_2, x_3, \ldots$ is a basis of H, an element $x \in H$ can be represented as

$$x = \sum_{k=1}^{\infty} c_k x_k, \qquad \|x\|^2 = \sum_{k=1}^{\infty} |c_k|^2,$$

and thus

$$\begin{aligned}(Ax, x) &= \left(A\sum_{k=1}^{\infty} c_k x_k, \sum_{k=1}^{\infty} c_k x_k\right) = \left(\sum_{k=1}^{\infty} c_k \lambda_k x_k, \sum_{k=1}^{\infty} c_k x_k\right) \\ &= \sum_{k=1}^{\infty} \lambda_k |c_k|^2.\end{aligned}$$

Let us take any n elements $y_1, \ldots, y_n$ of H and denote by Q_n the subspace spanned by these elements and by S_n its orthogonal complement in H. Let $Q_{n_{\text{eig}}}$ be the subspace of H spanned by the eigenvectors $x_1, x_2, \ldots, x_n$ of A that were determined in Section 2.14, and H_n the orthogonal complement of $Q_{n_{\text{eig}}}$ in H. We shall now prove the so-called minimax principle of Courant, which is

Theorem 2.16.1. μ_{n+1} of A is

$$\mu_{n+1} = \frac{1}{\lambda_{n+1}}, \qquad \lambda_{n+1} = \inf_{Q_n} \sup_{\substack{\|x\|\le 1 \\ x\in S_n}} (Ax, x) \tag{2.16.1}$$

and $\lambda_{n+1} = \sup_{\|x\|\le 1}(Ax, x)$ when $x \in H_n$.

Proof. We first recall that we have shown (Section 2.14) that

$$\mu_{n+1} = \frac{1}{\lambda_{n+1}}, \qquad \lambda_{n+1} = \sup_{\substack{\|x\|\le 1 \\ x\in H_n}} (Ax, x)$$

is an eigenvalue of A. Then we recall that $\lambda_n \ge \lambda_{n+1}$ for all $n \ge 1$.

Suppose $Q_n \neq Q_{n_{\text{eig}}}$. Then there is $x_0 = \sum_{k=1}^{n} c_k^0 x_k \in Q_{n_{\text{eig}}}$ such that $\|x_0\| = 1$, $x_0 \in S_n$, and

$$(Ax_0, x_0) = \sum_{k=1}^{n} \lambda_k |c_k^0|^2 \geq \lambda_n \sum_{k=1}^{n} |c_k^0|^2 = \lambda_n \geq \lambda_{n+1}$$

and so for any S_n

$$\sup_{\substack{\|x\| \leq 1 \\ x \in S_n}} (Ax, x) \geq \lambda_{n+1}.$$

This completes the proof. □

Some consequences of this principle are very important in mechanics.

Theorem 2.16.2. If an elastic body (a membrane, plate, or two- or three-dimensional body) is subjected to some additional geometrical constraints (fixed lines, surfaces, or their parts), then all corresponding eigenvalues μ_n can only grow; if some geometrical constraints are broken, then all μ_n can only be less than or equal to the original ones.

Proof. Additional constraints are assumed to be some linear restrictions of the type, say,

$$u\big|_{\gamma} = 0 \qquad \text{or} \qquad u\big|_{\Gamma} = 0$$

where γ, Γ are a line and a surface, respectively. The value (2.16.1) of the supremum of (Ax, x) in the definition of λ_n becomes less or the same and so this holds for their infimum, i.e., λ_n cannot be more than the old one, and so μ_n cannot be less under new restrictions. The second part of the statement of the theorem is now evident. □

Remark 2.16.1. In the energy spaces E_M and E_E for membranes and elastic bodies, a restriction

$$u\big|_{P} = 0$$

where P is a point in Ω, in accordance with Sobolev's imbedding theorems, is neglected, and so the fixing of several points of a membrane or an elastic body cannot increase the corresponding eigenfrequencies (this was shown by Vitt and Shubin [24]); but for eigenfrequencies of a plate, fixing of finite number of points of the plate increases eigenfrequencies.

3

Elements of Nonlinear Functional Analysis

From the viewpoint of functional analysis, nonlinear problems of mechanics are more complicated than linear problems; as in mechanics, they require new techniques for their study. Many of them, such as nonlinear elasticity in the general case, provide a wide field of investigation for mathematicians (see Antman [2]); the problem of existence of solutions in nonlinear elasticity in general is still open.

But some of the nonlinear problems of mechanics can be treated on a known background; as in the linear case, we consider only some of the known nonlinear results of functional analysis that are needed in what follows.

3.1 Fréchet and Gâteaux Derivatives

We begin nonlinear analysis of operators with definitions of differentiation. Let $F(x)$ be a nonlinear operator acting from $D(F) \subset X$ to $R(F) \subset Y$, where X and Y are real Banach spaces. Assume $D(F)$ is open.

Definition 3.1.1. $F(x)$ is *differentiable in the Fréchet sense* at $x_0 \in D(F)$ if there is a bounded linear operator, denoted by $F'(x_0)$, such that

$$F(x_0 + h) - F(x_0) = F'(x_0)h + \omega(x_0, h) \text{ for all } \|h\| < \varepsilon$$

with some $\varepsilon > 0$, where $\|\omega(x_0, h)\|/\|h\| \to 0$ as $\|h\| \to 0$. Then $F'(x_0)$ is called the *Fréchet derivative* of $F(x)$ at x_0, and $dF(x_0, h) = F'(x_0)h$ is

its *Fréchet differential*. $F(x)$ is Fréchet differentiable in an open domain $S \subset D(F)$ if it is Fréchet differentiable at every point of S.

It is clear that the Fréchet derivative of a continuous linear operator is the same operator.

Problem 3.1.1. Assume $\mathbf{y} = \mathbf{f}(\mathbf{x})$ is a vector function from $\mathbb{R}^m$ to $\mathbb{R}^n$ and $\mathbf{f}(\mathbf{x}) \in (C^{(1)}(\Omega))^n$. Show that its Fréchet derivative at $\mathbf{x}_0 \in \Omega$ is the Jacobi matrix $\left(\frac{\partial f_i(\mathbf{x}_0)}{\partial x_j}\right)_{\substack{i=1,\dots,n \\ j=1,\dots,m}}$.

In the construction of the Fréchet derivative, the reader can recognize a method of the calculus of variations, used to obtain the Euler equations of a functional. The following derivative by Gâteaux is yet closer to this.

Definition 3.1.2. Assume that for all $h \in D(F)$ we have

$$\lim_{t \to 0} \frac{F(x_0 + th) - F(x_0)}{t} = DF(x_0, h), \qquad x_0 \in D(F),$$

where $DF(x_0, h)$ is a linear operator with respect to h. Then $DF(x_0, h)$ is called the *Gâteaux differential* of $F(x)$ at x_0, and the operator is called *Gâteaux differentiable*. Denoting $DF(x_0, h) = F'(x_0)h$, we get the *Gâteaux derivative* $F'(x_0)$. An operator is differentiable in the Gâteaux sense in an open domain $S \subset X$ if it has a Gâteaux derivative at every point of S.

The definitions of derivatives are clearly valid for functionals. Suppose $\Phi(x)$ is a functional which is Gâteaux differentiable in a Hilbert space and that $D\Phi(x, h)$ is bounded at $x = x_0$ as a linear functional in h. Then, by the Riesz representation theorem, it can be represented in the form of an inner product; denoting the representing element by $\operatorname{grad} \Phi(x_0)$, we get

$$D\Phi(x_0, h) = (\operatorname{grad} \Phi(x_0), h).$$

By this, we have an operator $\operatorname{grad} \Phi(x_0)$ called the *gradient* of $\Phi(x)$ at x_0.

Theorem 3.1.1. If an operator $F(x)$ from X to Y is Fréchet differentiable at $x_0 \in D(F)$, then $F(x)$ is Gâteaux differentiable at x_0 and the Gâteaux derivative coincides with the Fréchet derivative.

Proof. Put th instead of h in Definition 3.1.1:

$$F(x_0 + th) - F(x_0) = F'(x_0)th + \omega(x_0, th).$$

It follows that

$$\lim_{t \to 0} \frac{F(x_0 + th) - F(x_0)}{t} = F'(x_0)h$$

since $\|\omega(x_0, th)\|/\|th\| \to 0$ as $t \to 0$. This means $F'(x_0)$ is a Gâteaux derivative of $F(x)$ at x_0. □

Gâteaux differentiability does not imply Fréchet differentiability. We formulate a sufficient condition as

Problem 3.1.2. Assume that the Gâteaux derivative of $F(x)$ exists in a neighborhood of x_0 and is continuous at x_0 in the uniform norm of $L(X, Y)$. Show that the Fréchet derivative exists and is equal to the Gâteaux derivative.

We consider an operator equation with a parameter μ being an element of a real Banach space M:

$$F(x, \mu) = 0$$

where $D(F(x, \mu)) \subseteq X$, $R(F(x, \mu)) \subseteq Y$.

In problems of mechanics, μ can represent loads or some parameters of a body or a process (say, disturbances of the thickness of a plate or its moduli).

There are different abstract analogs of the implicit function theorem; we present two of them.

Denote by $N(x_0, r; \mu_0, \rho)$ the following neighborhood of a pair:

$$N(x_0, r; \mu_0, \rho) = \{x \in X, \mu \in M \mid \|x - x_0\| < r,\ \|\mu - \mu_0\| < \rho\}.$$

Theorem 3.1.2. Assume:

(i) $F(x_0, \mu_0) = 0$;

(ii) $F(x_0, \mu)$ is continuous with respect to μ in a ball $\|\mu - \mu_0\| < \rho_1$;

(iii) there exist $r_1 > 0$ and $\rho_1 > 0$ and a continuous linear operator A from X to Y, being continuously invertible and such that in the neighborhood $N(x_0, r_1; \mu_0, \rho_1)$

$$\|F(x, \mu) - F(y, \mu) - A(x - y)\| \leq \alpha(r_1, \rho_1)\|x - y\|$$

where $\limsup_{r,\rho \to 0} |\alpha(r, \rho)| \, \|A^{-1}\| = q < 1$.

Then there exist $r_0 > 0$ and $\rho_0 > 0$ such that in $N(x_0, r_0; \mu_0, \rho_0)$ the equation

$$F(x, \mu) = 0 \tag{3.1.1}$$

has the unique solution $x = x(\mu)$ which depends continuously on μ: $x(\mu) \to x(\mu_0)$ as $\mu \to \mu_0$.

Proof. We reduce the equation to a form needed to apply the contraction mapping principle:

$$x = K(x, \mu), \qquad K(x, \mu) = x - A^{-1}F(x, \mu).$$

This equation is equivalent to (3.1.1) because A^{-1} is continuously invertible. $K(x,\mu)$ is a contraction operator with respect to x in some neighborhood of (μ_0, x_0). Indeed

$$\begin{aligned}\|K(x,\mu)-K(y,\mu)\| &= \|x-y-A^{-1}(F(x,\mu)-F(y,\mu))\| \\ &\le \|A^{-1}\|\,\|A(x-y)-(F(x,\mu)-F(y,\mu))\| \\ &\le \|A^{-1}\|\,|\alpha(r,\rho)|\,\|x-y\| \\ &\le (q+\varepsilon)\|x-y\|;\end{aligned}$$

by (iii), $q+\varepsilon<1$ if r and ρ are sufficiently small and $r<r_1$, $\rho<\rho_1$. Then there are r_0, ρ_0, $r_0 \le r_1$, $\rho_0 \le \rho_1$, such that $K(x,\mu)$ takes a ball $\|x-x_0\| \le r_0$ into itself when $\|\mu-\mu_0\| \le \rho_0$, indeed

$$\begin{aligned}\|K(x,\mu)-x_0\| &\le \|K(x,\mu)-K(x_0,\mu)\| + \|K(x_0,\mu)-x_0\| \\ &\le (q+\varepsilon)\|x-x_0\| + \|A^{-1}F(x_0,\mu)\| \\ &\le (q+\varepsilon)\|x-x_0\| + \|A^{-1}\|\,\|F(x_0,\mu)\|.\end{aligned}$$

Since $F(x_0,\mu) \to F(x_0,\mu_0)=0$ as $\mu\to\mu_0$, then

$$\|A^{-1}\|\,\|F(x_0,\mu)\| \le (1-q-\varepsilon)r_1 \quad \text{when} \quad \|\mu-\mu_0\| \le \rho_2 \text{ for some } \rho_2<\rho_1$$

and thus for any $r_0<r_1$, $\rho_0<\rho_2$, the ball $\|x-x_0\| \le r_0$ is taken by $K(x,\mu)$ into itself when $\|\mu-\mu_0\| \le \rho_0$.

By the contraction mapping principle, there is a solution $x=x(\mu)$ in $N(x_0, r_0; \mu_0, \rho_0)$. The continuity of $x(\mu)$ at μ_0 follows from the bound

$$\|x(\mu)-x_0\| \le \frac{\|A^{-1}\|}{1-q-\varepsilon}\|F(x_0,\mu)\|,$$

a consequence of the contraction mapping principle. □

To prove the other variant of the implicit function theorem, we need some properties of Fréchet derivatives as given by the next two lemmas.

Lemma 3.1.1. Assume an operator $F(x)$ from X to Y has a Fréchet derivative at $x=x_0$, and an operator $x=S(z)$ from a real Banach space Z to X also has a Fréchet derivative $S'(z_0)$ and $x_0=S(z_0)$. Then their composition $F(S(z))$ has a Fréchet derivative at $z=z_0$ and

$$(F(S(z_0)))' = F'(x_0)S'(z_0).$$

Proof. Substituting

$$x-x_0 = S(z)-S(z_0) = S'(z_0)(z-z_0)+\omega_1(z_0, z-z_0)$$

into

$$F(x)-F(x_0) = F'(x_0)(x-x_0)+\omega(x_0, x-x_0),$$

we get

$$F(x) - F(x_0) = F'(x_0)S'(z_0)(z - z_0) + F'(x_0)\omega_1(z_0, z - z_0) + \\ + \omega(x_0, S(z) - S(z_0)).$$

This completes the proof, since the last two terms on the right-hand side are of the order $o(\|z - z_0\|)$. □

The next lemma is the so-called *Lagrange identity*.

Lemma 3.1.2. Assume that $F(x)$ from X to Y is Fréchet differentiable in a neighborhood Ω of x_0. Then for $x \in \Omega$ we have

$$F(x) - F(x_0) = \int_0^1 F'(x_0 + \theta(x - x_0))\, d\theta\, (x - x_0).$$

Proof. By Lemma 3.1.1, the composition $F(S(\theta))$, where $S(\theta) = x_0 + \theta(x - x_0)$, has a Fréchet derivative

$$\frac{d}{d\theta} F(S(\theta)) = F'(x_0 + \theta(x - x_0))(x - x_0)$$

since $S'(\theta) = x - x_0$. Integrating this over $[0, 1]$ with regard for continuity of $F(S(\theta))$ in θ, we complete the proof. □

We can now present the more traditional version of the implicit function theorem. In preparation we introduce a partial Fréchet derivative $F_x(x, \mu)$ of $F(x, \mu)$ with respect to x as its Fréchet derivative with respect to x when μ is fixed.

Theorem 3.1.3. Assume:

(i) $F(x_0, \mu_0) = 0$;

(ii) for some $r > 0$ and $\rho > 0$, the operator $F(x, \mu)$ is continuous on the set $N(x_0, r; \mu_0, \rho)$;

(iii) $F_x(x, \mu)$ is continuous at (x_0, μ_0);

(iv) $F_x(x_0, \mu_0)$ has a continuous inverse linear operator.

Then there exist $r_0 > 0$, $\rho_0 > 0$ such that the equation $F(x, \mu) = 0$ has the unique solution $x = x(\mu)$ in a ball $\|x - x_0\| \le r_0$ when $\|\mu - \mu_0\| \le \rho_0$. If there is, in addition, $F_\mu(x, \mu)$ which is continuous at (x_0, μ_0) then $x(\mu)$ has a Fréchet derivative at $\mu = \mu_0$ and

$$x'(\mu_0) = -F_x^{-1}(x_0, \mu_0) F_\mu(x_0, \mu_0).$$

Proof. We verify that $A = F_x(x_0, \mu_0)$ meets condition (iii) of Theorem 3.1.2. Consider

$$\Psi(x, y, \mu) = \|F(x, \mu) - F(y, \mu) - F_x(x, \mu_0)(x - y)\|.$$

By Lemma 3.1.2,

$$F(x, \mu) - F(y, \mu) = \int_0^1 F_x(y + \theta(x - y), \mu)\, d\theta\, (x - y)$$

and so

$$\begin{aligned}\Psi(x, y, \mu) &= \left\| \int_0^1 (F_x(y + \theta(x - y), \mu) - F_x(x_0, \mu_0))\, d\,\theta(x - y) \right\| \\ &\le \int_0^1 \|F_x(y + \theta(x - y), \mu) - F_x(x_0, \mu_0)\|\, d\theta\, \|x - y\| \\ &\le \alpha(r, \rho)\|x - y\|\end{aligned}$$

where

$$\alpha(r, \rho) = \sup_{x,\mu} \|F_x(x, \mu) - F_x(x_0, \mu_0)\| \quad \text{ on } N(x_0, r; \mu_0, \rho)$$

is such that $\alpha(r, \rho) \to 0$ as $r, \rho \to 0$ since $F_x(x, \mu)$ is continuous at (x_0, μ_0). The other conditions of Theorem 3.1.2 are also satisfied and so a solution $x = x(\mu)$ actually exists. We leave the second part of the theorem on differentiability of $x(\mu)$ without proof. □

Using the implicit function theorem, we can determine whether a solution to a problem depends continuously and uniquely on some parameters.

We studied several linear problems of mechanics with constant parameters. The reader can now verify that small disturbances of elastic moduli or, say, the thickness of a plate, bring small disturbances in displacements (small in a corresponding energy norm). We note that for linear problems this can be shown more easily by using the contraction mapping principle, but in nonlinear problems using the implicit function theorem is more convenient.

3.2 Liapunov–Schmidt Method

We shall say that (x_0, μ_0) is a *regular point* of the equation $F(x, \mu) = 0$ if there is a neighborhood of (x_0, μ_0), say $N(x_0, r; \mu_0, \rho)$, in which there is a unique solution $x = x(\mu)$.

