

Eberhard Zeidler

Applied Functional Analysis

Main Principles and Their Applications

With 37 Illustrations



Springer-Verlag

New York Berlin Heidelberg London Paris
Tokyo Hong Kong Barcelona Budapest

Eberhard Zeidler
Department of Mathematics
University of Leipzig
Augustusplatz 10
04109 Leipzig
Germany

Editors

J.E. Marsden
Department of
Mathematics
University of California
Berkeley, CA 94720
USA

L. Sirovich
Division of
Applied Mathematics
Brown University
Providence, RI 02912
USA

Mathematics Subject Classification (1991): 34A12, 42A16, 35J05

Library of Congress Cataloging-in-Publication Data

Zeidler, Eberhard

Applied functional analysis : main principles and their
applications / Eberhard Zeidler.

p. cm. – (Applied mathematical sciences ; v. 109)

Includes bibliographical references and index.

ISBN 0-387-94422-2

1. Functional analysis. I. Title. II. Series: Applied
mathematical sciences (Springer-Verlag New York Inc.) ; v. 109.

QA1.A647 vol. 109

[QA320]

510 s—dc20

[515'.7]

94-41480

Printed on acid-free paper.

© 1995 Springer-Verlag New York, Inc.

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer-Verlag New York, Inc., 175 Fifth Avenue, New York, NY 10010, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use of general descriptive names, trade names, trademarks, etc., in this publication, even if the former are not especially identified, is not to be taken as a sign that such names, as understood by the Trade Marks and Merchandise Marks Act, may accordingly be used freely by anyone.

Production managed by Laura Carlson; manufacturing supervised by Joe Quatela.

Photocomposed copy prepared using L^AT_EX.

Printed and bound by R.R. Donnelley & Sons, Harrisonburg, VA.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1

ISBN 0-387-94422-2 Springer-Verlag New York Berlin Heidelberg

Dedicated in gratitude to my teacher
Professor Herbert Beckert
on the occasion of his 75th birthday

Everything should be made
as simple as possible, but not simpler.

Albert Einstein



David Hilbert
(1862–1943)



Stefan Banach
(1892–1945)



John von Neumann
(1903–1957)

Preface

A theory is the more impressive,
the simpler are its premises,
the more distinct are the things it connects,
and the broader is its range of applicability.

Albert Einstein

There are two different ways of teaching mathematics, namely,

- (i) the systematic way, and
- (ii) the application-oriented way.

More precisely, by (i), I mean a systematic presentation of the material governed by the desire for mathematical perfection and completeness of the results. In contrast to (i), approach (ii) starts out from the question “*What are the most important applications?*” and then tries to answer this question as quickly as possible. Here, one walks directly on the main road and does not wander into all the nice and interesting side roads.

The present book is based on the second approach. It is addressed to undergraduate and beginning graduate students of mathematics, physics, and engineering who want to learn how functional analysis elegantly solves mathematical problems that are related to our real world and that have played an important role in the history of mathematics. The reader should sense that the theory is being developed, not simply for its own sake, but for the *effective* solution of concrete problems.

Our introduction to applied functional analysis is divided into two parts:

Part I: Applications to Mathematical Physics (AMS Vol. 108);

Part II: Main Principles and Their Applications (AMS Vol. 109).

A detailed discussion of the contents can be found in the preface to AMS Vol. 108.

As prerequisites for the present volume, we only assume that the reader is familiar with some basic facts about normed spaces as summarized in Section 1.27 of AMS Vol. 108. The most important propositions are called theorems. A list of these theorems, along with the most important definitions, can be found at the end of the book.

The presentation of material takes into account that, in general, no book is read completely from beginning to end. We hope that even a quick skimming of the text will suffice to grasp the essential contents. To this end, we recommend reading the introductions to the individual chapters, the “theorems” (without proofs), and the examples (without proofs) as well as the motivations and comments in the text, which point out the meaning of the specific results. The proofs are worked out in great detail. Grasping the individual steps in the proofs as well as their essential ideas is made easier by the careful organization. It is a truism that only a precise study of the proofs enables one to penetrate more deeply into a mathematical theory.

The book is based on lectures I have given for students of mathematics and physics at Leipzig University. The manuscript has been finished during a stay at the “Sonderforschungsbereich 256” of Bonn University and at the Max Planck Institute for Mathematics in Bonn. I would like to thank Professors Stefan Hildebrandt and Friedrich Hirzebruch for the invitations and the kind hospitality. Finally, my special thanks are due to Springer-Verlag for the harmonious collaboration.