The implicit function theorem gives sufficient conditions for regularity of $F(x, \mu)$ at (x_0, μ_0).

In mechanics, the breakdown of the property of regularity of a solution is of great importance; it is usually connected with some qualitative change of the properties of a system under consideration: its behavior, stability, or type of motion.

We now consider an important class of non-regular points of an operator equation.

Definition 3.2.1. (x_0, μ_0) is a *bifurcation point* of the equation $F(x, \mu) = 0$ if for any $r > 0$, $\rho > 0$, in the ball $\|\mu - \mu_0\| \le \rho$ there exists μ such that in the ball $\|x - x_0\| \le r$ there are at least two solutions of the equation corresponding to μ.

Many problems of mechanics (in particular, in shell theory) are such that in an energy space a partial Fréchet derivative $F_x(x_0, \mu_0)$ of a corresponding operator of a problem may be reduced to the form $I - B$, $B = B(x_0, \mu_0)$, where B is a compact linear operator (as a rule it is self-adjoint) and so the results of the Fredholm–Riesz–Schauder theory are valid. In particular, $I - B$ is not continuously invertible if and only if there is a nontrivial solution to $(I - B)x = 0$, and this is the case when the implicit function theorem is not applicable. This case is now considered.

Without loss of generality, we assume $x_0 = 0$, $\mu_0 = 0$ (we can always change $x \mapsto x_0 + x$, $\mu \mapsto \mu_0 + \mu$) so let

$$F(0, 0) = 0.$$

Suppose F is an operator acting from $H \times M$ in H where H is a Hilbert space and M is a real Banach space. As we said, we suppose that $F_x(0, 0)$ takes the form

$$F_x(0, 0) = I - B_0$$

with B_0 a compact self-adjoint linear operator in H.

The equation $F(x, \mu) = 0$ can be rewritten in the form

$$(I - B_0)x = -F(x, \mu) + (I - B_0)x$$

or

$$(I - B_0)x = R(x, \mu), \qquad R(x, \mu) = -F(x, \mu) + (I - B_0)x. \tag{3.2.1}$$

We now consider the Liapunov–Schmidt method of determining the dependence of solution to (3.2.1) on μ when $\|\mu\|$ is small and there are nontrivial solutions to the equation $(I - B_0)x = 0$. As in Section 2.11, denote by N the set of these nontrivial solutions and let $x_1, \ldots, x_n$ be an orthonormal basis of N.

In the beginning of the proof of Theorem 2.11.4 we saw that the operator

$$Q_0 x = (I - B_0)x + \sum_{k=1}^{n} (x, x_k)x_k$$

is continuously invertible. Equation (3.2.1) can be written in the form

$$Q_0 x = R(x,\mu) + \sum_{k=1}^{n} \alpha_k x_k, \qquad \alpha_k = (x, x_k). \tag{3.2.2}$$

We now consider (3.2.2) as an equation with respect to x that has parameters $\mu, \alpha_1, \ldots, \alpha_n$, introducing, in preparation,

$$x = u + \sum_{k=1}^{n} \beta_k x_k, \qquad (u, x_j) = 0, \quad j = 1, \ldots, n.$$

Here $u \in M$, M being the orthogonal complement of N in H. As $(x, x_k) = \alpha_k$, then $x = u + \sum_{k=1}^{n} \alpha_k x_k$ and (3.2.2) is

$$Q_0 u = R\left(u + \sum_{k=1}^{n} \alpha_k x_k, \mu\right). \tag{3.2.3}$$

This equation defines u as a function of the variables $\mu, \alpha_1, \ldots, \alpha_n$. Since $R_x(0,0) = -F_x(0,0) + (I - B_0) = 0$ we get

$$\left(Q_0 x - R(u + \sum_{k=1}^{n} \alpha_k x_k, \mu)\right)_u \Big|_{\substack{u=0 \\ \mu=0, \alpha_1=\cdots=\alpha_n=0}} = Q_0$$

where Q_0 is a continuously invertible operator, so all the conditions of the implicit function theorem are fulfilled. Therefore (3.2.3) has a unique solution for every μ, $\alpha_1, \ldots, \alpha_n$ when $\|\mu\|$ and $|\alpha_k|$ are small:

$$u = u(\mu, \alpha_1, \ldots, \alpha_n).$$

This solution must be orthogonal to all x_k, $k = 1, \ldots, n$, and to define values $\alpha_1, \ldots, \alpha_n$ we have the system

$$(u(\mu, \alpha_1, \ldots, \alpha_n), x_k) = 0, \qquad k = 1, \ldots, n \tag{3.2.4}$$

which is called the Liapunov–Schmidt equation of branching.

Using the Liapunov–Schmidt method one can investigate so-called post-critical behavior of a system, say, post-buckling of a von Kármán plate.

3.3 Critical Points of a Functional

From now on, we shall consider operators and real-valued functionals given in a real Hilbert space H. So let $\Phi(x)$ be a functional on H.

Definition 3.3.1. $x_0 \in H$ is called a *local minimal* (*maximal*) point of $\Phi(x)$ if there is a ball $B = \{x \mid \|x - x_0\| \le \varepsilon\}$, $\varepsilon > 0$, such that for all $x \in B$ we have $\Phi(x) \ge \Phi(x_0)$ ($\Phi(x) \le \Phi(x_0)$). Minimal and maximal points are called *extreme points* of $\Phi(x)$. If $\Phi(x) \ge \Phi(x_0)$ for all $x \in H$, then x_0 is a point of *absolute minimum* of $\Phi(x)$.

We prove the following

Theorem 3.3.1. Assume:

(i) $\Phi(x)$ is given on an open set $S \subset H$;

(ii) there exists $\operatorname{grad} \Phi(x)$ at $x = x_0 \in S$;

(iii) x_0 is an extreme point of $\Phi(x)$.

Then $\operatorname{grad} \Phi(x_0) = 0$.

Proof. Let h be an arbitrary element of H. The functional $\Phi(x_0 + th)$ is a function in a real variable t that attains its minimum at $t = 0$. Since

$$\frac{d\Phi(x_0 + th)}{dt}\Big|_{t=0} = 0,$$

we have

$$(\operatorname{grad} \Phi(x_0), h) = 0. \tag{3.3.1}$$

Since h is arbitrary, the conclusion follows. □

Definition 3.3.2. A point x_0 at which $\operatorname{grad} \Phi(x_0) = 0$ is called a *critical point* of $\Phi(x)$.

In fact, we implicitly used this theorem for linear problems when $\Phi(x)$ was a (quadratic) functional of total energy of an elastic body and (3.3.1) was an equation defining a generalized solution of the corresponding problem. Similar results will be valid for some nonlinear problems in what follows.

In preparation, we introduce some definitions.

Definition 3.3.3. A functional $\Phi(x)$ is called *weakly continuous* at $x = x_0$ if for every sequence $\{x_k\}$ converging weakly to x_0 the numerical sequence $\Phi(x_k)$ tends to $\Phi(x_0)$ as $k \to \infty$. It is called weakly continuous on an open set $S \subset H$ if it is weakly continuous at every point of S.

Definition 3.3.4. A functional $\Phi(x)$ given on H is called *growing* if

$$\inf_{\|x\|=R} \Phi(x) \to \infty \text{ as } R \to \infty.$$

We obtained a necessary condition for existence of critical points of a functional. Now we point out some sufficient conditions for this that have important applications in mechanics.

Lemma 3.3.1. Assume Q is a weakly closed and bounded set in H. A weakly continuous functional $\Phi(x)$ is bounded on Q and attains its minimal and maximal values in it.

Proof. First we prove that the values of $\Phi(x)$ on Q are bounded from above. If not, there is a sequence $\{x_n\} \subset Q$ such that $\Phi(x_n) \to \infty$ as $n \to \infty$. By hypothesis $\{x_n\}$ contains a subsequence $\{x_{n_k}\}$ weakly convergent to $x_0 \in Q$ and so

$$\Phi(x_{n_k}) \to \Phi(x_0) \neq \infty \text{ as } n_k \to \infty,$$

which contradicts the assumption. Boundedness from below is thus clearly seen.

Let $d = \inf_{x \in Q} \Phi(x)$. By definition of infimum there is a sequence $\{z_n\}$ for which $\Phi(z_n) \to d$ as $n \to \infty$. As above, it contains a subsequence $\{z_{n_k}\}$ converging weakly to $z_0 \in Q$. By weak continuity of $\Phi(x)$ we get $\Phi(z_0) = d$. The proof for the maximal value is similar. □

Note that a ball $B(R) = \{x \mid \|x\| \leq R\}$ has the properties of Q of the lemma.

In what follows, some problems of mechanics can be reduced to a problem of finding critical points of the functional

$$\Psi(x) = \|x\|^2 + \Phi(x)$$

with $\Phi(x)$ a weakly continuous functional. The functional $\Psi(x)$ is not weakly continuous because of the term $\|x\|^2$ and so Lemma 3.3.1 does not apply.

Theorem 3.3.2. Let $\Phi(x)$ be a weakly continuous functional. On a ball $B(R) = \{x \mid \|x\| \leq R\}$, the functional $\Psi(x) = \|x\|^2 + \Phi(x)$ attains its minimal value.

Proof. By Lemma 3.3.1, $\Phi(x)$ and hence $\Psi(x)$ is bounded from below on $B(R)$. Let $d = \inf \Psi(x)$ on $B(R)$ and $\{x_n\}$ be a sequence in $B(R)$ such that $\Psi(x_n) \to d$ as $n \to \infty$. By weak compactness of $B(R)$ we can produce a subsequence $\{x_{n_k}\}$ which converges weakly to $x_0 \in B(R)$. Moreover, from the bounded numerical sequence $\{\|x_{n_k}\|\}$ we can take a subsequence which tends to some number a, $a \leq R$. Redenote the last subsequence as $\{x_n\}$ again.

We show that $\|x_0\| \leq a$. Indeed, since $x_n \rightharpoonup x_0$ then $\lim_{n\to\infty}(x_n, x_0) = \|x_0\|^2$ and we have

$$\|x_0\|^2 = \lim_{n\to\infty} |(x_n, x_0)| \leq \lim_{n\to\infty} \|x_n\| \, \|x_0\| = a\|x_0\|$$

which gives $\|x_0\| \leq a$.

By weak continuity of $\Phi(x)$, we get $\Phi(x_n) \to \Phi(x_0)$ as $n \to \infty$ and $\Psi(x_n) \to d = a^2 + \Phi(x_0)$ simultaneously. Since $x_0 \in B(R)$,

$$\Psi(x_0) = \|x_0\|^2 + \Phi(x_0) \geq \inf_{x \in B(R)} \Psi(x) = d = a^2 + \Phi(x_0),$$

and so $\|x_0\| \geq a$. With the above, this implies $\|x_0\| = a$ and thus x_0 is a point at which $\Psi(x)$ takes its minimal value on $B(R)$. □

Remark 3.3.1. Since $\{x_{n_k}\}$ from the proof converges weakly to x_0 and the sequence $\{\|x_{n_k}\|\}$ converges to $\|x_0\| = a$, this sequence converges to x_0 strongly in H.

Definition 3.3.5. Assume $\inf \Phi(x) = d > -\infty$ on H. A sequence $\{x_n\}$ is called a *minimizing sequence* of $\Phi(x)$ if $\Phi(x_n) \to d$ as $n \to \infty$.

In the proof of Theorem 3.3.2 we have established that under the conditions of that theorem any sequence minimizing $\Psi(x)$ contains a subsequence that converges strongly to an element at which the minimum of $\Psi(x)$ occurs. Now we can formulate

Theorem 3.3.3. Assume that a functional $\Psi(x) = \|x\|^2 + \Phi(x)$, where $\Phi(x)$ is weakly continuous on H, is growing. Then:

(i) there exists $x_0 \in H$ at which $\Psi(x)$ takes its minimal value;

(ii) any minimizing sequence of $\Psi(x)$ contains a subsequence which converges strongly to a point at which $\Psi(x)$ takes its minimal value: moreover, every weakly convergent subsequence of $\{x_n\}$ converges strongly to a minimizer of $\Psi(x)$;

(iii) if a point x_0 at which $\Psi(x)$ takes its minimal value is unique, then a minimizing sequence converges to x_0 strongly;

(iv) if $\operatorname{grad} \Phi(x_0)$ exists at a point of minimum x_0, then

$$2x_0 + \operatorname{grad} \Phi(x_0) = 0.$$

Proof. By Theorem 3.3.2, on a ball $\|x\| \leq R$ the functional $\Psi(x)$ takes its minimal value. Since $\Psi(x)$ is growing we can take R so large that the minimum is attained inside the open ball $\|x\| < R$. So statements (i) and (ii) follow from Theorem 3.3.3 and Remark 3.3.1. Statement (iv) follows from Theorem 3.3.1. The proof of (iii) is carried out in a way similar to that given in Section 1.23. □

Now we consider the application of the Ritz method to solve the problem of minimizing $\Psi(x)$ under the restrictions of Theorem 3.3.3. First we state the equations of Ritz's method. Let $g_1, g_2, g_3, \ldots$ be a complete system in H such that every finite subsystem is linearly independent. Denote by H_n a subspace of H which is spanned by $g_1, \ldots, g_n$.

The approximation of the Ritz method to minimize the functional $\Psi(x)$ is now formulated as follows:

- Find a minimizer x_n of $\Psi(x)$ on H_n.

- Note that if $\Psi(x)$ has $\operatorname{grad}\Psi(x)$ then the equations to find the nth Ritz approximation are

$$(\operatorname{grad}\Psi(x_n), g_k) = 0, \qquad k = 1, \dots, n, \quad x_n \in H_n.$$

Theorem 3.3.4. Under the restrictions of Theorem 3.3.3, the following hold:

(i) for each n there exists a solution $x_n \in H_n$, the nth Ritz approximation of the minimizer of $\Psi(x)$;

(ii) the sequence of Ritz approximations is a minimizing sequence of $\Psi(x)$, and thus

(iii) the sequence $\{x_n\}$ contains at least one weakly convergent subsequence whose weak limit is a minimizer of $\Psi(x)$ — in fact, this subsequence converges strongly to the minimizer;

(iv) every weakly convergent subsequence of $\{x_n\}$ converges strongly to a minimizer of $\Psi(x)$; if a minimizer of $\Psi(x)$ is unique, then the whole sequence $\{x_n\}$ converges to it strongly.

Proof. (i) Solvability of the problem for the nth approximation of solution by the Ritz method follows from Theorem 3.3.3.

(ii) Let x_0 be a solution to the main problem

$$\Psi(x_0) = d = \inf_{x \in H} \Psi(x).$$

As the system $g_1, g_2, g_3, \dots$ is complete, there is $x^{(n)} \in H_n$ such that

$$\|x_0 - x^{(n)}\| = \delta_n \to 0 \text{ as } n \to \infty.$$

Since $\Psi(x)$ is continuous we get

$$|\Psi(x^{(n)}) - \Psi(x_0)| = \varepsilon_n \to 0 \text{ as } n \to \infty.$$

But x_n is a minimizer of $\Psi(x)$ on H_n, so

$$d = \Psi(x_0) \le \Psi(x_n) = \inf_{x \in H_n} \Psi(x) \le \Psi(x^{(n)}).$$

Therefore

$$|\Psi(x_n) - \Psi(x_0)| \le \varepsilon_n \to 0 \text{ as } n \to \infty$$

and thus $\{x_n\}$ is a minimizing sequence of $\Psi(x)$.

The other statements follow from Theorem 3.3.3. □

Note that Theorem 3.3.4 can be applied to linear and nonlinear problems of mechanics.

3.4 Von Kármán Equations of a Plate

Theorem 3.3.4 will be applied to the boundary value problem of equilibrium of a plate described by the von Kármán equations, which are

$$\Delta^2 w = [f, w] + q \quad \text{in } \Omega \subset \mathbb{R}^2, \tag{3.4.1}$$

$$\Delta^2 f = -[w, w] \quad \text{in } \Omega, \tag{3.4.2}$$

where $w(x, y)$ is the normal displacement of the middle surface Ω of the plate, $f(x, y)$ is the Airy function, $q = q(x, y)$ is the transverse external load, and

$$[u, v] = \frac{\partial^2 u}{\partial x^2}\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\frac{\partial^2 v}{\partial y^2} - 2\frac{\partial^2 u}{\partial x \partial y}\frac{\partial^2 v}{\partial x \partial y}.$$

We consider the Dirichlet problem for these equations:

$$w\big|_{\partial\Omega} = \frac{\partial w}{\partial n}\Big|_{\partial\Omega} = 0, \tag{3.4.3}$$

$$f\big|_{\partial\Omega} = \frac{\partial f}{\partial n}\Big|_{\partial\Omega} = 0. \tag{3.4.4}$$

Let us consider the integro-differential equations

$$a(w, \varphi) = B(f, w, \varphi) + \int_\Omega q\varphi \, d\Omega, \tag{3.4.5}$$

$$a(f, \eta) = -B(w, w, \eta), \tag{3.4.6}$$

where

$$a(w, \varphi) = \int_\Omega \left\{ \frac{\partial^2 w}{\partial x^2}\left(\frac{\partial^2 \varphi}{\partial x^2} + \nu\frac{\partial^2 \varphi}{\partial y^2}\right) + 2(1-\nu)\frac{\partial^2 w}{\partial x \partial y}\frac{\partial^2 \varphi}{\partial x \partial y} + \right.$$
$$\left. + \frac{\partial^2 w}{\partial y^2}\left(\frac{\partial^2 \varphi}{\partial y^2} + \nu\frac{\partial^2 \varphi}{\partial x^2}\right)\right\} d\Omega,$$

$$B(f, w, \varphi) = \int_\Omega \left\{ \left(\frac{\partial^2 f}{\partial x \partial y}\frac{\partial w}{\partial y} - \frac{\partial^2 f}{\partial y^2}\frac{\partial w}{\partial x}\right)\frac{\partial \varphi}{\partial x} + \right.$$
$$\left. + \left(\frac{\partial^2 f}{\partial x \partial y}\frac{\partial w}{\partial x} - \frac{\partial^2 f}{\partial x^2}\frac{\partial w}{\partial y}\right)\frac{\partial \varphi}{\partial y}\right\} d\Omega,$$

ν being Poisson's ratio, $0 < \nu < 1/2$.

Note that $a(u, v)$ is the scalar product (1.10.4) (with an omitted multiplier — the bending rigidity) of the energy space E_{PC} for an isotropic plate, and we shall use this notation in this section.

Suppose that (3.4.5) and (3.4.6), with respect to the unknown function w, f, being smooth (of $C^{(4)}(\overline{\Omega})$) and satisfying the boundary conditions (3.4.3) and (3.4.4), are valid for every φ, η which also satisfy (3.4.3) for these

functions and their normal derivatives on the boundary. The usual tools of the calculus of variations show that the pair (w, f) is a classical solution to the von Kármán equations (3.4.1) and (3.4.2). This means that we can use (3.4.5) and (3.4.6) to define a generalized solution to the problem under consideration. We note that (3.4.5) expresses the virtual work principle for the plate, and (3.4.6) is the equation of compatibility. So we introduce

Definition 3.4.1. A pair (w, f), $w, f \in E_{PC}$, is called a generalized solution to the problem (3.4.1)–(3.4.4) if it satisfies the integro-differential equations (3.4.5)–(3.4.6) for any (φ, η), $\varphi, \eta \in E_{PC}$.

For correctness of the definition the load $q = q(x, y)$ must be such that the term $\int_\Omega q\varphi \, d\Omega$ is a continuous linear functional in E_{PC}; for this it suffices that, say, q be of $L^1(\Omega)$ (cf., Section 1.14).

Under the restrictions of the definition, all terms in (3.4.5) and (3.4.6) make sense as each of the first derivatives of any of the functions under consideration are of $L^p(\Omega)$ with any $p < \infty$. Indeed, a typical term which is not present in a linear statement of the plate problem is bounded as

$$\left| \int_\Omega \frac{\partial^2 f}{\partial x^2} \frac{\partial w}{\partial y} \frac{\partial \varphi}{\partial y} \, d\Omega \right| \le \left(\int_\Omega \left| \frac{\partial^2 f}{\partial x^2} \right|^2 d\Omega \right)^{1/2} \cdot \cdot \left(\int_\Omega \left| \frac{\partial w}{\partial y} \right|^4 d\Omega \right)^{1/4} \left(\int_\Omega \left| \frac{\partial \varphi}{\partial y} \right|^4 d\Omega \right)^{1/4}, \quad (3.4.7)$$

and hence is finite.

We could present a functional whose gradient in the space $E_{PC} \times E_{PC}$ is defined by (3.4.5) and (3.4.6); unfortunately it is not of the form required by Theorem 3.3.4. That is why we shall reformulate the problem with respect to the only unknown function w, defining f as an operator with respect to w and construct a functional of w whose critical point is a generalized solution of the problem. We now embark on this program.

So let w be a fixed but arbitrary element of E_{PC}. Consider $B(w, w, \eta)$ as a functional with respect to η in E_{PC}. It is clearly linear. By (3.4.7) written for a typical term with $f = w$, thanks to the imbedding theorem in E_{PC}, we get

$$|B(w, w, \eta)| \le m \|w\|^2_{E_P} \|\eta\|_{E_P},$$

i.e., the functional is continuous and so we can apply the Riesz representation theorem to get

$$-B(w, w, \eta) = (c, \eta)_{E_P} = a(c, \eta).$$

Being uniquely defined by $w \in E_{PC}$, the element $c \in E_{PC}$ can be considered as a value of a nonlinear operator

$$c = C(w), \qquad a(C(w), \eta) = -B(w, w, \eta). \quad (3.4.8)$$

Before studying the properties of C we introduce

Definition 3.4.2. An operator A mapping from a Banach space X to a Banach space Y is called *compact* if it is continuous in X and takes every bounded set of X into a precompact set in Y. An operator is called *completely continuous* if it takes every weakly convergent sequence of X, $x_n \rightharpoonup x_0$, into a sequence $A(x_n)$ converging strongly to $A(x_0)$.

Lemma 3.4.1. A completely continuous operator F mapping a Hilbert space X into a Banach space Y is compact.

Proof. F is continuous since when a sequence $\{x_n\}$ converges to x_0 strongly in X then it converges to x_0 weakly, too.

Next we take a bounded set S in X and let $\{x_n\}$ be a sequence lying in S. From $\{x_n\}$, thanks to its boundedness, we can choose a subsequence $\{x_{n_k}\}$ converging weakly to $x_0 \in X$. Then, by definition of complete continuity, we get the sequence $\{F(x_{n_k})\}$ converging to $F(x_0)$ strongly. This means $F(S)$ is precompact, hence F is compact. □

It is known that there are compact operators in a Hilbert space which are not completely continuous.

Corollary 3.4.1. If $F(x)$ is a completely continuous operator, then the functional $\|F(x)\|^2$ is a weakly continuous functional in X.

The proof is evident. Now we can prove

Lemma 3.4.2. The operator $C(w)$ defined by (3.4.8) is completely continuous.

Proof. When the functions $u, v, w \in E_{PC}$ are smooth, direct integration by parts gives

$$B(u, v, w) = B(v, u, w) = B(v, w, u) = B(w, u, v); \tag{3.4.9}$$

the limit passage shows that this is valid for $u, v, w \in E_{PC}$. So

$$-B(w, w, \eta) = \int_\Omega \left\{ \left(\frac{\partial w}{\partial x}\right)^2 \frac{\partial^2 \eta}{\partial y^2} + \left(\frac{\partial w}{\partial y}\right)^2 \frac{\partial^2 \eta}{\partial x^2} - 2\frac{\partial w}{\partial x}\frac{\partial w}{\partial y}\frac{\partial^2 \eta}{\partial x \partial y} \right\} d\Omega.$$

Next we take an arbitrary sequence $\{w_n\}$ converging weakly to w_0 in E_{PC} and consider

$$|a(C(w_n) - C(w_0), \eta)| = |B(w_n, w_n, \eta) - B(w_0, w_0, \eta)|.$$

Using the Hölder inequality, we bound a typical term of the right-hand side of this equality as follows:

$$\begin{aligned} d_n &= \left| \int_\Omega \left[\left(\frac{\partial w_n}{\partial x} \right)^2 - \left(\frac{\partial w_0}{\partial x} \right)^2 \right] \frac{\partial^2 \eta}{\partial y^2} \, d\Omega \right| \\ &= \left| \int_\Omega \left(\frac{\partial w_n}{\partial x} - \frac{\partial w_0}{\partial x} \right) \left(\frac{\partial w_n}{\partial x} + \frac{\partial w_0}{\partial x} \right) \frac{\partial^2 \eta}{\partial y^2} \, d\Omega \right| \\ &\le \left\| \frac{\partial w_n}{\partial x} - \frac{\partial w_0}{\partial x} \right\|_{L^4(\Omega)} \left(\left\| \frac{\partial w_n}{\partial x} \right\|_{L^4(\Omega)} + \left\| \frac{\partial w_0}{\partial x} \right\|_{L^4(\Omega)} \right) \left\| \frac{\partial^2 \eta}{\partial y^2} \right\|_{L^2(\Omega)}. \end{aligned}$$

By the imbedding theorem in E_{PC}, which is a subspace of $W^{2,2}(\Omega)$, we get

$$d_n \le m_1 \left\| \frac{\partial w_n}{\partial x} - \frac{\partial w_0}{\partial x} \right\|_{L^4(\Omega)} (\|w_n\|_{E_P} + \|w_0\|_{E_P}) \|\eta\|_{E_P}$$

and, thanks to the boundedness of a weakly convergent sequence,

$$d_n \le m_2 \|w_n - w_0\|_{W^{1,4}(\Omega)} \|\eta\|_{E_P}$$

where m_1 and m_2 are constants.

Gathering all such bounds, we obtain

$$|a(C(w_n) - C(w_0), \eta)| \le m_3 \|w_n - w_0\|_{W^{1,4}(\Omega)} \|\eta\|_{E_P}.$$

Putting $\eta = C(w_n) - C(w_0)$, we finally obtain

$$\|C(w_n) - C(w_0)\|_{E_P} \le m_3 \|w_n - w_0\|_{W^{1,4}(\Omega)} \to 0 \text{ as } n \to \infty$$

since the imbedding operator of $W^{2,2}(\Omega)$ into $W^{1,4}(\Omega)$ is completely continuous (a particular case of Sobolev's imbedding theorems in $W^{2,2}(\Omega)$). The last limit passage shows that C is completely continuous. □

From this lemma we see that (3.4.6) with a given $w \in E_{PC}$ has the unique solution

$$f = C(w). \tag{3.4.10}$$

If $\{w_n\}$ converges to w_0 weakly in E_{PC}, then $\{f_n\} = \{C(w_n)\}$ converges to $f_0 = C(w_0)$ strongly in E_{PC}.

From now on we consider f in (3.4.5) to be determined by (3.4.10).

For a fixed $w \in E_{PC}$, by bounds of the type (3.4.7), we see that the functional

$$B(f, w, \varphi) + \int_\Omega q\varphi \, d\Omega$$

is linear and continuous with respect to $\varphi \in E_{PC}$. So applying the Riesz representation theorem, we have a representation

$$B(f, w, \varphi) + \int_\Omega q\varphi \, d\Omega = a(U, \varphi)$$

where $U \in E_{PC}$ is uniquely determined by $w \in E_{PC}$; so we define an operator G, $U = G(w)$, acting in E_{PC}, by

$$B(f, w, \varphi) + \int_\Omega q\varphi \, d\Omega = a(G(w), \varphi). \tag{3.4.11}$$

In much the same way that Lemma 3.4.2 is proved we can establish

Lemma 3.4.3. G is a completely continuous operator in E_{PC}.

Now the following is evident:

Lemma 3.4.4. The system of equations (3.4.5)–(3.4.6) defining a generalized solution of the problem under consideration is equivalent to the operator equation

$$w = G(w) \tag{3.4.12}$$

with a completely continuous operator G acting in E_{PC}.

Now we introduce a functional

$$I(w) = \frac{1}{2}a(w, w) + \frac{1}{4}a(f, f) - \int_\Omega qw \, d\Omega$$

where, as we said, f is defined by (3.4.8).

The decisive point of this section is

Lemma 3.4.5. For every $w \in E_{PC}$, we have

$$\operatorname{grad} I(w) = w - G(w). \tag{3.4.13}$$

Proof. In accordance with the definition of the gradient of a functional, we consider

$$\frac{dI(w + t\varphi)}{dt}\Big|_{t=0} = \frac{1}{2}\frac{d}{dt}a(w + t\varphi, w + t\varphi)\Big|_{t=0} + \frac{1}{2}a\left(f, \frac{df}{dt}\right)\Big|_{t=0} - \int_\Omega q\varphi \, d\Omega$$

where $f = C(w + t\varphi)$. It is clear that

$$\frac{1}{2}\frac{d}{dt}a(w + t\varphi, w + t\varphi)\Big|_{t=0} = a(w, \varphi).$$

Using the definition (3.4.8) of C, with regard for the equality $B(w, \varphi, \eta) = B(\varphi, w, \eta)$, a particular case of (3.4.9), we calculate directly that

$$a\left(\frac{df}{dt}\Big|_{t=0}, \eta\right) = -2B(w, \varphi, \eta)$$

and so

$$a\left(f, \frac{df}{dt}\right)\Big|_{t=0} = -2B(w, \varphi, f) = -2B(f, w, \varphi).$$

It follows that

$$\frac{dI(w+t\varphi)}{dt}\Big|_{t=0} = a(w,\varphi) - B(f,w,\varphi) - \int_\Omega q\varphi\, d\Omega$$

and, thanks to (3.4.11),

$$\frac{dI(w+t\varphi)}{dt}\Big|_{t=0} = a(w,\varphi) - a(G(w),\varphi) = a(w-G(w),\varphi).$$

This, by definition of the gradient of a functional, means that (3.4.13) holds. □

Combining Lemmas 3.4.3 and 3.4.4, we have

Lemma 3.4.6. A critical point w of $I(w)$ defines the pair $(w, G(w))$ that is a generalized solution of the problem under consideration.

So we reduce the problem of finding a generalized solution of the problem to the problem of the minimum of a functional (it is not equivalent as there are in general solutions which are not points of minimum of the functional).

To apply Theorem 3.3.3, it remains to verify

Lemma 3.4.7. The functional $2I(w)$ is growing and has the form

$$\|w\|^2_{E_P} + \Phi_1(w)$$

where

$$\Phi_1(w) = \frac{1}{2}a(f,f) - 2\int_\Omega qw\, d\Omega$$

is a weakly continuous functional, f being defined by (3.4.10).

Proof. $2I(w)$ is growing since

$$2I(w) \geq a(w,w) - 2\left|\int_\Omega qw\, d\Omega\right| = \|w\|^2_{E_P} - 2\left|\int_\Omega qw\, d\Omega\right|$$

and

$$2I(w) \geq \|w\|^2_{E_P} - m\|w\|_{E_P} \to \infty \quad \text{if } \|w\|_{E_P} \to \infty$$

as q is assumed to be such that $\int_\Omega qw\, d\Omega$ is a continuous functional with respect to $w \in E_{PC}$.

Weak continuity of $\Phi_1(w)$ is a consequence of Corollary 3.4.1 and Lemma 3.4.2 for $a(f,f) = \|C(w)\|^2_{E_P}$ and the fact that the continuous linear functional $\int_\Omega qw\, d\Omega$ is weakly continuous (by definition). □

So we can reformulate Theorem 3.3.3 in the case of the plate problem as follows

Theorem 3.4.1. Assume q is such that $\int_\Omega qw\,d\Omega$ is a continuous linear functional with respect to w in E_{PC}. Then any critical point of the growing functional $I(w)$ which has at least one point of absolute minimum is a generalized solution of the plate problem in the sense of Definition 3.4.1; any minimizing sequence of $I(w)$ contains at least one subsequence which converges strongly to a generalized solution of the problem; each of the weak limit points of the minimizing sequence, which are strong limit points simultaneously, is a generalized solution to the problem under consideration.

The reader can also reformulate Theorem 3.3.4 in the present case to justify application of the Ritz method (and thus the method of finite elements) to von Kármán equations. Note that in this modification of the method we must find f exactly from (3.4.6). But it is not too difficult to show that f can be found approximately, also by the Ritz method, and the corresponding theorem on convergence remains valid in the present case.

3.5 Buckling of a Thin Elastic Shell

Following an article by I.I. Vorovich [27] (and [28]), we now consider a buckling problem for a shallow elastic shell described by equations of von Kármán's type. We want to study stability of the momentless state (here $w = 0$) of the shell. Assume the external load to be proportional to a parameter λ. For every λ, existence of the momentless state of the shell is seen. We formulate the equations of equilibrium as follows:

$$\begin{aligned}\Delta^2 w &= -\lambda\left(T_1\frac{\partial^2 w}{\partial x^2}+T_2\frac{\partial^2 w}{\partial y^2}+2T_{12}\frac{\partial^2 w}{\partial x\partial y}-F_1\frac{\partial w}{\partial x}-F_2\frac{\partial w}{\partial y}\right)+\\ &\quad+[f,w+z],\\ \Delta^2 f &= -\{2[z,w]+[w,w]\}.\end{aligned}\tag{3.5.1}$$

We study a problem with Dirichlet conditions

$$w\big|_{\partial\Omega}=\frac{\partial w}{\partial n}\Big|_{\partial\Omega}=f\big|_{\partial\Omega}=\frac{\partial f}{\partial n}\Big|_{\partial\Omega}=0.\tag{3.5.2}$$

Here $z=z(x,y)\in C^{(3)}(\overline{\Omega})$ is the equation of mid-surface of the shell. It is supposed that the tangential stresses T_1, T_2, T_{12} are given, belong to $L^2(\Omega)$ and, as assumed during derivation of the equations, satisfy equations of the two-dimensional theory of elasticity with forces (F_1, F_2). Other bits of notation are taken from the previous section.

The equations of a generalized statement of the problem under consideration are as follows:

$$a(w,\varphi)=\lambda\int_\Omega\left[T_1\frac{\partial w}{\partial x}\frac{\partial\varphi}{\partial x}+T_2\frac{\partial w}{\partial y}\frac{\partial\varphi}{\partial y}+T_{12}\left(\frac{\partial w}{\partial x}\frac{\partial\varphi}{\partial y}+\frac{\partial w}{\partial y}\frac{\partial\varphi}{\partial x}\right)\right]dx\,dy + B(f,w+z,\varphi), \tag{3.5.3}$$

$$a(f,\eta)=-2B(z,w,\eta)-B(w,w,\eta). \tag{3.5.4}$$

Using standard variational tools, we can derive from these the equations (3.5.1) if a solution is assumed to be sufficiently smooth; conversely, we can derive (3.5.1) from (3.5.3)–(3.5.4). So we can take the latter equations to formulate

Definition 3.5.1. A pair w, f from E_{PC} is called a generalized solution to the problem (3.5.1)–(3.5.2) if it satisfies the integro-differential equations (3.5.3)–(3.5.4) for any $\varphi,\eta\in E_{PC}$

The problem under consideration has a trivial solution $w=f=0$. We are interested in when there exists a nontrivial solution, i.e., in solving a nonlinear eigenvalue problem.

First we mention that, as in Section 3.4, we solve the equation (3.5.4) and then exclude $f\in E_{PC}$ from the equation (3.5.3) using the solution f of (3.5.4) when $w\in E_{PC}$ is given. It is clear that

$$f=f_1+f_2$$

where the f_i are defined by the equations

$$a(f_1,\eta)=-2B(z,w,\eta),\qquad a(f_2,\eta)=-B(w,w,\eta).$$

Using the Riesz representation theorem we can find from these that

$$f_1=Lw,\qquad f_2=C(w). \tag{3.5.5}$$

In Section 3.4 it was shown that $C(w)$ is a completely continuous operator. The same is valid for the linear operator L (we leave it to the reader to show this).

In Section 2.5, we introduced the self-adjoint bounded operator C that is now redenoted as K. It is defined by

$$a(Kw,\varphi)=\int_\Omega\left[T_1\frac{\partial w}{\partial x}\frac{\partial\varphi}{\partial x}+T_{12}\left(\frac{\partial w}{\partial x}\frac{\partial\varphi}{\partial y}+\frac{\partial w}{\partial y}\frac{\partial\varphi}{\partial x}\right)+T_2\frac{\partial w}{\partial y}\frac{\partial\varphi}{\partial y}\right]dx\,dy.$$

K is compact in E_{PC} as follows from Sobolev's imbedding theorem.