I hope that the reader of this book enjoys getting a feel for the unity of mathematics by discovering interrelations between apparently completely different subjects.

Leipzig
Spring 1995

Eberhard Zeidler

Contents

Preface	vii
Contents of AMS Volume 108	xiii
1 The Hahn–Banach Theorem Optimization Problems	1
1.1 The Hahn–Banach Theorem	2
1.2 Applications to the Separation of Convex Sets	6
1.3 The Dual Space $C[a, b]^*$	10
1.4 Applications to the Moment Problem	13
1.5 Minimum Norm Problems and Duality Theory	15
1.6 Applications to Čebyšev Approximation	19
1.7 Applications to the Optimal Control of Rockets	20
2 Variational Principles and Weak Convergence	39
2.1 The n th Variation	43
2.2 Necessary and Sufficient Conditions for Local Extrema and the Classical Calculus of Variations	45
2.3 The Lack of Compactness in Infinite-Dimensional Banach Spaces	48
2.4 Weak Convergence	49
2.5 The Generalized Weierstrass Existence Theorem	53
2.6 Applications to the Calculus of Variations	56
2.7 Applications to Nonlinear Eigenvalue Problems	59
2.8 Reflexive Banach Spaces	61

2.9 Applications to Convex Minimum Problems and Variational Inequalities	66
2.10 Applications to Obstacle Problems in Elasticity	71
2.11 Saddle Points	72
2.12 Applications to Duality Theory	73
2.13 The von Neumann Minimax Theorem on the Existence of Saddle Points	75
2.14 Applications to Game Theory	81
2.15 The Ekeland Principle about Quasi-Minimal Points	83
2.16 Applications to a General Minimum Principle via the Palais–Smale Condition	86
2.17 Applications to the Mountain Pass Theorem	87
2.18 The Galerkin Method and Nonlinear Monotone Operators .	93
2.19 Symmetries and Conservation Laws (The Noether Theorem)	98
2.20 The Basic Ideas of Gauge Field Theory	102
2.21 Representations of Lie Algebras	107
2.22 Applications to Elementary Particles	112
3 Principles of Linear Functional Analysis	167
3.1 The Baire Theorem	169
3.2 Application to the Existence of Nondifferentiable Continuous Functions	171
3.3 The Uniform Boundedness Theorem	172
3.4 Applications to Cubature Formulas	175
3.5 The Open Mapping Theorem	178
3.6 Product Spaces	180
3.7 The Closed Graph Theorem	181
3.8 Applications to Factor Spaces	183
3.9 Applications to Direct Sums and Projections	188
3.10 Dual Operators	199
3.11 The Exactness of the Duality Functor	205
3.12 Applications to the Closed Range Theorem and to Fredholm Alternatives	210
4 The Implicit Function Theorem	225
4.1 m -Linear Bounded Operators	227
4.2 The Differential of Operators and the Fréchet Derivative .	228
4.3 Applications to Analytic Operators	233
4.4 Integration	238
4.5 Applications to the Taylor Theorem	243
4.6 Iterated Derivatives	244
4.7 The Chain Rule	247
4.8 The Implicit Function Theorem	250
4.9 Applications to Differential Equations	254
4.10 Diffeomorphisms and the Local Inverse Mapping Theorem .	258

4.11	Equivalent Maps and the Linearization Principle	260
4.12	The Local Normal Form for Nonlinear Double Splitting Maps	264
4.13	The Surjective Implicit Function Theorem	268
4.14	Applications to the Lagrange Multiplier Rule	270
5	Fredholm Operators	281
5.1	Duality for Linear Compact Operators	284
5.2	The Riesz–Schauder Theory on Hilbert Spaces	286
5.3	Applications to Integral Equations	291
5.4	Linear Fredholm Operators	292
5.5	The Riesz–Schauder Theory on Banach Spaces	295
5.6	Applications to the Spectrum of Linear Compact Operators	296
5.7	The Parametrix	298
5.8	Applications to the Perturbation of Fredholm Operators . .	300
5.9	Applications to the Product Index Theorem	301
5.10	Fredholm Alternatives via Dual Pairs	303
5.11	Applications to Integral Equations and Boundary-Value Problems	305
5.12	Bifurcation Theory	309
5.13	Applications to Nonlinear Integral Equations	313
5.14	Applications to Nonlinear Boundary-Value Problems	315
5.15	Nonlinear Fredholm Operators	317
5.16	Interpolation Inequalities	322
5.17	Applications to the Navier–Stokes Equations	329
References		371
List of Symbols		385
List of Theorems		391
List of Most Important Definitions		393
Subject Index		399

Contents of AMS Volume 108

Preface

Prologue

Contents of AMS Volume 109

1 Banach Spaces and Fixed-Point Theorems

- 1.1 Linear Spaces and Dimension
- 1.2 Normed Spaces and Convergence
- 1.3 Banach Spaces and the Cauchy Convergence Criterion
- 1.4 Open and Closed Sets
- 1.5 Operators
- 1.6 The Banach Fixed-Point Theorem and the Iteration Method
- 1.7 Applications to Integral Equations
- 1.8 Applications to Ordinary Differential Equations
- 1.9 Continuity
- 1.10 Convexity
- 1.11 Compactness
- 1.12 Finite-Dimensional Banach Spaces and Equivalent Norms
- 1.13 The Minkowski Functional and Homeomorphisms
- 1.14 The Brouwer Fixed-Point Theorem
- 1.15 The Schauder Fixed-Point Theorem
- 1.16 Applications to Integral Equations

- 1.17 Applications to Ordinary Differential Equations
- 1.18 The Leray–Schauder Principle and a priori Estimates
- 1.19 Sub- and Supersolutions, and the Iteration Method in Ordered Banach Spaces
- 1.20 Linear Operators
- 1.21 The Dual Space
- 1.22 Infinite Series in Normed Spaces
- 1.23 Banach Algebras and Operator Functions
- 1.24 Applications to Linear Differential Equations in Banach Spaces
- 1.25 Applications to the Spectrum
- 1.26 Density and Approximation
- 1.27 Summary of Important Notions

2 Hilbert Spaces, Orthogonality, and the Dirichlet Principle

- 2.1 Hilbert Spaces
- 2.2 Standard Examples
- 2.3 Bilinear Forms
- 2.4 The Main Theorem on Quadratic Variational Problems
- 2.5 The Functional Analytic Justification of the Dirichlet Principle
- 2.6 The Convergence of the Ritz Method for Quadratic Variational Problems
- 2.7 Applications to Boundary-Value Problems, the Method of Finite Elements, and Elasticity
- 2.8 Generalized Functions and Linear Functionals
- 2.9 Orthogonal Projection
- 2.10 Linear Functionals and the Riesz Theorem
- 2.11 The Duality Map
- 2.12 Duality for Quadratic Variational Problems
- 2.13 The Linear Orthogonality Principle
- 2.14 Nonlinear Monotone Operators
- 2.15 Applications to the Nonlinear Lax–Milgram Theorem and the Nonlinear Orthogonality Principle

3 Hilbert Spaces and Generalized Fourier Series

- 3.1 Orthonormal Series
- 3.2 Applications to Classical Fourier Series
- 3.3 The Schmidt Orthogonalization Method
- 3.4 Applications to Polynomials
- 3.5 Unitary Operators
- 3.6 The Extension Principle
- 3.7 Applications to the Fourier Transformation
- 3.8 The Fourier Transform of Tempered Generalized Functions

4 Eigenvalue Problems for Linear Compact Symmetric Operators

- 4.1 Symmetric Operators
- 4.2 The Hilbert–Schmidt Theory
- 4.3 The Fredholm Alternative
- 4.4 Applications to Integral Equations
- 4.5 Applications to Boundary-Eigenvalue Value Problems