Applying the Riesz representation theorem to the relation (3.5.3) wherein f is defined by (3.5.5), we find an operator equation for a generalized solution of the problem under consideration

$$w-G(\lambda,w)=0. \tag{3.5.6}$$

The next point is to define a functional whose critical points are solutions to (3.5.6). It is

$$I(\lambda, w) = \frac{1}{2}a(w, w) + \frac{1}{4}a(f, f) - \lambda J(w)$$

where

$$J(w) = \frac{1}{2}\int_\Omega \left[T_1 \left(\frac{\partial w}{\partial x}\right)^2 + 2T_{12}\frac{\partial w}{\partial x}\frac{\partial w}{\partial y} + T_2 \left(\frac{\partial w}{\partial y}\right)^2 \right] dx\, dy.$$

$I(\lambda, w)$ is the total energy of the system "shell-load."

Lemma 3.5.1. For every $w \in E_{PC}$ we have

$$\operatorname{grad} I(\lambda, w) = w - G(\lambda, w). \tag{3.5.7}$$

The proof is similar to that for Lemma 3.4.4 and is omitted, as is the proof that $G(\lambda, w)$ is a completely continuous operator in $w \in E_{PC}$.

Next we consider the functional $a(f, f)$. It is seen that

$$\begin{aligned} a(f, f) &= a(f_1, f_1) + A_3(w) + A_4(w), \\ A_3(w) &= 2a(f_1, f_2) = -4B(z, w, f_2), \\ A_4(w) &= a(f_2, f_2) = \frac{1}{2}B(f_2, w, w). \end{aligned}$$

Here $A_k(w)$ is a homogeneous function of order k with respect to w, i.e.,

$$A_k(tw) = t^k A_k(w).$$

We leave it to the reader to show that $a(f, f)$, along with each of its parts, is a weakly continuous functional on E_{PC} (for $a(f, f)$, this is a consequence of Corollary 3.4.1).

It is evident that $J(w)$ is a weakly continuous functional in E_{PC}. So we have

Lemma 3.5.2. For every real number λ, the functional $I(\lambda, w)$ takes the form

$$I(\lambda, w) = \frac{1}{2}\|w\|^2_{E_{PC}} + \Psi(\lambda, w), \qquad \Psi(\lambda, w) = \frac{1}{4}a(f, f) - \lambda J(w),$$

where $\Psi(\lambda, w)$ is a weakly continuous functional.

From now on, we assume that

$$J(w) > 0 \quad \text{if } w \neq 0,\ w \in E_{PC}. \tag{3.5.8}$$

This assumption has the physical implication that almost everywhere in the shell the stress state of the shell is compressive.

To study stability of the non-buckled state of the shell (that is, when $w = 0$), beginning from L. Euler's work on stability of a bar, one solves the linearized (here around zero state) eigenvalue problem that is now

$$\operatorname{grad}\left[\frac{1}{2}a(w,w) + \frac{1}{4}a(f_1,f_1)\right] = \lambda \operatorname{grad} J(w). \tag{3.5.9}$$

The lowest eigenvalue of the latter, denoted λ_E and called the Euler lowest critical value, is usually considered as a value when the main, trivial form of equilibrium of the shell becomes unstable. We shall analyze this method for the shell.

We begin with the eigenvalue problem (3.5.9).

Lemma 3.5.3. There is a countable set λ_k of eigenvalues $\lambda_k > 0$ of the equation (3.5.9) considered in E_{PC}.

Proof. We first mention that the scalar product

$$\langle w, \varphi\rangle = a(w,\varphi) + \frac{1}{2}a(Lw, L\varphi), \qquad f_1 = Lw,$$

induces the norm in E_{PC} which is equivalent to the usual one since

$$a(w,w) \le \langle w, w\rangle \le m\, a(w,w).$$

Using the new norm, we can rewrite (3.5.9) in the form

$$w = \lambda K_1 w$$

where K_1 is determined, thanks to the Riesz representation theorem, by the equality

$$\langle K_1 w, \varphi\rangle = \int_\Omega \left[T_1 \frac{\partial w}{\partial x}\frac{\partial \varphi}{\partial x} + T_{12}\left(\frac{\partial w}{\partial x}\frac{\partial \varphi}{\partial y} + \frac{\partial w}{\partial y}\frac{\partial \varphi}{\partial x}\right) + T_2 \frac{\partial w}{\partial y}\frac{\partial \varphi}{\partial y}\right] dx\, dy.$$

It is easily seen that K_1, as well as K, is strictly positive, self-adjoint, and compact, and thus we can use Theorem 2.14.2 which gives even more than the lemma states. □

For the trivial solution $w = f = 0$, the total energy $I(\lambda, w) = 0$. A state of the shell at which $I(\lambda, w)$ takes its minimal value is, in a certain sense, stable. So it is of interest what is the range of λ in which $I(\lambda, w)$ can take negative values.

Theorem 3.5.1. Assume $T_1, T_{12}, T_2 \in L^2(\Omega)$ and w_E is an eigenfunction of the linearized boundary value problem (3.5.9) corresponding to its smallest eigenvalue λ_E, the Euler critical value. Then for every λ of the half-line

$$\lambda > \lambda^* \equiv \lambda_E - \frac{A_3^2(w_E)}{4A_4(w_E)J(w_E)} \tag{3.5.10}$$

there exists at least one nontrivial solution of the nonlinear boundary value problem (3.5.6) at which $I(\lambda, w)$ is negative.

The proof is a consequence of the following three lemmas. The first of them is auxiliary.

Lemma 3.5.4. Assume that $w \in E_{PC}$ satisfies

$$\frac{\partial^2 w}{\partial x^2}\frac{\partial^2 w}{\partial y^2} - \left(\frac{\partial^2 w}{\partial x \partial y}\right)^2 = 0 \tag{3.5.11}$$

in the sense of $L^1(\Omega)$ (almost everywhere in Ω). Then $w = 0$.

Proof. If $w \in C^{(2)}(\Omega)$, then (3.5.11) means the Gaussian curvature of the surface $z = w(x, y)$ vanishes so the surface is developable and, thanks to the boundary conditions (3.5.2), $w = 0$.

If $w \notin C^{(2)}(\Omega)$, we take another route. For arbitrary $w \in E_{PC}$, $F \in W^{2,2}(\Omega)$, the following formula holds:

$$\int_\Omega \left[\left(\frac{\partial^2 F}{\partial x \partial y}\frac{\partial w}{\partial y} - \frac{\partial^2 F}{\partial y^2}\frac{\partial w}{\partial x}\right)\frac{\partial w}{\partial x} + \left(\frac{\partial^2 F}{\partial x \partial y}\frac{\partial w}{\partial x} - \frac{\partial^2 F}{\partial x^2}\frac{\partial w}{\partial y}\right)\frac{\partial w}{\partial y}\right] dx\, dy$$
$$= 2\int_\Omega \left[\frac{\partial^2 w}{\partial x^2}\frac{\partial^2 w}{\partial y^2} - \left(\frac{\partial^2 w}{\partial x \partial y}\right)^2\right] F\, dx\, dy. \tag{3.5.12}$$

(This is easily seen after integrating by parts for smooth functions; the limit passage shows that it is valid for the needed classes.) In (3.5.12) we put

$$F = \frac{1}{2}(x^2 + y^2)$$

which gives for w satisfying (3.5.11)

$$\int_\Omega \left[\left(\frac{\partial w}{\partial x}\right)^2 + \left(\frac{\partial w}{\partial y}\right)^2\right] dx\, dy = 0.$$

This, together with the boundary conditions for w, completes the proof. □

Lemma 3.5.5. The functional $I(\lambda, w)$ is growing for every $\lambda > 0$; that is, we have $I(\lambda, w) \to \infty$ as $\|w\|_{E_P} \to \infty$.

Proof. On the unit sphere $S = \{w\colon a(w, w) = 1\}$ of E_{PC} consider the set S_1 defined by

$$\frac{1}{2}a(w, w) - \lambda J(w) > \frac{1}{4}.$$

Then on the image of S_1 under the mapping $w \mapsto Rw$, we get

$$\begin{aligned} I(\lambda, Rw) &\geq \frac{1}{2}a(Rw, Rw) - \lambda J(Rw) \\ &= R^2\left[\frac{1}{2}a(w, w) - \lambda J(w)\right] \\ &> \frac{1}{4}R^2, \qquad w \in S_1. \end{aligned} \tag{3.5.13}$$

Next consider $I(\lambda, Rw)$ when $w \in S_2 = S \setminus S_1$. Here

$$\frac{1}{2}a(w,w) - \lambda J(w) \le \frac{1}{4}. \tag{3.5.14}$$

Let us introduce the weak closure of S_2 in E_{PC}, denoted by $\operatorname{Cl} S_2$. First we show that $\operatorname{Cl} S_2$ does not contain zero. If to the contrary it does contain zero then there is a sequence $\{w_n\} \in \operatorname{Cl} S_2$ such that $a(w_n, w_n) = 1$ and $w_n \rightharpoonup 0$ in E_{PC} (or, equivalently, in $W^{2,2}(\Omega)$). By the imbedding theorem in $W^{2,2}(\Omega)$, the sequences of first derivatives of $\{w_n\}$ tend to zero strongly in $L^p(\Omega)$ for any $p < \infty$ and thus $J(w_n) \to 0$, which contradicts (3.5.14) since

$$\frac{1}{2} \equiv \frac{1}{2}a(w_n, w_n) \le \frac{1}{4} + \lambda J(w_n).$$

Next we show that for all $w \in \operatorname{Cl} S_2$,

$$A_4(w) \ge c_* \tag{3.5.15}$$

wherein c_* is a positive constant. Indeed, if (3.5.15) is not valid there is a sequence $\{w_n\} \in \operatorname{Cl} S_2$ such that $A_4(w_n) \to 0$ as $n \to \infty$. This sequence contains a subsequence which converges weakly to w_0 belonging to $\operatorname{Cl} S_2$ too. Since A_4 is a weakly continuous functional,

$$A_4(w_0) = 0.$$

This means that

$$a(f_2, f_2) = 0, \qquad f_2 = C(w_0).$$

Returning to (3.5.5), we get

$$B(w_0, w_0, \eta) = 0$$

or, equivalently,

$$\int_\Omega \left[\frac{\partial^2 w_0}{\partial x^2} \frac{\partial^2 w_0}{\partial y^2} - \left(\frac{\partial^2 w_0}{\partial x \partial y} \right)^2 \right] \eta \, dx \, dy = 0$$

for any $\eta \in E_{PC}$. As E_{PC} is dense in $L^2(\Omega)$,

$$\frac{\partial^2 w_0}{\partial x^2} \frac{\partial^2 w_0}{\partial y^2} - \left(\frac{\partial^2 w_0}{\partial x \partial y} \right)^2 = 0$$

almost everywhere in Ω and, by Lemma 3.5.3, it follows that $w_0(x, y) = 0$. This contradicts the fact that w_0 belongs to $\operatorname{Cl} S_2$ which does not contain zero.

Since $|A_3(w)| \le c_1$ on S, we get, thanks to (3.5.15),

$$I(\lambda, Rw) \ge c_* R^4 - \left(\frac{1}{4} R^2 + c_1 R^3 \right)$$

when $w \in \mathrm{Cl}\, S_2$ and so for sufficiently large R, with regard for (3.5.13), we obtain

$$I(\lambda, Rw) \geq \frac{1}{4}R^2$$

for all $w \in S$. This means that $I(\lambda, w)$ is growing. □

By Theorem 3.3.3 it follows that, for any λ, the functional $I(\lambda, w)$ takes its minimal value in E_{PC}. But $w = 0$ is also a critical point of the functional, so to conclude the proof of Theorem 3.5.1 we formulate

Lemma 3.5.6. Under the conditions of Theorem 3.5.1, the minimal value of $I(\lambda, w)$ is negative if λ satisfies (3.5.10).

Proof. Consider $I(\lambda, cw_E)$ where c is a constant. It is seen that

$$I(\lambda, cw_E) = c^2 \left[\frac{1}{2}a(w_E, w_E) + \frac{1}{4}a(Lw_E, Lw_E) - \lambda J(w_E)\right] + \\ + c^3 A_3(w_E) + c^4 A_4(w_E), \qquad (f_1 = Lw_E).$$

Further, from (3.5.9) it follows that

$$\frac{1}{2}a(w_E, w_E) + \frac{1}{4}a(Lw_E, Lw_E) = \lambda_E J(w_E).$$

Hence

$$I(\lambda, cw_E) = c^2 \left[(\lambda_E - \lambda)J(w_E) + cA_3(w_E) + c^2 A_4(w_E)\right].$$

The minimum of $I(\lambda, cw_E)/c^2$ considered as a function of the real variable c is taken at

$$c_0 = -\frac{1}{2}A_3(w_E)/A_4(w_E);$$

this minimum is equal to

$$\min_c(c^{-2}I(\lambda, cw_E)) = (\lambda_E - \lambda)J(w_E) - A_3^2(w_E)/A_4(w_E).$$

So for λ satisfying (3.5.10), we get

$$I(\lambda, c_0 w_E) < 0$$

and thus at w_0, a minimizer of $I(\lambda, w)$ at the same λ,

$$I(\lambda, w_0) < 0.$$

This completes the proof of the lemma, and therefore of Theorem 3.5.1. □

A very important result follows from Theorem 3.5.1.

Corollary 3.5.1. Assume that there is an eigenfunction w_E corresponding to the Euler critical value λ_E such that

$$A_3(w_E) \neq 0.$$

In this case we have a sharp inequality $\lambda^* < \lambda_E$.

This result is of fundamental importance in the theory of stability of shells, since from it we have that if $A_3(w_E) \neq 0$, then the problem of stability cannot be solved by linearization in the neighborhood of a momentless state of stress, used since Euler in the theory of stability of rods. If $A_3(w_E) \neq 0$, then we must investigate the problem of stability of a shell in its nonlinear formulation.

Theorem 3.5.2. Let $T_1, T_2, T_{12} \in L^2(\Omega)$. Then there is a value $\lambda_l \leq \lambda^*$ such that for any $\lambda < \lambda_l$ the nonlinear problem (3.5.6) has the unique solution $w = 0$.

Proof. Assume w is a solution of (3.5.6), i.e., the pair $w, f = Lw + C(w)$ from E_{PC} satisfies (3.5.3)–(3.5.4) for arbitrary $\varphi, \eta \in E_{PC}$. Setting $\varphi = w$ and $\eta = f$ in (3.5.3)–(3.5.4) we get

$$\begin{aligned} a(w,w) &= 2\lambda J(w) + B(f,w,w) + B(f,z,w), \\ a(f,f) &= -2B(z,w,f) - B(w,w,f). \end{aligned}$$

Summing these equalities term by term, we have the identity

$$a(w,w) + a(f,f) = 2\lambda J(w) - B(z,f,w). \tag{3.5.16}$$

Using the elementary inequality $|ab| \leq a^2 + \frac{1}{4}b^2$, we get an estimate

$$\begin{aligned} |B(z,f,w)| &= \left| \int_\Omega \left(\frac{\partial^2 f}{\partial x^2}\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}\frac{\partial^2 z}{\partial y^2} - 2\frac{\partial^2 f}{\partial x \partial y}\frac{\partial^2 z}{\partial x \partial y} \right) w \, dx\, dy \right| \\ &\leq \int_\Omega \left[\left(\frac{\partial^2 f}{\partial x^2} \right)^2 + \left(\frac{\partial^2 f}{\partial y^2} \right)^2 + 2\left(\frac{\partial^2 f}{\partial x \partial y} \right)^2 \right] dx\, dy + \\ &+ \frac{1}{4} \int_\Omega \left[\left(\frac{\partial^2 z}{\partial x^2} \right)^2 + \left(\frac{\partial^2 z}{\partial y^2} \right)^2 + 2\left(\frac{\partial^2 z}{\partial x \partial y} \right)^2 \right] w^2 \, dx\, dy. \end{aligned}$$

Integrating by parts in the expression for $a(f,f)$ gives

$$a(f,f) = \int_\Omega \left[\left(\frac{\partial^2 f}{\partial x^2} \right)^2 + \left(\frac{\partial^2 f}{\partial y^2} \right)^2 + 2\left(\frac{\partial^2 f}{\partial x \partial y} \right)^2 \right] dx\, dy$$

and thus, from (3.5.16), it follows that

$$a(w,w) \leq 2\lambda J(w) + \frac{1}{4} \int_\Omega \left[\left(\frac{\partial^2 z}{\partial x^2} \right)^2 + \left(\frac{\partial^2 z}{\partial y^2} \right)^2 + 2\left(\frac{\partial^2 z}{\partial x \partial y} \right)^2 \right] w^2 \, dx\, dy. \tag{3.5.17}$$

Now we need a lemma which will be proved later.

Lemma 3.5.7. On the surface $S = \{w \mid J(w) = 1\}$ in E_{PC}, the functional

$$I_1(w) = a(w,w) - \frac{1}{4}\int_\Omega \left[\left(\frac{\partial^2 z}{\partial x^2}\right)^2 + \left(\frac{\partial^2 z}{\partial y^2}\right)^2 + 2\left(\frac{\partial^2 z}{\partial x \partial y}\right)^2\right] w^2\, dx\, dy$$

has finite minimum denoted by $2\lambda^{**}$.

We are continuing the proof. From this lemma, it follows that

$$I_1(w) \geq 2\lambda^{**} J(w)$$

since all of the functionals are homogeneous with respect to w of order 2. Thus, from (3.5.17), we get

$$(2\lambda^{**} - 2\lambda)J(w) \leq 0,$$

from which it follows that if $\lambda \leq \lambda^{**}$ then

$$J(w) \leq 0.$$

This is possible only at $w = 0$, and the proof is complete. □

Proof of Lemma 3.5.7. Assume $\{w_n\}$ is a minimizing sequence of $I_1(w)$ on S and, by contradiction, that the minimum on S is not finite, i.e., $I_1(w_n) \to -\infty$ as $n \to \infty$. It is quite obvious that $\|w_n\|_{E_P} \to \infty$.