5 Self-Adjoint Operators, the Friedrichs Extension and the Partial Differential Equations of Mathematical Physics

- 5.1 Extensions and Embeddings
- 5.2 Self-Adjoint Operators
- 5.3 The Energetic Space
- 5.4 The Energetic Extension
- 5.5 The Friedrichs Extension of Symmetric Operators
- 5.6 Applications to Boundary-Eigenvalue Problems for the Laplace Equation
- 5.7 The Poincaré Inequality and Rellich’s Compactness Theorem
- 5.8 Functions of Self-Adjoint Operators
- 5.9 Semigroups, One-Parameter Groups, and Their Physical Relevance
- 5.10 Applications to the Heat Equation
- 5.11 Applications to the Wave Equation
- 5.12 Applications to the Vibrating String and the Fourier Method
- 5.13 Applications to the Schrödinger Equation
- 5.14 Applications to Quantum Mechanics
- 5.15 Generalized Eigenfunctions
- 5.16 Trace Class Operators
- 5.17 Applications to Quantum Statistics
- 5.18 C^* -Algebras and the Algebraic Approach to Quantum Statistics
- 5.19 The Fock Space in Quantum Field Theory and the Pauli Principle
- 5.20 A Look at Scattering Theory
- 5.21 The Language of Physicists in Quantum Physics and the Justification of the Dirac Calculus
- 5.22 The Euclidean Strategy in Quantum Physics
- 5.23 Applications to Feynman’s Path Integral
- 5.24 The Importance of the Propagator in Quantum Physics
- 5.25 A Look at Solitons and Inverse Scattering Theory

Epilogue

Appendix

References

Hints for Further Reading

List of Symbols

List of Theorems

List of the Most Important Definitions

Subject Index

1

The Hahn–Banach Theorem and Optimization Problems

The most practical solution is a good theory.

Albert Einstein

True optimization is the revolutionary contribution of modern research to decision processes.

George Bernhard Dantzig
(born 1914)

The Hahn–Banach theorem is the most important theorem about the structure of linear continuous functionals on normed spaces. In terms of geometry, the Hahn–Banach theorem guarantees the separation of convex sets in normed spaces by hyperplanes. Figure 1.1 describes a number of important consequences of the Hahn–Banach theorem that will be studied in this chapter and the following one.

In this chapter we want to show that the Hahn–Banach theorem represents a fundamental *existence principle* in linear functional analysis that allows the solution of variational problems without using any compactness. In the next chapter, we will study variational problems by employing weak convergence, which is related to a generalized compactness concept.

The Hahn–Banach theorem was proved independently by Hahn in 1926 and by Banach in 1929. The discovery of this theorem was closely related to the famous classical moment problem.

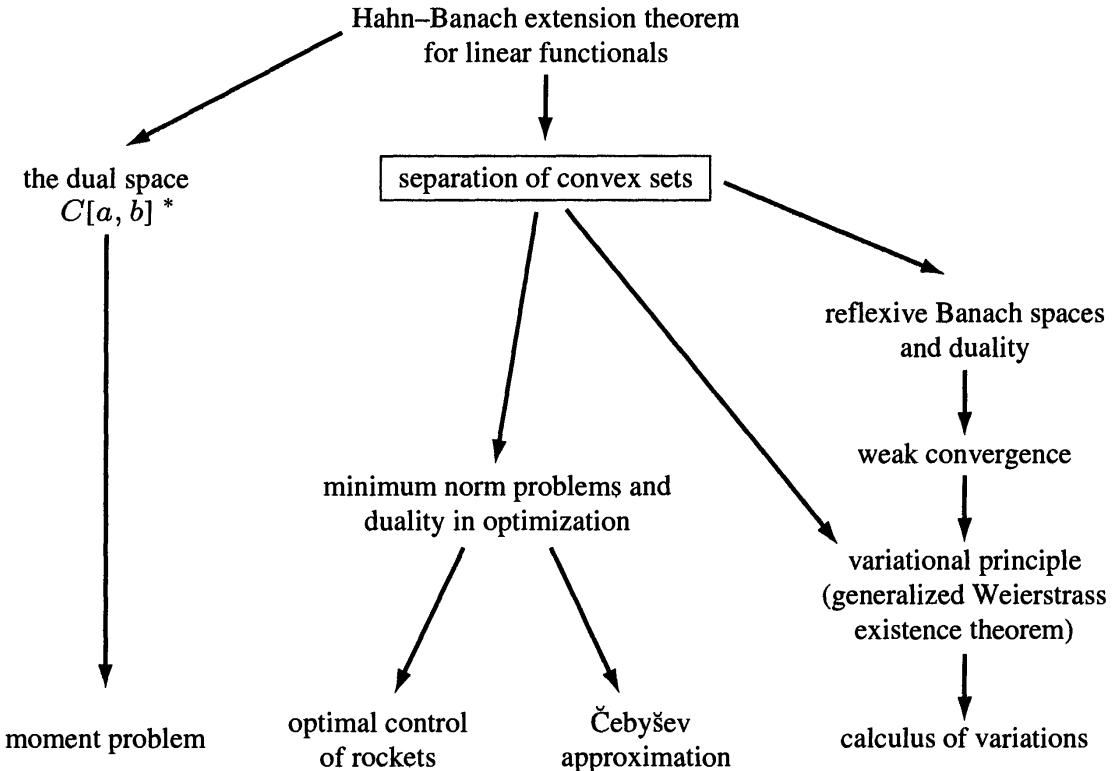


FIGURE 1.1.

1.1 The Hahn–Banach Theorem

Theorem 1.A (The Hahn–Banach theorem for linear spaces). *We assume that*

- (i) L is a linear subspace of the real linear space X .
- (ii) $p: X \rightarrow \mathbb{R}$ is a sublinear functional, that is, for all $u, v \in X$ and all $\alpha \geq 0$,

$$p(u + v) \leq p(u) + p(v) \quad \text{and} \quad p(\alpha u) = \alpha p(u).$$

- (iii) $F: L \rightarrow \mathbb{R}$ is a linear functional such that

$$F(u) \leq p(u) \quad \text{for all } u \in L. \tag{1}$$

Then, F can be extended to a linear functional $f: X \rightarrow \mathbb{R}$ such that

$$f(u) \leq p(u) \quad \text{for all } u \in X. \tag{1^*}$$

Proof. *Step 1:* We first prove the statement in the special case where

$$X = L + \text{span}\{v\} \quad \text{with fixed } v \notin L.$$

To this end, we set

$$f(u + \lambda v) := F(u) + c\lambda \quad \text{for all } u \in L, \lambda \in \mathbb{R},$$

where c is a fixed real number that satisfies the following condition:

$$\sup_{u \in L} (F(u) - p(u - v)) \leq c \leq \inf_{w \in L} (p(w + v) - F(w)). \quad (2)$$

We have to show that such a number c exists. In fact, for all $u, w \in L$, we get

$$\begin{aligned} F(u) + F(w) &= F(u + w) \leq p(u + w) \\ &= p(u - v + w + v) \leq p(u - v) + p(w + v), \end{aligned}$$

and hence

$$F(u) - p(u - v) \leq p(w + v) - F(w) \quad \text{for all } u, w \in L.$$

This proves (2).

Obviously, the functional $f: X \rightarrow \mathbb{R}$ is linear. Thus, it remains to show that

$$F(u) + c\lambda \leq p(u + \lambda v) \quad \text{for all } u \in L, \lambda \in \mathbb{R}. \quad (3)$$

In fact, this is true for $\lambda = 0$. Let $\lambda > 0$. By (2),

$$c \leq p(\lambda^{-1}u + v) - F(\lambda^{-1}u) = \lambda^{-1}(p(u + \lambda v) - F(u)).$$

This is (3). In the case where $\lambda < 0$, it follows from (2) that

$$c \geq F(-\lambda^{-1}u) - p(-\lambda^{-1}u - v) = -\lambda^{-1}(F(u) - p(u + \lambda v)),$$

and again we get (3).

Step 2: Induction. Suppose that there exists a sequence (L_n) of linear subspaces of X such that $L = L_1 \subseteq L_2 \subseteq \dots$ along with

$$X = \bigcup_n L_n,$$

where

$$L_{n+1} = L_n + \text{span}\{v_n\} \quad \text{for some fixed } v_n \in X \text{ and } v_n \notin L_n,$$

and for all n . Using Step 1, a simple induction argument shows that F can be extended to L_n for all n . This yields the desired extension f of F .

Step 3: If the situation from Step 2 is not at hand, then we can use the *Zorn lemma* from the appendix of AMS Vol. 108. To this end, let \mathcal{C} denote the set of all the linear functionals

$$g: D(g) \subseteq X \rightarrow \mathbb{K}$$

that are an extension of F such that

$$g(u) \leq p(u) \quad \text{for all } u \in D(g).$$

We write

$$g \leq h \quad \text{iff } h: D(h) \rightarrow \mathbb{K} \text{ is an extension of } g: D(g) \rightarrow \mathbb{K}.$$