Define $w_n^* = w_n/\|w_n\|_{E_P}$. We can consider the sequence $\{w_n^*\}$ to be weakly convergent to an element $w_0^* \in E_{PC}$. In this case

$$J(w_n) = \|w_n\|_{E_P}^2 J(w_n^*)$$

so

$$J(w_n^*) = J(w_n)/\|w_n\|_{E_P}^2 \to 0 \text{ as } n \to \infty.$$

Since J is weakly continuous then $J(w_0^*) = 0$ and thus $w_0^* = 0$. This means that $w_n^* \rightharpoonup 0$.

By the imbedding theorem we get

$$\sup_\Omega |w_n^*(x,y)| \to 0$$

and so

$$a_n = \int_\Omega \left[\left(\frac{\partial^2 z}{\partial x^2}\right)^2 + \left(\frac{\partial^2 z}{\partial y^2}\right)^2 + 2\left(\frac{\partial^2 z}{\partial x \partial y}\right)^2\right] (w_n^*)^2\, dx\, dy \to 0 \text{ as } n \to \infty.$$

Thus

$$\lim_{n\to\infty} I_1(w_n) = \lim_{n\to\infty} \|w_n\|_{E_P}^2 \left(1 - \frac{1}{4}a_n\right) = +\infty,$$

a contradiction. Similar considerations demonstrate that a minimizing sequence $\{w_n\}$ of I_1 is bounded. Then there is a subsequence that converges weakly to an element w_0. This element belongs to S since $J(w)$ is weakly continuous. The structure of I_1 provides that $I_1(w_0) = \lambda^{**}$. □

As a result of Theorem 3.5.2 we get the estimates

$$-\infty < \lambda^{**} \le \lambda_l \le \lambda^* \le \lambda_E < \infty. \tag{3.5.18}$$

From the statement of Lemma 3.5.7, it is seen that λ^{**} can be defined as the lowest eigenvalue of the boundary value problem

$$\operatorname{grad} I_1(w) = 2\lambda \operatorname{grad} J(w). \tag{3.5.19}$$

Let us consider a particular case of a von Kármán plate. Here $z(x, y) = 0$ and thus the problem (3.5.19) takes the form

$$\operatorname{grad}(a(w, w)) = 2\lambda \operatorname{grad} J(w).$$

But the equation (3.5.9) determining the λ_E for the plate coincides with this one as $f_1 = Lw = 0$ for a plate. Thus $\lambda_E = \lambda^{**}$ and (3.5.18) states that $\lambda_l = \lambda_E$ for the plate. This implies an important

Theorem 3.5.3. In the case of a plate ($z(x, y) = 0$), under the conditions of Theorem 3.5.1, the equality $\lambda_l = \lambda_E$ is satisfied. In other words, for $\lambda \le \lambda_E$ there is a unique generalized solution, $w = 0$, of the problem under consideration; if $\lambda > \lambda_E$ then there is another solution of the problem, at which the functional of total energy of the plate is strictly negative.

This theorem establishes the possibility of applying Euler's method of linearization to the problem of stability of a plate.

We note that many works (not mentioned here) are devoted to mathematical questions in the theory of von Kármán's plates and shells. The corresponding boundary value problems of the theory are a touchstone of abstract nonlinear mathematical theory because of their importance in applications, as well as their not too complicated form.

3.6 The Nonlinear Problem of Equilibrium of the Theory of Elastic Shallow Shells

We consider another simple modification of the nonlinear theory of elastic shallow shells when the geometry of the mid-surface of the shell is identified with the geometry of a plane. This modification of the theory is widely applied in engineering calculations. Nonlinear theory of shallow shells in curvilinear coordinates is considered in [26] in detail.

We express the equations describing the behavior of the shell in a notation which is commonly used along with this version of the theory. Namely, we let x, y denote the coordinates on the plane that is identified with the mid-surface of the shell, u, v denote the tangential components of the vector of displacements of the mid-surface, w denote the transverse displacement

of the mid-surface, and subscripts x, y denote partial derivatives with respect to x and y. The equations of equilibrium of the shell are

$$D\nabla^4 w + N_1(k_1 - w_{xx}) + N_2(k_2 - w_{yy}) - 2N_{12}w_{xy} - F = 0, \tag{3.6.1}$$

$$\begin{aligned}
&\nabla^2 u + (1+\mu)/(1-\mu)(u_x + v_y)_x + \\
&\quad + 2/(1-\mu)[(k_1 w)_x + w_x w_{xx} + \mu(k_2 w)_x + \mu w_y w_{xy}] + \\
&\quad + w_y w_{xy} + w_x w_{yy} = 0, \\
&\nabla^2 v + (1+\mu)/(1-\mu)(u_x + v_y)_y + \\
&\quad + 2/(1-\mu)[(k_2 w)_y + w_y w_{yy} + \mu(k_1 w)_y + \mu w_x w_{xy}] + \\
&\quad + w_x w_{xy} + w_y w_{xx} = 0,
\end{aligned} \tag{3.6.2}$$

D, E, μ being the elastic constants, $0 < \mu < 1/2$. We consider the shell under the action of a transverse load F. The components of the tangential strain tensor are

$$\varepsilon_1 = u_x + k_1 w + \frac{1}{2}w_x^2, \quad \varepsilon_2 = v_y + k_2 w + \frac{1}{2}w_y^2, \quad \varepsilon_{12} = u_y + v_x + w_x w_y. \tag{3.6.3}$$

Let us formulate the conditions under which we justify application of Ritz's method to a boundary value problem for the shell, and so for the finite element method as well, and establish an existence theorem.

We suppose Ω, the domain occupied by the shell, satisfies the same conditions we imposed earlier for the von Kármán plate. Let the shell be clamped against the transverse translation at three points (x_i, y_i), $i = 1, 2, 3$, that do not lie on the same straight line:

$$w(x_i, y_i) = 0. \tag{3.6.4}$$

It is sufficient (but not necessary) to assume that

$$w\big|_{\Gamma_1} = 0 \tag{3.6.5}$$

holds on a portion Γ_1 of the boundary.

Let us call C_4 the set of functions w belonging to $C^{(4)}(\Omega)$ and satisfying the conditions (3.6.4)–(3.6.5).

For the tangential displacements u, v, the minimal restrictions in this consideration must be such that Korn's inequality of two-dimensional elasticity holds. That is (see Mikhlin [19]), we must have

$$\int_\Omega (u^2 + v^2 + u_x^2 + u_y^2 + v_x^2 + v_y^2)\,dx\,dy \le m \int_\Omega [u_x^2 + (u_y + v_x)^2 + v_y^2]\,dx\,dy. \tag{3.6.6}$$

One of the possible restrictions under which (3.6.6) holds for all u, v with the unique constant m is

$$u\big|_{\Gamma_2} = 0, \qquad v\big|_{\Gamma_2} = 0, \tag{3.6.7}$$

Γ_2 being some part of the boundary of Ω.

Let us introduce the set C_2 of vector functions $\omega = (u, v)$ with the components belonging to $C^{(2)}(\Omega)$ and satisfying (3.6.7).

We may suppose that some part of the boundary of the shell is elastically supported (the corresponding term of the energy of the system should be included into the expression of the energy norm) or that on some part of the boundary there is given a transverse load (here the term that is the work of the load on the boundary must be included into the energy functional). We will not place these conditions in the differential form; they are well known and can be derived from the variational statement of the problem. The presence of these conditions has no practical impact on the way in which we consider the problem.

Let us introduce energy spaces. Let E_1 be a subspace of $W^{1,2}(\Omega) \times W^{1,2}(\Omega)$ that is the completion of the set C_2 in the norm of $W^{1,2}(\Omega) \times W^{1,2}(\Omega)$. The Korn inequality (3.6.6) implies that on E_1 the following norm is equivalent:

$$\|\omega\|^2_{E_1} = \frac{Eh}{2(1-\mu^2)} \int_\Omega [e_1^2 + e_2^2 + 2\mu e_1 e_2 + \frac{1}{2}(1-\mu)e_{12}^2]\, dx\, dy,$$

where

$$e_1 = u_x, \qquad e_2 = v_y, \qquad e_{12} = u_y + v_x,$$

and h is the shell thickness.

E_2, a subspace of $W^{2,2}(\Omega)$, is the completion of C_4 in the norm of $W^{2,2}(\Omega)$. On E_2 there is an equivalent norm (the energy norm we introduced for the problem of bending of the plate):

$$\|w\|^2_{E_2} = \frac{1}{2} D \int_\Omega [(\nabla^2 w)^2 + 2(1-\mu)(w_{xy}^2 - w_{xx} w_{yy})]\, dx\, dy.$$

The norms on E_i induce the inner products that are denoted with use of the names of corresponding spaces. Denote $E_1 \times E_2$ by E.

Definition 3.6.1. $\mathbf{u} = (u, v, w) \in E$ is called a generalized solution of the problem of equilibrium of a shallow shell if for an arbitrary $\delta\mathbf{u} = (\delta u, \delta v, \delta w) \in E$ it satisfies the equation

$$\int_\Omega (M_1 \delta\kappa_1 + M_2 \delta\kappa_2 + 2M_{12}\delta\chi + N_1 \delta\epsilon_1 + N_2 \delta\epsilon_2 + N_{12}\delta\epsilon_{12})\, dx\, dy = \int_\Omega F \delta w\, dx\, dy + \int_{\partial\Omega} f \delta w\, ds, \tag{3.6.8}$$

where

$$M_1 = D(\kappa_1 + \mu\kappa_2), \qquad M_2 = D(\kappa_2 + \mu\kappa_1), \qquad M_{12} = D(1-\mu)\chi,$$

$$N_1 = \frac{Eh}{1-\mu^2}(\epsilon_1 + \mu\epsilon_2), \quad N_2 = \frac{Eh}{1-\mu^2}(\epsilon_2 + \mu\epsilon_1), \quad N_{12} = \frac{Eh}{2(1+\mu)}\epsilon_{12},$$

$$\kappa_1 = -w_{xx}, \qquad \kappa_2 = -w_{yy}, \qquad \chi = -w_{xy},$$

f being the external load on the edge of the shell.

We note that on the part of the boundary where $\delta w = 0$, it is not necessary to show f. However we shall assume that on this part of the boundary the function $f = 0$.

It is seen that all the stationary points of the energy functional

$$I(\mathbf{u}) = \|w\|^2_{E_2} + \frac{1}{2}\int_\Omega (N_1\epsilon_1 + N_2\epsilon_2 + N_{12}\epsilon_{12})\,dx\,dy - \\ - \int_\Omega Fw\,dx\,dy - \int_{\partial\Omega} fw\,ds \tag{3.6.9}$$

are solutions to (3.6.8) since moving all the terms of (3.6.8) to the left-hand side we get on the left in (3.6.8) the expression for the first variation of the functional $I(\mathbf{u})$.

Let us note that for the correctness of Definition 3.6.1 it is necessary to impose an additional requirement: the terms

$$\int_\Omega F\delta w\,dx\,dy + \int_{\partial\Omega} f\delta w\,ds$$

must make sense for any $\delta w \in E_2$. The set of these loads is called E^*. By Sobolev's imbedding theorems, sufficient conditions for the loads to belong to E^* are:

$$F = F_0 + F_1$$

where $F_0 \in L(\Omega)$ and F_1 is a finite sum of δ-functions (point transverse forces);

$$f = f_0 + f_1$$

where $f_0 \in L(\partial\Omega)$ and f_1 is a finite sum of δ-functions (point transverse forces on $\partial\Omega$). Under these conditions, the functional

$$\int_\Omega F\delta w\,dx\,dy + \int_{\partial\Omega} f\delta w\,ds$$

is linear and continuous in $\delta w \in E_2$.

By the Riesz representation theorem there exists the unique element $g \in E_2$ such that

$$\int_\Omega F\delta w\,dx\,dy + \int_{\partial\Omega} f\delta w\,ds = (g, \delta w)_{E_2}. \tag{3.6.10}$$

Now we can represent $I(\mathbf{u})$ in a more compact form:

$$I(\mathbf{u}) = \|w\|_{E_2}^2 + \frac{1}{2}\int_\Omega (N_1\epsilon_1 + N_2\epsilon_2 + N_{12}\epsilon_{12})\,dx\,dy - (g,w)_{E_2}. \quad (3.6.11)$$

Let us find the tangential displacements u_1, u_2 through w. For this consider the equation

$$\int_\Omega (N_1\delta\epsilon_1 + N_2\delta\epsilon_2 + N_{12}\delta\epsilon_{12})\,dx\,dy = 0$$

in E_1. Reasoning as was done earlier, we can easily establish that this equation is uniquely solvable in E_1 with respect to $\omega = (u_1, u_2)$; the solution can be written as

$$\omega = G(w),$$

where G is a completely continuous operator. Let us put this ω into the expression of $I(\mathbf{u})$. After this substitution, the functional $I(\mathbf{u})$ depends only on w; it is denoted by $\aleph(w)$. Standard reasoning leads us to the statement that any stationary point of $\aleph(w)$ is a generalized solution of the problem under consideration.

The functional $\aleph(w)$ has a structure that is suitable for application of Theorem 3.3.4. To justify the Ritz method it is enough to show that $\aleph(w)$ is growing. Let us demonstrate this.

Lemma 3.6.1. Let the external load belong to E^*. Then $\aleph(w)$ is growing; that is, $\aleph(w) \to \infty$ when $\|w\|_{E_2} \to \infty$.

Proof. The proof follows from considering the form of $\aleph(w)$. Indeed, under the above assumptions, we have

$$N_1\epsilon_1 + N_2\epsilon_2 + 2N_{12}\epsilon_{12} \geq 0.$$

Then

$$|(g, \delta w)_{E_2}| \leq \|g\|_{E_2}\|w\|_{E_2},$$

so

$$\aleph(w) \geq \|w\|_{E_2}^2 - \|g\|_{E_2}\|w\|_{E_2}.$$

From this the lemma follows. □

Thus we have

Theorem 3.6.1. Let the conditions of Lemma 3.6.1 hold. Then

(i) there is a generalized solution of the problem of equilibrium of the shell that belongs to E_2 and admits a minimum of the functional $\aleph(w)$;

(ii) any sequence $\{w_n\}$ minimizing the functional $\aleph(w)$ in E_2 contains a subsequence that converges strongly to a generalized solution of the problem;

(iii) the equations of the Ritz method (and thus of Galerkin's method and so of any conforming modification of the finite element method) have a solution in each approximation; the set of approximations contains a subsequence that converges strongly to a generalized solution of the problem in E_2; moreover, any weakly converging subsequence of the Ritz approximations converges strongly to a generalized solution of the problem.

3.7 Degree Theory

This is only a sketch of degree theory of a map, which will be used in what follows. We begin with an illuminating example.

Let $f(z)$ be a function holomorphic on a closed domain D of the complex plane, and let ∂D, the boundary of D, be smooth and let it not contain zeros of $f(z)$. Then, as is well known, the number defined by the integral

$$n = \oint_{\partial D} \frac{f'(z)}{f(z)}\, dz$$

is equal to the number of zeros of $f(z)$ inside D with regard for their multiplicity.

This is extended to more general classes of maps; this is the so-called degree theory, a full presentation of which can be found in Schwartz [21].

The degree of a finite-dimensional vector-field $\mathbf{\Phi}(\mathbf{x})\colon \mathbb{R}^n \to \mathbb{R}^n$, originally due to L.E.J. Brouwer, is defined as follows. Let $\mathbf{\Phi}(\mathbf{x}) = (\Phi_1(\mathbf{x}), \dots, \Phi_n(\mathbf{x}))$ be continuously differentiable on a bounded open domain D with the boundary ∂D in $\mathbb{R}^n$. Suppose $\mathbf{p} \in \mathbb{R}^n$ does not belong to ∂D, then the set $\Phi^{-1}(\mathbf{p})$, the preimage of $\mathbf{p}$ in D, is discrete and, finally, at each $\mathbf{x} \in \mathbf{\Phi}^{-1}(\mathbf{p})$, the Jacobian

$$J_{\mathbf{\Phi}}(\mathbf{x}) = \det\left(\frac{\partial \Phi_i}{\partial x_j}\right)$$

does not vanish. Then the degree of $\mathbf{\Phi}$ with respect to $\mathbf{p}$ and D is

$$\deg(\mathbf{p}, \mathbf{\Phi}, D) = \sum_{\substack{\mathbf{\Phi}(\mathbf{x})=\mathbf{p} \\ \mathbf{x}\in D}} \operatorname{sign} J_{\mathbf{\Phi}}(\mathbf{x})$$

where $\operatorname{sign} J_{\mathbf{\Phi}}(\mathbf{x})$ is the signum of $J_{\mathbf{\Phi}}(\mathbf{x})$.

If $\deg(\mathbf{p}, \mathbf{\Phi}, D) \neq 0$, then there are solutions of the equation $\mathbf{\Phi}(\mathbf{x}) = \mathbf{p}$ in D. If $\mathbf{p} \notin \mathbf{\Phi}(D)$ then $\deg(\mathbf{p}, \mathbf{\Phi}, D) = 0$ and so $\deg(\mathbf{p}, \mathbf{\Phi}, D)$ determines, in a certain way, the number of solutions of the equation $\mathbf{\Phi}(\mathbf{x}) = \mathbf{p}$.

If there are points $\mathbf{x}$ at which $\mathbf{\Phi}(\mathbf{x}) = \mathbf{p}$ and $J_{\mathbf{\Phi}}(\mathbf{x}) = 0$, then we can introduce the degree of the map using the limit passage. We can always take a sequence of points $\mathbf{p}_k \to \mathbf{p}$ such that $J_{\mathbf{\Phi}}(\mathbf{x}) \neq 0$ at any $\mathbf{x} \in \mathbf{\Phi}^{-1}(\mathbf{p}_k)$;

the degree of $\mathbf{\Phi}$ is now defined by

$$\deg(\mathbf{p}, \mathbf{\Phi}, D) = \lim_{k\to\infty} \deg(\mathbf{p}_k, \mathbf{\Phi}, D).$$

It is shown that this number does not depend on the choice of the sequence $\{\mathbf{p}_k\}$ and also characterizes the number of solutions of the equation $\mathbf{\Phi}(\mathbf{x}) = \mathbf{p}$ in D.