This way \mathcal{C} becomes an *ordered set*. Let \mathcal{T} be a totally ordered subset of \mathcal{C} , that is, $g, h \in \mathcal{T}$ implies

$$g \leq h \quad \text{or} \quad h \leq g.$$

Then, there exists an upper bound $b \in \mathcal{C}$ for \mathcal{T} , that is,

$$g \leq b \quad \text{for all } g \in \mathcal{T}.$$

To show this, let $D(b)$ be the union of all the sets $D(g)$ with $g \in \mathcal{T}$ and define

$$b(u) := g(u) \quad \text{on } D(g).$$

Since \mathcal{T} is totally ordered, the linear functional $b: D(b) \rightarrow \mathbb{K}$ is well defined, and $b(u) \leq p(u)$ for all $u \in D(b)$.

By the Zorn lemma, there is a maximal element f in \mathcal{C} . That is, the linear functional $f: D(f) \subseteq X \rightarrow \mathbb{K}$ has no proper extension in the sense of \mathcal{C} . This implies $D(f) = X$ and $f(u) \leq p(u)$ for all $u \in X$. In fact, suppose that $D(f) \neq X$. Then, there exists an extension of f in the sense of \mathcal{C} , by Step 1. This contradicts the maximality of f . Thus, $f: X \rightarrow \mathbb{R}$ is the desired extension of F . \square

Theorem 1.B (The Hahn–Banach theorem for normed spaces¹). *We assume that*

- (i) *L is a linear subspace of the normed space X over \mathbb{K} .*
- (ii) *$F: L \rightarrow \mathbb{K}$ is a linear functional such that*

$$|F(u)| \leq \alpha \|u\| \quad \text{for all } u \in L \text{ and fixed } \alpha \geq 0. \quad (4)$$

Then, F can be extended to a linear continuous functional $f: X \rightarrow \mathbb{K}$ such that

$$|f(u)| \leq \alpha \|u\| \quad \text{for all } u \in X. \quad (4^*)$$

Proof. *Step 1:* Let $\mathbb{K} = \mathbb{R}$. Set

$$p(u) := \alpha \|u\| \quad \text{for all } u \in X.$$

¹Basic notions on normed spaces are summarized in Section 1.27 of AMS Vol. 108.

By Theorem 1.A, the functional F can be extended to a linear functional $f: X \rightarrow \mathbb{R}$ such that

$$f(u) \leq \alpha \|u\| \quad \text{for all } u \in X.$$

Since $f(\pm u) = \pm f(u)$, we get (4*). Thus, f is continuous.

Step 2: Let $\mathbb{K} = \mathbb{C}$. Define

$$H(u) := \operatorname{Re} F(u) \quad \text{for all } u \in L.$$

Then,

$$\begin{aligned} F(u) &= \operatorname{Re} F(u) + i \operatorname{Im} F(u) = \operatorname{Re} F(u) - i \operatorname{Re} F(iu), \\ &= H(u) - iH(iu) \quad \text{for all } u \in L, \end{aligned}$$

and

$$|H(u)| \leq \alpha \|u\| \quad \text{for all } u \in L.$$

If we regard X as a real normed space, then it follows from Step 1 that there exists a linear continuous functional $h: X \rightarrow \mathbb{R}$ such that $h(u) = H(u)$ for all $u \in L$ and

$$|h(u)| \leq \alpha \|u\| \quad \text{for all } u \in X.$$

Define

$$f(u) := h(u) - ih(iu) \quad \text{for all } u \in X.$$

Hence $h(u) = \operatorname{Re} f(u)$. We want to show that f is the desired functional.

Obviously, $f: X \rightarrow \mathbb{C}$ is an extension of F . Moreover, f is linear. This follows from

$$f(iu) = h(iu) - ih(-u) = if(u) \quad \text{for all } u \in X,$$

and from the linearity of h with respect to \mathbb{R} . Finally, we have to show that

$$|f(u)| \leq \alpha \|u\| \quad \text{for all } u \in X.$$

In fact, for each $u \in X$, we get $f(u) = re^{i\beta}$ with $r \geq 0$. Hence

$$\begin{aligned} |f(u)| &= r = \operatorname{Re}(e^{-i\beta} f(u)) = \operatorname{Re} f(e^{-i\beta} u) \\ &= h(e^{-i\beta} u) \leq \alpha \|e^{-i\beta} u\| = \alpha \|u\|. \end{aligned}$$
□

Standard Example 1. Let X be a normed space over \mathbb{K} . Then, for each given $u_0 \in X$ with $u_0 \neq 0$, there exists a functional $f \in X^*$ such that²

$$f(u_0) = \|u_0\| \quad \text{and} \quad \|f\| = 1.$$

²Recall from Section 1.21 of AMS Vol. 108 that the dual space X^* to X consists of all linear continuous functionals $f: X \rightarrow \mathbb{K}$.