The next step of the theory is to state it for $\mathbf{\Phi}(\mathbf{x})$ being of $C(\overline{D})$ (each of the components of $\mathbf{\Phi}(\mathbf{x})$ being of $C(\overline{D})$). This is done by using a limit passage. Namely, for $\mathbf{\Phi}(\mathbf{x})$, there is a sequence $\{\mathbf{\Phi}_k(\mathbf{x})\}$ such that $\mathbf{\Phi}_k(\mathbf{x}) \in C^{(1)}(\overline{D})$ and each component of $\mathbf{\Phi}_k(\mathbf{x})$ converges uniformly on $\overline{D}$ to a corresponding component of $\mathbf{\Phi}(\mathbf{x})$. Then as is shown, there exists

$$\lim_{k\to\infty} \deg(\mathbf{p}, \mathbf{\Phi}_k, D)$$

which does not depend on the choice of $\{\mathbf{\Phi}_k(\mathbf{x})\}$; it is, by definition, the degree of $\mathbf{\Phi}(x)$ with respect to $\mathbf{p}$ and D.

As there is a one-to-one correspondence between $\mathbb{R}^n$ and n-dimensional real Banach space, the notion of degree of a map is transferred to continuous maps in the latter space. Moreover, it is seen how it can be determined for a continuous map whose range is a finite-dimensional subspace of a Banach space.

In the case of general operator in a Banach space, the notion was extended to operators of the form $I + F$ with a compact operator F on a real Banach space X by J. Leray and J. Schauder [15]. To do this, they introduce an approximate operator as follows.

Let D be a bounded open domain in X with the boundary ∂D. As F is a compact operator, $F(\overline{D})$, the image of $\overline{D}$, is compact. So, by the Hausdorff criterion on compactness, there is a finite ε-net $N_\varepsilon = \{x_k \mid x_k \in F(\overline{D}); k = 1, \ldots, n\}$, such that for every $x \in \overline{D}$ there is an integer k such that $\|F(x) - x_k\| < \varepsilon$. Finally, the approximate operator F_ε is defined by

$$F_\varepsilon(x) = \frac{\sum_{k=1}^n \mu_k(x) x_k}{\sum_{k=1}^n \mu_k(x)}, \qquad x \in \overline{D}$$

where $\mu_k(x) = 0$ if $\|F(x) - x_k\| > \varepsilon$ and $\mu_k = \varepsilon - \|F(x) - x_k\|$ if $\|F(x) - x_k\| \le \varepsilon$. This operator is called the Schauder projection operator.

It is easily seen that the range of $F_\varepsilon(x)$ is a domain in a finite dimensional subspace X_n of X, the operator F_ε is continuous, and, moreover,

$$\|F(x) - F_\varepsilon(x)\| \le \varepsilon$$

when $x \in \overline{D}$.

By the above, we can introduce the degree of $I + F_\varepsilon$ with respect to p and $D_n = D \cap X_n$ if $p \notin (I + F_\varepsilon)(\partial D_n)$. As is shown in Schwartz [21], for sufficiently small $\varepsilon > 0$ the degree $\deg(p, I + F_\varepsilon, D_n)$ is the same and thus it is defined as the degree of the operator $I + F$ with respect to p and D.

The following properties of the degree of an operator $I+F$ with compact operator F hold:

1. If $x + F(x) \neq p$ in $\overline{D}$, then $\deg(p, I + F, D) = 0$;
2. if $\deg(p, I + F, D) \neq 0$, then in D there is at least one solution to the equation $x + F(x) = p$;
3. $\deg(p, I, D) = +1$ if $p \in D$;
4. if $D = \cup_i D_i$ where the family $\{D_i\}$ is disjoint and $\partial D_i \subset \partial D$, then
$$\deg(p, I + F, D) = \sum_i \deg(p, I + F, D_i);$$
5. $\deg(p, I + F, D)$ is continuous with respect to p and F;
6. (invariance under homotopy) Let $\Phi(x, t) = x + \Psi(x, t)$. Assume that for every $t \in [a, b]$ the operator $\Phi(x, t)$ is compact with respect to $x \in X$ and continuous in $t \in [a, b]$ uniformly with respect to $x \in \overline{D}$. Then the operators $\Psi_a = \Psi(\cdot, a)$ and $\Psi_b = \Psi(\cdot, b)$ are said to be compact homotopic. Let Ψ_a and Ψ_b be compact homotopic and $p \neq x+\Psi(x, t)$ for every $x \in \partial D$ and $t \in [a, b]$; then
$$\deg(p, I + \Psi_a, D) = \deg(p, I + \Psi_b, D).$$

The sixth and third properties give a result that is frequently used to establish existence of solution of the equation

$$x + F(x) = 0. \tag{3.7.1}$$

We formulate it as

Lemma 3.7.1. Assume $F(x)$ is a compact operator in a Banach space X and the equation $x + tF(x) = 0$ has no solutions on a sphere $\|x\| = R$ for any $t \in [a, b]$. Then in the ball $B = \{x \mid \|x\| < R\}$ there exists at least one solution to (3.7.1) and

$$\deg(0, I + F, B) = +1.$$

In the next section we demonstrate an application of the lemma.

3.8 Steady-State Flow of Viscous Liquid

Following I.I. Vorovich and V.I. Yudovich [27], we consider the steady-state flow of a viscous incompressible liquid described by the Navier–Stokes equations

$$\nu\Delta\mathbf{v} = (\mathbf{v} \cdot \nabla)\mathbf{v} + \nabla p + \mathbf{f}, \tag{3.8.1}$$

$$\nabla \cdot \mathbf{v} = 0. \tag{3.8.2}$$

Let $\nu > 0$. We are treating a problem with boundary condition

$$\mathbf{v}\big|_{\partial\Omega} = \boldsymbol{\alpha}. \tag{3.8.3}$$

From now on, we assume:

(i) Ω is a bounded domain in $\mathbb{R}^2$ or $\mathbb{R}^3$ whose boundary $\partial\Omega$ consists of r closed curves or surfaces S_k, $k = 1, \ldots, r$ with continuous curvature.

(ii) There is a continuously differentiable vector-function

$$\mathbf{a}(\mathbf{x}) = (a_1(\mathbf{x}), a_2(\mathbf{x}), a_3(\mathbf{x}))$$

such that

$$a_k(\mathbf{x}) \in C^{(1)}(\overline{\Omega}), \qquad \nabla \cdot \mathbf{a} = 0 \text{ in } \Omega, \qquad \mathbf{a}\big|_{\partial\Omega} = \boldsymbol{\alpha}.$$

(iii) On each S_k, $k = 1, \ldots, r$, we have

$$\int_{S_k} \boldsymbol{\alpha} \cdot \mathbf{n}\, dS = 0 \tag{3.8.4}$$

where $\mathbf{n}$ is the unit outward normal at a point of S_k.

We note that the condition

$$\sum_{k=1}^{r} \int_{S_k} \boldsymbol{\alpha} \cdot \mathbf{n}\, dS = 0$$

is necessary for solvability of the problem.

Let $H(\Omega)$ be the completion of the set $S^0(\Omega)$ of all smooth solenoidal vector-functions $\mathbf{u}(\mathbf{x})$ satisfying the boundary condition, in the norm induced by the scalar product

$$(\mathbf{u}, \mathbf{v})_{H(\Omega)} = \int_\Omega \nabla\mathbf{u} \cdot \nabla\mathbf{v}\, d\Omega \equiv \int_\Omega \operatorname{rot} \mathbf{u} \cdot \operatorname{rot} \mathbf{v}\, d\Omega$$

and so each of the components of $\mathbf{u}(\mathbf{x}) \in H(\Omega)$ is of $W^{1,2}(\Omega)$. Thus in the three dimensional case, the imbedding operator of $H(\Omega)$ into $(L^p(\Omega))^3$ is continuous when $1 \le p \le 6$ and compact when $1 \le p < 6$; in the two dimensional case, the imbedding operator is compact into $(L^p(\Omega))^2$ for any $1 \le p < \infty$.

We assume

(iv) $f_k(\mathbf{x}) \in L^p(\Omega)$, $p \ge 6/5$ in the three dimensional case ($k = 1, 2, 3$), $p > 1$ in the two dimensional case ($k = 1, 2$).

Definition 3.8.1. $\mathbf{v}(\mathbf{x}) = \mathbf{a}(\mathbf{x}) + \mathbf{u}(\mathbf{x})$ is called a generalized solution to the problem (3.8.1)–(3.8.3) if $\mathbf{u}(\mathbf{x}) \in H(\Omega)$ and satisfies the integro-differential equation

$$\nu(\mathbf{u}, \mathbf{\Phi})_{H(\Omega)} = -\int_\Omega [(\mathbf{u}\cdot\nabla)\mathbf{u}\cdot\mathbf{\Phi} + (\mathbf{u}\cdot\nabla)\mathbf{a}\cdot\mathbf{\Phi} + (\mathbf{a}\cdot\nabla)\mathbf{u}\cdot\mathbf{\Phi} + \\ + (\mathbf{a}\cdot\nabla)\mathbf{a}\cdot\mathbf{\Phi} + \nu \operatorname{rot}\mathbf{a}\cdot\operatorname{rot}\mathbf{\Phi} + \mathbf{f}\cdot\mathbf{\Phi}]\, d\Omega \qquad (3.8.5)$$

for any $\mathbf{\Phi} \in H(\Omega)$.

It is easily seen that if $\mathbf{a}(\mathbf{x})$ and $\mathbf{u}(\mathbf{x})$ belong to $C^{(2)}(\overline{\Omega})$ then $\mathbf{v}(\mathbf{x})$ is a classical solution to the problem (3.8.1)–(3.8.3).

Note that there are infinitely many vectors $\mathbf{a}(\mathbf{x})$ satisfying the assumption (ii) if there is one, but the set of generalized solutions does not depend on the choice of $\mathbf{a}(\mathbf{x})$.

To use Lemma 3.7.1, we reduce equation (3.8.5) to the operator form $\mathbf{u} + F(\mathbf{u}) = 0$, defining F with use of the Riesz representation theorem from the equality

$$\nu(F(\mathbf{u}), \mathbf{\Phi})_{H(\Omega)} = \int_\Omega [(\mathbf{u}\cdot\nabla)\mathbf{u}\cdot\mathbf{\Phi} + (\mathbf{u}\cdot\nabla)\mathbf{a}\cdot\mathbf{\Phi} + (\mathbf{a}\cdot\nabla)\mathbf{u}\cdot\mathbf{\Phi} + \\ + (\mathbf{a}\cdot\nabla)\mathbf{a}\cdot\mathbf{\Phi} + \nu \operatorname{rot}\mathbf{a}\cdot\operatorname{rot}\mathbf{\Phi} + \mathbf{f}\cdot\mathbf{\Phi}]\, d\Omega. \qquad (3.8.6)$$

The estimates needed to prove that the right-hand side of (3.8.6) is a continuous linear functional in $H(\Omega)$ with respect to $\mathbf{\Phi}$ follow from traditional estimates of the terms using the Hölder inequality. But we now show a sharper result; namely,

Lemma 3.8.1. F is a completely continuous operator in $H(\Omega)$.

Proof. Let $\{\mathbf{u}_n(x)\}$ be a weakly convergent sequence in $H(\Omega)$. Then it converges strongly in $(L^4(\Omega))^k$ ($k = 2$ or 3). From (3.8.6), we get

$$\begin{aligned}
\nu|(F(\mathbf{u}_m) - F(\mathbf{u}_n), \mathbf{\Phi})_{H(\Omega)}| &= \\
&= \Big| \int_\Omega \{[(\mathbf{u}_m - \mathbf{u}_n)\cdot\nabla]\mathbf{u}_m\cdot\mathbf{\Phi} - (\mathbf{u}_n\cdot\nabla)(\mathbf{u}_m - \mathbf{u}_n)\cdot\mathbf{\Phi} + \\
&\quad + [(\mathbf{u}_m - \mathbf{u}_n)\cdot\nabla]\mathbf{a}\cdot\mathbf{\Phi} + (\mathbf{a}\cdot\nabla)(\mathbf{u}_m - \mathbf{u}_n)\cdot\mathbf{\Phi}\}\, d\Omega \Big| \\
&\le M\|\mathbf{u}_m - \mathbf{u}_n\|_{L^4(\Omega)}\|\mathbf{\Phi}\|_{H(\Omega)}
\end{aligned}$$

with a constant M which does not depend on m, n, or $\mathbf{\Phi}$. Setting

$$\mathbf{\Phi} = F(\mathbf{u}_m) - F(\mathbf{u}_n)$$

in the inequality, we obtain

$$\nu\|F(\mathbf{u}_m) - F(\mathbf{u}_n)\|_{H(\Omega)} \le M\|\mathbf{u}_m - \mathbf{u}_n\|_{(L^4(\Omega))^k} \to 0$$

when $m, n \to \infty$, and so F is completely continuous. □

From Definition 3.8.1 it follows that

Lemma 3.8.2. A generalized solution of the problem under consideration in the sense of Definition 3.8.1 satisfies the operator equation

$$\mathbf{u} + F(\mathbf{u}) = 0; \tag{3.8.7}$$

conversely, a solution to (3.8.7) is a generalized solution of the problem.

By Lemma 3.7.1, it now suffices to show that all solutions of the equation $\mathbf{u}+tF(\mathbf{u}) = 0$, for all $t \in [0,1]$, lie in a sphere $\|\mathbf{u}\|_{H(\Omega)} \le R$ for some $R < \infty$. First we show this in the simpler case of homogeneous boundary condition (3.8.3). Here $\boldsymbol{\alpha} = 0$ and thus $\mathbf{a}(\mathbf{x}) = 0$. □

Theorem 3.8.1. The problem (3.8.1)–(3.8.3) with $\boldsymbol{\alpha} = 0$ has at least one generalized solution in the sense of Definition 3.8.1. Each generalized solution $\mathbf{u}(\mathbf{x})$ is bounded, $\|\mathbf{u}\|_{H(\Omega)} < R$ for some $R < \infty$ and the degree of $I + F$ with respect to 0 and $D = \{\mathbf{u} \in H(\Omega) \mid \|\mathbf{u}\| < R\}$ is +1.

Proof. As was said, it suffices to show an a priori estimate for solutions to the equation $\mathbf{u} + tF(\mathbf{u}) = 0$ for $t \in [0,1]$. For a solution, there holds the identity

$$(\mathbf{u} + tF(\mathbf{u}), \mathbf{u})_{H(\Omega)} = 0$$

or, the same,

$$\nu(\mathbf{u},\mathbf{u})_{H(\Omega)} + t\int_\Omega (\mathbf{u}\cdot\nabla)\mathbf{u}\cdot\mathbf{u}\,d\Omega = -t\int_\Omega \mathbf{f}\cdot\mathbf{u}\,d\Omega.$$

Integration by parts gives

$$\begin{aligned}\int_\Omega (\mathbf{u}\cdot\nabla)\mathbf{u}\cdot\mathbf{u}\,d\Omega &= \frac{1}{2}\int_\Omega \sum_k u_k \frac{\partial}{\partial x_k}(\mathbf{u}\cdot\mathbf{u})\,d\Omega \\ &= -\frac{1}{2}\int_\Omega (\mathbf{u}\cdot\mathbf{u})(\nabla\cdot\mathbf{u})\,d\Omega = 0 \end{aligned} \tag{3.8.8}$$

since $\nabla\cdot\mathbf{u} = 0$ and thus, for a solution $\mathbf{u}$, we get

$$|\nu(\mathbf{u},\mathbf{u})_{H(\Omega)}| = \left|t\int_\Omega \mathbf{f}\cdot\mathbf{u}\,d\Omega\right| \le \frac{\nu R}{2}\|\mathbf{f}\|_{L^p(\Omega)}\|\mathbf{u}\|_{H(\Omega)}$$

with some constant R, or

$$\|\mathbf{u}\|_{H(\Omega)} < R.$$

This completes the proof. □

Now we consider the more complicated case of nonhomogeneous boundary conditions (3.8.3). We need some auxiliary results.

Let ω_ε be a domain in $\overline{\Omega}$ which consists of points covered by all inward normals to $\partial\Omega$ of the length ε. For sufficiently small $\varepsilon > 0$, these normals

do not intersect and thus in ω_ε we can use a coordinate system pointing out for a $\mathbf{x} \in \omega_\varepsilon$ a point Q on $\partial\Omega$ and a number s, the distance from Q to $\mathbf{x}$ along the corresponding normal. So for a function $g(\mathbf{x})$ given on ω_ε, we write down $g(s, Q)$.

Lemma 3.8.3. There is a solenoidal vector function $\mathbf{a}_\varepsilon(\mathbf{x}) \in (C^{(1)}(\overline{\Omega}))^k$ such that $\mathbf{a}_\varepsilon(\mathbf{x}) = 0$ in $\Omega \setminus \omega_\varepsilon$,

$$\mathbf{a}_\varepsilon(\mathbf{x})\big|_{\partial\Omega} = \boldsymbol{\alpha}, \qquad \text{and} \qquad |\mathbf{a}_\varepsilon(\mathbf{x})| \le M_1/\varepsilon \text{ in } \overline{\Omega} \tag{3.8.9}$$

with a constant M_1 not depending on ε.

Proof. Let us introduce a function $q(\mathbf{x})$ by

$$q(s, Q) = \begin{cases} (\varepsilon^2 - s^2)^2/\varepsilon^4, & 0 \le s \le \varepsilon, \\ 0, & s > \varepsilon. \end{cases}$$

Let $\mathbf{a}(\mathbf{x})$ be a solenoidal vector-function satisfying the assumption (ii) of the beginning of the section. Under the taken assumptions, there is a vector-function $\mathbf{p}(\mathbf{x})$ such that

$$\mathbf{a}(\mathbf{x}) = \operatorname{rot} \mathbf{p}(\mathbf{x}).$$

It is seen that the vector function $\mathbf{a}_\varepsilon(\mathbf{x}) = \operatorname{rot}(q\mathbf{p})$ is needed.

Note that in the plane case, this is a vector $(0, 0, q\psi)$ where $\psi(x_1, x_2)$ is the flow function of $\mathbf{a}(\mathbf{x})$. □

Lemma 3.8.4. For $\mathbf{u} \in H(\Omega)$, we have

$$\int_{\omega_\varepsilon} |\mathbf{u}|^2 \, d\Omega \le M_2^2 \varepsilon^2 \int_{\omega_\varepsilon} \sum_{i,j} \left| \frac{\partial u_i}{\partial x_j} \right|^2 d\Omega \tag{3.8.10}$$

with a constant M_2 not depending on $\mathbf{u}$ or ε.

Proof. We show (3.8.10) for a smooth function. The limit passage will prove the general case. So for points of ω_ε we have

$$\mathbf{u}(s, Q) = \int_0^s \frac{\partial \mathbf{u}(t, Q)}{\partial t} \, dt.$$

By the Cauchy inequality

$$\begin{aligned} \int_0^\varepsilon |\mathbf{u}(t, Q)|^2 \, dt &= \int_0^\varepsilon \left| \int_0^s \frac{\partial \mathbf{u}(t, Q)}{\partial t} \, dt \right|^2 ds \\ &\le \int_0^\varepsilon s \int_0^s \left| \frac{\partial \mathbf{u}(t, Q)}{\partial t} \right|^2 dt \, ds \\ &\le \frac{\varepsilon^2}{2} \int_0^\varepsilon \left| \frac{\partial \mathbf{u}(t, Q)}{\partial t} \right|^2 dt. \end{aligned}$$

It is easily seen that for any $g(\mathbf{x})$

$$m_1 \int_0^\varepsilon \int_{\partial\Omega} g^2(s,Q)\,ds\,dS \le \int_{\omega_\varepsilon} g^2\,d\Omega \le m_2 \int_0^\varepsilon \int_{\partial\Omega} g^2(s,Q)\,ds\,dS$$

and so

$$\begin{aligned}\int_{\omega_\varepsilon} |\mathbf{u}|^2\,d\Omega &\le m_2 \int_{\partial\Omega}\int_0^\varepsilon |\mathbf{u}(s,Q)|^2\,ds\,dS \\ &\le m_2 \int_{\partial\Omega} \frac{\varepsilon^2}{2} \int_0^\varepsilon \left|\frac{\partial \mathbf{u}}{\partial t}\right|^2 dt\,dS \\ &\le \frac{m_2}{2m_1}\varepsilon^2 \int_{\omega_\varepsilon} \sum_{i,j} \left|\frac{\partial u_i}{\partial x_j}\right|^2 d\Omega.\end{aligned}$$

□

To apply degree theory to the problem under consideration, it remains to establish

Lemma 3.8.5. All solutions of the equation

$$\mathbf{u} + tF(\mathbf{u}) = 0 \tag{3.8.11}$$

for all $t \in [0,1]$, are in a ball $\|\mathbf{u}\|_{H(\Omega)} < R$ whose radius R depends only on $\mathbf{f}$, $\partial\Omega$, $\mathbf{a}$, and ν.

Proof. Suppose that the set of solutions to (3.8.11) is unbounded. This means there is a sequence $\{t_k\} \subset [0,1]$ and a corresponding sequence $\{\mathbf{u}_k\}$ such that $\mathbf{u}_k + t_k F(\mathbf{u}_k) = 0$ and

$$\|\mathbf{u}_k\|_{H(\Omega)} \to \infty \text{ as } k \to \infty. \tag{3.8.12}$$

Without loss of generality, we can consider $\{t_k\}$ to be convergent to $t_0 \in [0,1]$ and, moreover, the sequence $\{\mathbf{u}_k^*\}$, $\mathbf{u}_k^* = \mathbf{u}_k/\|\mathbf{u}_k\|_{H(\Omega)}$, to be weakly convergent to an element $\mathbf{u}_0 \in H(\Omega)$ since $\{\mathbf{u}_k^*\}$ is bounded.

Let us consider the identity $(\mathbf{u}_k + t_k F(\mathbf{u}_k), \mathbf{u}_k) = 0$, namely,

$$\begin{aligned}-\nu\|\mathbf{u}_k\|^2_{H(\Omega)} = t_k \int_{\omega_\varepsilon} (\mathbf{a}_\varepsilon \cdot \nabla)\mathbf{u}_k \cdot \mathbf{u}_k\,d\Omega + \\ + t_k \int_\Omega [(\mathbf{a}_\varepsilon \cdot \nabla)\mathbf{a}_\varepsilon \cdot \mathbf{u}_k + \nu \operatorname{rot}\mathbf{a}_\varepsilon \cdot \operatorname{rot}\mathbf{u}_k + \mathbf{f}\cdot\mathbf{u}_k]\,d\Omega\end{aligned} \tag{3.8.13}$$

which is valid because of (3.8.8) and a similar equality

$$\int_{\omega_\varepsilon} (\mathbf{u}_k \cdot \nabla)\mathbf{a}_\varepsilon \cdot \mathbf{u}_k\,d\Omega = 0.$$

The first integral on the right-hand side of (3.8.13) is a weakly continuous functional with respect to $\mathbf{u}_k$, and for the second integral we have

$$\left| \int_\Omega [(\mathbf{a}_\varepsilon \cdot \nabla)\mathbf{a}_\varepsilon \cdot \mathbf{u}_k + \nu \operatorname{rot} \mathbf{a}_\varepsilon \cdot \operatorname{rot} \mathbf{u}_k + \mathbf{f} \cdot \mathbf{u}_k]\, d\Omega \right| \le M_3 \|\mathbf{u}_k\|_{H(\Omega)}$$

where M_3 does not depend on $\mathbf{u}_k$. Dividing both sides of (3.8.13) by $\|\mathbf{u}_k\|^2_{H(\Omega)}$, it follows that

$$-\nu = t_0 \int_{\omega_\varepsilon} (\mathbf{a}_\varepsilon \cdot \nabla)\mathbf{u}_0 \cdot \mathbf{u}_0 \, d\Omega. \tag{3.8.14}$$

We note that this holds for any small positive $\varepsilon < \varepsilon_0$ with a fixed ε_0 for which the above construction of the frame for ω_{ε_0} is valid. To prove it, take $\varepsilon = \eta < \varepsilon_0$

$$\mathbf{w}_k = \mathbf{u}_k + \mathbf{a}_{\varepsilon_0} - \mathbf{a}_\eta$$

and consider the identity

$$(\mathbf{u}_k + t_k F(\mathbf{u}_k), \mathbf{w}_k)_{H(\Omega)} = 0$$

which takes the form

$$\begin{aligned} -\nu \|\mathbf{w}_k\|^2_{H(\Omega)} = t_k \int_{\omega_\eta} (\mathbf{a}_\eta \cdot \nabla)\mathbf{w}_k \cdot \mathbf{w}_k \, d\Omega + \\ + t_k \int_\Omega [(\mathbf{a}_\eta \cdot \nabla)\mathbf{a}_\eta \cdot \mathbf{w}_k + \nu \operatorname{rot} \mathbf{a}_\eta \cdot \operatorname{rot} \mathbf{w}_k + \mathbf{f} \cdot \mathbf{w}_k]\, d\Omega. \end{aligned}$$

Divide this equality by $\|\mathbf{u}_k\|^2_{H(\Omega)}$ term by term. Consider the sequence

$$\mathbf{w}^*_k = \mathbf{u}^*_k + (\mathbf{a}_\varepsilon - \mathbf{a}_\eta)/\|\mathbf{u}_k\|_{H(\Omega)}.$$

Since $\|\mathbf{u}_k\|_{H(\Omega)} \to \infty$, we have $(\mathbf{a}_\varepsilon - \mathbf{a}_\eta)/\|\mathbf{u}_k\|_{H(\Omega)} \to 0$ strongly. Since $\|\mathbf{u}^*_k\|_{H(\Omega)} = 1$, we have that $\|\mathbf{w}^*_k\|_{H(\Omega)} \to 1$. Besides, it is clear that $\mathbf{w}^*_k \to \mathbf{u}_0$ weakly and thus we get the needed equality (3.8.14) again.

Now we show that the limit of the integral on the right-hand side of (3.8.14) is zero. Thanks to (3.8.9) and (3.8.10), we obtain

$$\left| \int_{\omega_\varepsilon} (\mathbf{a}_\varepsilon \cdot \nabla)\mathbf{u}_0 \cdot \mathbf{u}_0 \, d\Omega \right| \le M_1 M_2 \int_{\omega_\varepsilon} |\operatorname{rot} \mathbf{u}_0|^2 \, d\Omega \to 0$$

as $\varepsilon \to 0$. Since $\nu > 0$, we have a contradiction which completes the proof. □

Now we can formulate

Theorem 3.8.2. Under assumptions (i)–(iv), there exists at least one generalized solution of the problem (3.8.1)–(3.8.3) in the sense of Definition 3.8.1. All generalized solutions of the problem are bounded in the energy space and the degree of the operator $I + F$ of the problem with respect to zero and a ball about zero with sufficiently large radius is $+1$.

Problem 3.8.1. Check which of the assumptions (i)–(iv) are not necessary in proving Theorem 3.8.1.

Appendix
Hints for Selected Problems

Problem 1.1.1 (page 8)

Setting $\mathbf{y} = \mathbf{x}$ in the triangle inequality we obtain $d(\mathbf{x}, \mathbf{x}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{x})$. By axioms D2 and D3 then, we have $0 \leq 2d(\mathbf{x}, \mathbf{z})$.

Problem 1.1.2 (page 8)

The metrics d_S and d_p are equivalent with

$$1 \leq \frac{d_p(\mathbf{x}, \mathbf{y})}{d_S(\mathbf{x}, \mathbf{y})} \leq N^{1/p},$$

because

$$\begin{aligned}
\max_{1 \leq i \leq n} |x_i - y_i| &= \left(\max_{1 \leq i \leq n} |x_i - y_i|^p \right)^{1/p} \\
&\leq \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} \\
&\leq \left(\sum_{i=1}^{n} \left(\max_{1 \leq j \leq n} |x_j - y_j| \right)^p \right)^{1/p} \\
&= N^{1/p} \max_{1 \leq j \leq n} |x_j - y_j|.
\end{aligned}$$

The metrics d_E and d_k are equivalent with

$$\left(\min_{1\le i\le n} k_i\right)^{1/2} \le \frac{d_k(\mathbf{x},\mathbf{y})}{d_E(\mathbf{x},\mathbf{y})} \le \left(\max_{1\le i\le n} k_i\right)^{1/2}.$$

Problem 1.1.3 (page 11)

Axioms D1–D3 are obviously fulfilled in each case. For the expression (1.1.3), axiom D4 holds because

$$|x_i - y_i| \le |x_i - z_i| + |z_i - y_i|$$

for each i, and therefore

$$\begin{aligned} d(\mathbf{x},\mathbf{y}) &= \sup_i |x_i - y_i| \\ &\le \sup_i (|x_i - z_i| + |z_i - y_i|) \\ &\le \sup_i |x_i - z_i| + \sup_i |z_i - y_i| \\ &= d(\mathbf{x},\mathbf{z}) + d(\mathbf{z},\mathbf{y}). \end{aligned}$$

To verify D4 for the expressions (1.1.4) and (1.1.5), we need the Minkowski inequality

$$\left(\sum_{i=1}^{\infty} |a_i + b_i|^p\right)^{1/p} \le \left(\sum_{i=1}^{\infty} |a_i|^p\right)^{1/p} + \left(\sum_{i=1}^{\infty} |b_i|^p\right)^{1/p}.$$

For (1.1.4) we have, starting with the triangle inequality,

$$\begin{aligned} d(\mathbf{x},\mathbf{y}) &= \left(\sum_{i=1}^{\infty} |x_i - z_i + z_i - y_i|^p\right)^{1/p} \\ &\le \left(\sum_{i=1}^{\infty} [|x_i - z_i| + |z_i - y_i|]^p\right)^{1/p} \\ &\le \left(\sum_{i=1}^{\infty} |x_i - z_i|^p\right)^{1/p} + \left(\sum_{i=1}^{\infty} |z_i - y_i|^p\right)^{1/p} \\ &= d(\mathbf{x},\mathbf{z}) + d(\mathbf{z},\mathbf{y}). \end{aligned}$$

For (1.1.5) we have

$$\begin{aligned} d(\mathbf{x},\mathbf{y}) &= \left(\sum_{k=1}^{\infty} k^2 |x_k - z_k + z_k - y_k|^2\right)^{1/2} \\ &\le \left(\sum_{k=1}^{\infty} k^2 \left(|x_k - z_k| + |z_k - y_k|\right)^2\right)^{1/2} \\ &= \left(\sum_{k=1}^{\infty} \left(k|x_k - z_k| + k|z_k - y_k|\right)^2\right)^{1/2} \\ &\le \left(\sum_{k=1}^{\infty} (k|x_k - z_k|)^2\right)^{1/2} + \left(\sum_{k=1}^{\infty} (k|z_k - y_k|)^2\right)^{1/2} \\ &= d(\mathbf{x},\mathbf{z}) + d(\mathbf{z},\mathbf{y}). \end{aligned}$$

Problem 1.2.1 (page 13)

The only aspect of D1–D3 worthy of close examination is the implication

$$d(f,g) = 0 \implies f = g$$

of D2. Note that $d(f,g) = 0$ implies

$$\max_{\mathbf{x}\in\Omega} |D^\alpha f(\mathbf{x}) - D^\alpha g(\mathbf{x})| = 0 \text{ for all } \alpha \text{ such that } |\alpha| \le k.$$

In particular this holds for $\alpha = (0,\ldots,0)$, giving

$$\max_{\mathbf{x}\in\Omega} |f(\mathbf{x}) - g(\mathbf{x})| = 0.$$

This implies that $f(\mathbf{x}) = g(\mathbf{x})$ for all $\mathbf{x} \in \Omega$. Fulfillment of D4 follows from the triangle inequality

$$|D^\alpha f(\mathbf{x}) - D^\alpha g(\mathbf{x})| \le |D^\alpha f(\mathbf{x}) - D^\alpha h(\mathbf{x})| + |D^\alpha h(\mathbf{x}) - D^\alpha g(\mathbf{x})|$$

for real numbers.

Problem 1.2.2 (page 14)

The constant functions $f(x) \equiv 0$ and $g(x) \equiv 1$ on $[0,1]$ are not equal, but the proposed distance function would give $d(f,g) = 0$. Hence d is not a metric (cf., axiom D2).

To generate a metric space under this metric, we could narrow our consideration to the functions satisfying a condition such as $f(0) = 0$.

Problem 1.3.1 (page 15)

Requested here is a general setup of the problem; we cannot, in the narrow sense, "find" a solution. Rather, we can investigate the setup of the variational problem in the form of a boundary value problem for a partial differential equation. Of course, from a logical viewpoint the initial setup as a problem of minimum is neither better nor worse than the latter one, since in both cases we cannot find explicit solutions in general. Historically, however, the theory of partial differential equations was well developed and variational problems were always reduced to the solution of corresponding boundary value problems.

Let us suppose that u has more smoothness than is required by the problem, namely that u belongs to $C^{(2)}$. We employ the usual methods of the calculus of variations. In the functional $J(u)$ we consider variations $u = u(x,y) + t\varphi(x,y)$ where t is a real variable and $\varphi|_{\partial\Omega} = 0$. For fixed φ the functional $J(u + t\varphi)$ can be regarded as an ordinary function of t, taking its minimum at $t = 0$. It is necessary that

$$\begin{aligned} 0 &= \frac{d}{dt} J(u + t\varphi)\Big|_{t=0} \\ &= \iint_\Omega \left[\left(\frac{\partial u}{\partial x} + t\frac{\partial \varphi}{\partial x} \right) \frac{\partial \varphi}{\partial x} + \left(\frac{\partial u}{\partial y} + t\frac{\partial \varphi}{\partial y} \right) \frac{\partial \varphi}{\partial y} - f\varphi \right] \Bigg|_{t=0} dx\, dy \\ &= \iint_\Omega \left[\frac{\partial u}{\partial x}\frac{\partial \varphi}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial \varphi}{\partial y} - f\varphi \right] dx\, dy. \end{aligned}$$

We now integrate by parts using the formula

$$\iint_\Omega u \frac{\partial v}{\partial x_i}\, dx\, dy = -\iint_\Omega \frac{\partial u}{\partial x_i} v\, dx\, dy + \oint_{\partial\Omega} uv\, n_i\, ds,$$

where s parameterizes $\partial\Omega$ and n_i is the cosine of the angle between the outward normal $\mathbf{n}$ to Ω and the x_i axis ($x_i = x, y$ for $i = 1, 2$, respectively). This gives us

$$-\iint_\Omega \left[\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + f \right] \varphi\, dx\, dy + \oint_{\partial\Omega} \left[\frac{\partial u}{\partial x}\, n_x + \frac{\partial u}{\partial y}\, n_y \right] \varphi\, ds = 0.$$

Because $\varphi|_{\partial\Omega} = 0$, we have

$$\iint_\Omega \left[\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + f \right] \varphi\, dx\, dy = 0$$

for any "admissible" φ. It follows (by the "fundamental lemma") that

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + f = 0.$$

This is Poisson's equation for the unknown u. It can be argued that this equation continues to hold even when the condition $\varphi|_{\partial S} = 0$ is removed, allowing us to write

$$\oint_{\partial\Omega} \left[\frac{\partial u}{\partial x}\, n_x + \frac{\partial u}{\partial y}\, n_y\right] \varphi\, ds = 0$$

and thereby deduce the natural boundary condition

$$\left[\frac{\partial u}{\partial x}\, n_x + \frac{\partial u}{\partial y}\, n_y\right]\Big|_{\partial\Omega} = \frac{\partial u}{\partial n}\Big|_{\partial\Omega} = 0.$$

Problem 1.5.1 (page 19)

The functions

$$f_n(x) = \begin{cases} 0, & 0 \le x \le \frac{1}{2}, \\ nx - \frac{n}{2}, & \frac{1}{2} \le x \le \frac{1}{2} + \frac{1}{n}, \\ 1 & \frac{1}{2} + \frac{1}{n} \le x \le 1, \end{cases} \qquad (n = 2, 3, 4, \ldots)$$

are each continuous on $[0, 1]$. To see that $\{f_n\}$ is a Cauchy sequence, we assume $m > n$ and calculate

$$\begin{aligned} d(f_n, f_m) &= \int_{\frac{1}{2}}^{\frac{1}{2}+\frac{1}{m}} \left|\left(mx - \frac{m}{2}\right) - \left(nx - \frac{n}{2}\right)\right| dx + \\ &\quad + \int_{\frac{1}{2}+\frac{1}{m}}^{\frac{1}{2}+\frac{1}{n}} \left|1 - \left(nx - \frac{n}{2}\right)\right| dx \\ &= \frac{1}{2}\left(\frac{1}{n} - \frac{1}{m}\right) \to 0 \text{ as } m, n \to \infty. \end{aligned}$$

However, we have $f_n \to f$ where

$$f = \begin{cases} 0, & 0 \le x \le \frac{1}{2}, \\ 1, & \frac{1}{2} < x \le 1, \end{cases}$$

because

$$d(f_n, f) = \int_{\frac{1}{2}}^{\frac{1}{2}+\frac{1}{n}} \left|1 - \left(nx - \frac{n}{2}\right)\right| dx = \frac{1}{2n} \to 0 \text{ as } n \to \infty.$$

Problem 1.9.1 (page 28)

We can show that any norm is equivalent to the Euclidean norm $\|\cdot\|_e$. Take any basis i_k that is orthonormal in the Euclidean inner product. We can

express any x as $x = \sum_{k=1}^{n} c_k i_k$. Then

$$\|x\|_e = \left(\sum_{k=1}^{n} c_k^2 \right)^{1/2}.$$

For an arbitrary norm $\|\cdot\|$ we have

$$\|x\| = \left\| \sum_{k=1}^{n} c_k i_k \right\| \le \sum_{k=1}^{n} |c_k| \|i_k\| \le \sum_{k=1}^{n} \left(\sum_{j=1}^{n} |c_j|^2 \right)^{1/2} \|i_k\| = m\|x\|_e$$

where $m = \sum_{k=1}^{n} \|i_k\|$ is finite. So one side is proved. For the other side, consider $\|x\|$ as a function of the n variables c_k. Because of the above inequality it is a continuous function in the usual sense. Indeed

$$|\|x_1\| - \|x_2\|| \le \|x_1 - x_2\| \le m\|x_1 - x_2\|_e.$$

It is enough to show that on the sphere $\|x\|_e = 1$ we have $\inf \|x\| = a > 0$ (because of homogeneity of norms). Being a continuous function, $\|x\|$ achieves its minimum on the compact set $\|x\|_e = 1$ at a point x_0. So $\|x_0\| = a$. If $a = 0$ then $x_0 = 0$ and thus x_0 does not belong to the unit sphere. Thus $a > 0$ and for any x, $\|x\|/\|x\|_e \ge a$.

Problem 1.9.2 (page 29)

By N3 with x replaced by $x - y$, we have $\|x\| - \|y\| \le \|x - y\|$. Swapping x and y we have, by N2, $\|y\| - \|x\| \le \|y - x\| = \|(-1)(x - y)\| = \|x - y\|$. Therefore $\|x - y\| \ge |\|x\| - \|y\||$, as desired.

Problem 1.9.3 (page 31)

$$\begin{aligned}
\|x+y\|^2 + \|x-y\|^2 &= (x+y, x+y) + (x-y, x-y) \\
&= (x, x+y) + (y, x+y) + (x, x-y) - (y, x-y) \\
&= \overline{(x+y, x)} + \overline{(x+y, y)} + \overline{(x-y, x)} - \overline{(x-y, y)} \\
&= \overline{(x, x)} + \overline{(y, x)} + \overline{(x, y)} + \overline{(y, y)} + \\
&\quad + \overline{(x, x)} - \overline{(y, x)} - \overline{(x, y)} + \overline{(y, y)} \\
&= 2(x, x) + 2(y, y) \\
&= 2\|x\|^2 + 2\|y\|^2.
\end{aligned}$$

Problem 1.10.1 (page 44)

$$\iint_\Omega F(x,y)\,dx\,dy + \sum_k F_k(x_k,y_k) + \int_\gamma F_l(x,y)\,ds = 0,$$
$$\iint_\Omega xF(x,y)\,dx\,dy + \sum_k x_k F_k(x_k,y_k) + \int_\gamma xF_l(x,y)\,ds = 0,$$
$$\iint_\Omega yF(x,y)\,dx\,dy + \sum_k y_k F_k(x_k,y_k) + \int_\gamma yF_l(x,y)\,ds = 0.$$

Problem 1.13.1 (page 54)

Let $m \to \infty$ in the inequality $d(x_n, x_{n+m}) \le q^n d(x_0, x_m)$. The result is less useful than (1.13.4) because the right member involves the unknown quantity x_*.

Problem 1.13.2 (page 54)

We show that A^N is a contraction for some N if A acts in $C[0,T]$ and is given by

$$Ay(t) = \int_0^t g(t-\tau)y(\tau)\,d\tau.$$

Let M be the maximum value attained by $g(t)$ on $[0,T]$. For any $t \in [0,T]$ we have

$$\begin{aligned}
|Ay_2(t) - Ay_1(t)| &\le \int_0^t |g(t-\tau)|\,|y_2(\tau) - y_1(\tau)|\,d\tau \\
&\le \max_{\tau\in[0,t]} |g(t-\tau)| \max_{\tau\in[0,t]} |y_2(\tau) - y_1(\tau)| \int_0^t d\tau \\
&\le \max_{\tau\in[0,T]} |g(t-\tau)| \max_{\tau\in[0,T]} |y_2(\tau) - y_1(\tau)|\,t \\
&= M\,t\,d(y_2,y_1).
\end{aligned}$$

Then

$$\begin{aligned}
|A^2y_2(t) - A^2y_1(t)| &\le \int_0^t |g(t-\tau)|\,|Ay_2(\tau) - Ay_1(\tau)|\,d\tau \\
&\le M \cdot M\,d(y_2,y_1) \int_0^t \tau\,d\tau \\
&= M^2 \frac{t^2}{1\cdot 2}\,d(y_2,y_1).
\end{aligned}$$

Continuing in this manner we can show that

$$|A^k y_2(t) - A^k y_1(t)| \le M^k \frac{t^k}{k!}\,d(y_2,y_1)$$

for any positive integer k and any $t \in [0,T]$. Thus

$$\begin{aligned} d(A^k y_2\,,\, A^k y_1) &= \max_{t\in[0,T]} |A^k y_2(t) - A^k y_1(t)| \\ &\le \max_{t\in[0,T]} M^k \frac{t^k}{k!}\, d(y_2, y_1) \\ &= \frac{(MT)^k}{k!}\, d(y_2, y_1). \end{aligned}$$

Finally, we choose N so large that $(MT)^N/N! < 1$.

Problem 1.15.1 (page 63)

Fix $n \in \mathbb{N}$ and let P_r^n denote the set of all polynomials of degree n having rational coefficients. Denote by $\mathbb{Q}$ the set of all rational numbers. The set P_r^n can be put into one-to-one correspondence with the countable set

$$\mathbb{Q}^{n+1} = \underbrace{\mathbb{Q} \times \mathbb{Q} \times \cdots \mathbb{Q}}_{n+1 \text{ times}}.$$

The set P_r of all polynomials having rational coefficients is given by

$$P_r = \bigcup_{n=0}^{\infty} P_r^n$$

and this is a countable union of countable sets.

Problem 1.17.1 (page 71)

Yes. M bounded in $C(\Omega)$ means

$$\exists R \text{ such that } \forall f \in M, \quad \|f\|_{C(\Omega)} = \max_{\mathbf{x}\in\Omega} |f(\mathbf{x})| \le R.$$

Based on (i) we can assert this with $R = c$.

Problem 1.22.2 (page 96)

Let x_k, $k = 1, \ldots, n$, be orthonormal. Then

$$\sum_{k=1}^{n} \alpha_k x_k = 0 \implies \left(\sum_{k=1}^{n} \alpha_k x_k, x_j \right) = \sum_{k=1}^{n} \alpha_k (x_k, x_j) = \alpha_j = 0$$

for $j = 1, \ldots, n$.

Problem 2.1.2 (page 123)

We have

$$\begin{aligned}\|A_nB_n - AB\| &= \|A_nB_n - A_nB + A_nB - AB\| \\ &= \|A_n(B_n - B) + (A_n - A)B\| \\ &\le \|A_n(B_n - B)\| + \|(A_n - A)B\| \\ &\le \|A_n\| \cdot \|B_n - B\| + \|A_n - A\| \cdot \|B\|\end{aligned}$$

where $\|A_n\|$ is bounded since A_n is convergent.

Problem 2.4.1 (page 132)

Let A map an element x from the space $(X, \|\cdot\|_2)$ into the same element x regarded as an element of the space $(X, \|\cdot\|_1)$. This operator is linear and, by hypothesis ($\|x\|_1 \le c_1\|x\|_2$) it is bounded (continuous), hence it is closed. It is also one-to-one and onto. By Theorem 2.4.4, A^{-1} is continuous on $(X, \|\cdot\|_1)$; this gives the inequality $\|x\|_2 \le c_2\|x\|_1$, as desired.

Problem 2.6.1 (page 140)

In order to invoke Arzelà's theorem we first show that B takes $L^2(0,1)$ into $C(0,1)$. We have

$$\begin{aligned}|(Bf)(t) - (Bf)(t_0)| &\le \int_0^1 |K(t,s) - K(t_0,s)|\,|f(s)|\,ds \\ &\le \max_{s\in[0,1]} |K(t,s) - K(t_0,s)| \int_0^1 |f(s)|\,ds \\ &\le \max_{s\in[0,1]} |K(t,s) - K(t_0,s)| \|f\|_{L^2(0,1)}\end{aligned}$$

by application of Schwarz's inequality in the form

$$\int_0^1 1 \cdot |f(s)|\,ds \le \left(\int_0^1 1^2\,ds\right)^{1/2} \left(\int_0^1 |f(s)|^2\,ds\right)^{1/2} = \|f\|_{L^2(0,1)}.$$

By continuity of K we can make $|(Bf)(t) - (Bf)(t_0)|$ as small as desired for sufficiently small $|t - t_0|$, uniformly with respect to s.

Following the argument in the text, we show that B takes the unit ball of $L^2(0,1)$ into a precompact subset S of $C(0,1)$. First S is uniformly bounded: by the inequality displayed above we have

$$\|(Bf)(t)\|_{C(0,1)} \le \max_{t\in[0,1]} \int_0^1 |K(t,s)|\,|f(s)|\,ds \le M\|f\|_{L^2(0,1)},$$

where

$$M = \max_{t,s\in[0,1]} |K(t,s)|.$$

Equicontinuity follows from the inequality

$$|(Bf)(t+\delta) - (Bf)(t)| \le \max_{s\in[0,1]} |K(t+\delta,s) - K(t,s)|$$

on the unit ball in $L^2(0,1)$, and the uniform continuity of K.

Finally, we observe that a precompact set in $C(0,1)$ is precompact in $L^2(0,1)$. Indeed, if S is precompact in $C(0,1)$ then every sequence $\{f_j\} \subset S$ contains a Cauchy subsequence $\{f_{j_k}\}$: for every $\varepsilon > 0$ there exists N such that

$$\|f_{j_m} - f_{j_n}\|_{C(0,1)} = \max_{x\in[0,1]} |f_{j_m}(x) - f_{j_n}(x)| < \varepsilon$$

for $m, n > N$. Then $\{f_{j_k}\}$ is also a Cauchy sequence in $L^2(0,1)$:

$$\|f_{j_m} - f_{j_n}\|_{L^2(0,1)} = \left(\int_0^1 |f_{j_m}(x) - f_{j_n}(x)|^2 \, dx \right)^{1/2} \le \varepsilon$$

for $m, n > N$.

Problem 3.1.1 (page 178)

Given $\mathbf{f}\colon \mathbb{R}^m \to \mathbb{R}^n$, we wish to examine the difference

$$\mathbf{f}(\mathbf{x}_0 + \mathbf{h}) - \mathbf{f}(\mathbf{x}_0).$$

Let us introduce the standard orthonormal bases of $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively:

$$\tilde{\mathbf{e}}_1, \ldots, \tilde{\mathbf{e}}_m, \qquad\qquad \mathbf{e}_1, \ldots, \mathbf{e}_n.$$

Then

$$\mathbf{f}(\mathbf{x}) = \sum_{i=1}^{n} f_i(\mathbf{x})\, \mathbf{e}_i$$

and we have

$$\mathbf{f}(\mathbf{x}_0 + \mathbf{h}) - \mathbf{f}(\mathbf{x}_0) = \sum_{i=1}^{n} [f_i(\mathbf{x}_0 + \mathbf{h}) - f_i(\mathbf{x}_0)]\, \mathbf{e}_i.$$

But Taylor expansion to first order gives, for each i,

$$f_i(\mathbf{x}_0 + \mathbf{h}) - f_i(\mathbf{x}_0) = \sum_{j=1}^{m} \frac{\partial f_i(\mathbf{x}_0)}{\partial x_j} h_j + o(\|\mathbf{h}\|),$$

where the h_j are the components of $\mathbf{h}$:

$$\mathbf{h} = \sum_{j=1}^{m} h_j \tilde{\mathbf{e}}_j.$$

So we identify

$$\mathbf{f}'(\mathbf{x}_0)(\mathbf{h}) = \sum_{i=1}^{n} \mathbf{e}_i \sum_{j=1}^{m} \frac{\partial f_i(\mathbf{x}_0)}{\partial x_j} h_j,$$

and observe that the right-hand side is represented in matrix-vector notation as

$$\begin{pmatrix} \frac{\partial f_1(\mathbf{x}_0)}{\partial x_1} h_1 + \cdots + \frac{\partial f_1(\mathbf{x}_0)}{\partial x_m} h_m \\ \vdots \\ \frac{\partial f_n(\mathbf{x}_0)}{\partial x_1} h_1 + \cdots + \frac{\partial f_n(\mathbf{x}_0)}{\partial x_m} h_m \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1(\mathbf{x}_0)}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x}_0)}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n(\mathbf{x}_0)}{\partial x_1} & \cdots & \frac{\partial f_n(\mathbf{x}_0)}{\partial x_m} \end{pmatrix} \begin{pmatrix} h_1 \\ \vdots \\ h_m \end{pmatrix}.$$

This page intentionally left blank

References

[1] Adams, R.A. *Sobolev Spaces.* Academic Press, New York, 1975.

[2] Antman, S.S. The influence of elasticity on analysis: modern developments, Bull. Amer. Math. Soc. (New Series), 1983, 9, 267–291.

[3] Antman, S.S. *Nonlinear Problems of Elasticity.* Springer–Verlag, New York, 1996.

[4] Banach, S. *Théories des opérations linéaires.* Chelsea Publishing Company, New York, 1978.

[5] Bramble, J.H., and Hilbert, S.R. Bounds for a class of linear functionals with applications to Hermite interpolation. Numer. Math., 1971, 16, 362–369.

[6] Ciarlet, P.G. *The Finite Element Method for Elliptic Problems.* North Holland Publ. Company, 1978.

[7] Ciarlet, P.G. *Mathematical Elasticity*, vol. 1–3. North Holland, 1988–2000.

[8] Courant, R., and Hilbert, D. *Methods of Mathematical Physics.* Interscience Publishers, New York, 1953–62.

[9] Fichera, G. Existence theorems in elasticity (XIII.15), and Boundary value problems of elasticity with unilateral constraints (YII.8, XIII.15, XIII.6), in Handbuch der Physik YIa/2, C. Truesdell, ed., Springer–Verlag, 1972.

[10] Friedrichs, K.O. The identity of weak and strong extensions of differential operators. Trans. Amer. Math. Soc., 1944, vol. 55, pp. 132–151.

[11] Gokhberg, I.Ts, and Krejn, M.G. *Theory of the Volterra Operators in Hilbert Space and Its Applications*. Nauka, Moscow, 1967.

[12] Hardy, G.H., Littlewood, J.E., and Pólya, G. *Inequalities*. Cambridge University Press, 1952.

[13] Il'yushin, A.A. *Plasticity*. Gostekhizfat, Moscow, 1948 (in Russian).

[14] Lebedev, L.P., Vorovich, I.I., and Gladwell, G.M.L. *Functional Analysis: Applications in Mechanics and Inverse Problems*. Kluwer Academic Publishers, Dordrecht, 1996.

[15] Leray, J., and Schauder, J. Topologie et équations fonctionnelles. Ann. S.E.N., 1934, 51, 45–78.

[16] Kantorovich, L.V., and Akilov, G.P. *Functional Analysis*. Pergamon, 1982.

[17] Lax, P.D., and Milgram, A.N. "Parabolic equations" in *Contributions to the Theory of Partial Differential Equations*. Princeton, 1954.

[18] Lions, J.-L., and Magenes, E. Problèmes aux Limites Non Homogènes et Applications, Tome 1. Dunod, Paris, 1968.

[19] Mikhlin, S.G. *The Problem of Minimum of a Quadratic Functional*. Holden–Day, San Francisco, 1965.

[20] Mikhlin, S.G. *Variational Methods in Mathematical Physics*. Pergamon Press, Oxford, 1964.

[21] Schwartz, J.T. *Nonlinear Functional Analysis*. Gordon and Breach Sc. Publ. Inc., 1969.

[22] Sobolev, S.L. *Some Applications of Functional Analysis to Mathematical Physics*. LGU, 1951.

[23] Struwe, M. *Variational Methods*, 2nd ed. Springer–Verlag, Berlin, 1996.

[24] Vitt, A., and Shubin, S. On tones of a membrane fixed in a finite number of points. Zhurn. Tekhn. Fiz., I (1931), no. 2–3, 163–175.

[25] Vorovich, I.I., and Krasovskij, Yu.P. On the method of elastic solutions. Doklady Akad. Nauk SSSR, 126 (1959), no. 4, 740–743.

[26] Vorovich, I.I. *Mathematical Problems of Nonlinear Theory of Shallow Shells*. Nauka, Moscow, 1989. (Translated as *Nonlinear Theory of Shallow Shells*. Springer–Verlag, New York, 1999.)

[27] Vorovich, I.I., and Yudovich, V.I. Steady flow of viscous incompressible liquid. Mat.Sbornik, 1961, vol. 53 (95), no. 4, 361–428.

[28] Vorovich, I.I. The problem of non-uniqueness and stability in the nonlinear mechanics of continuous media, Applied Mechanics. Proc. Thirteenth Intern. Congr. Theor. Appl. Mech., Springer–Verlag, 1973, 340–357.

[29] Yosida, K. *Functional Analysis*. Springer–Verlag, New York, 1965.

[30] Zeidler, E. *Nonlinear Functional Analysis and Its Applications, Parts 1–4*. Springer–Verlag, New York, 1985–1988.

This page intentionally left blank

Index