Proof. Set $L := \text{span}\{u_0\}$ and

$$F(u) := \lambda \|u_0\| \quad \text{for all } u \in L, \text{ where } u = \lambda u_0.$$

Obviously, $|F(u)| = \|u\|$ for all $u \in L$. By Theorem 1.B, there exists a functional $f \in X^*$ such that $f(u) = F(u)$ on L and

$$|f(u)| \leq \|u\| \quad \text{for all } u \in X.$$

Hence $\|f\| = 1$. □

Corollary 2. Let X be a normed space over \mathbb{K} . Then, for all $u_0 \in X$,

$$\|u_0\| = \max_{f \in X^*, \|f\| \leq 1} |f(u_0)|.$$

Proof. Since $|f(u_0)| \leq \|f\| \|u_0\|$ for all $f \in X^*$, the assertion follows from Standard Example 1. □

Corollary 3. Let X be a normed space over \mathbb{K} . Then, it follows from $u \in X$ and

$$f(u) = 0 \quad \text{for all } f \in X^*$$

that $u = 0$.

This is an immediate consequence of Standard Example 1.

1.2 Applications to the Separation of Convex Sets

Definition 1. By a *closed hyperplane* H in the real normed space X , we understand a set

$$H := \{u \in X : f(u) = \alpha\},$$

where $f: X \rightarrow \mathbb{R}$ is a linear continuous functional and α is a fixed real number. We also define the *half-spaces* H_{\leq} and $H_{>}$ of H through

$$H_{\leq} := \{u \in X : f(u) \leq \alpha\} \quad \text{and} \quad H_{>} := \{u \in X : f(u) > \alpha\}.$$

Let A and B be two subsets of X . Then, we say that the closed hyperplane H strictly *separates* the sets A and B iff

$$A \subseteq H_{\leq} \quad \text{and} \quad B \subseteq H_{>}.$$

Furthermore, we say that the closed hyperplane H separates the sets A and B iff $A \subseteq H_{\leq}$ and $B \subseteq H_{\geq}$.

Example 2. Let $X := \mathbb{R}^2$. Then, every closed hyperplane H in X is given through

$$H := \{(\xi, \eta) \in \mathbb{R}^2 : a\xi + b\eta = \alpha\},$$

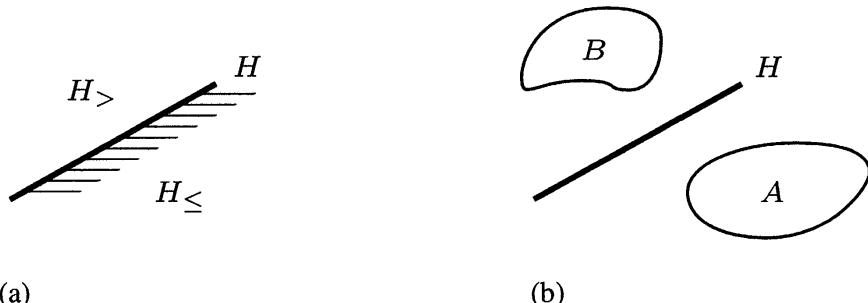


FIGURE 1.2.

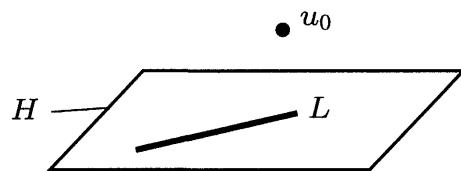


FIGURE 1.3.

where a , b , and α are fixed real numbers. In Figure 1.2, the sets A and B are strictly separated by H .

Proposition 3. Let L be a linear subspace of the normed space X over \mathbb{K} . Then, for each point $u_0 \in X$ with

$$\text{dist}(u_0, L) > 0,$$

there exists a linear continuous functional $f: X \rightarrow \mathbb{K}$ such that

$$f(u) = 0 \quad \text{for all } u \in L,$$

along with $\|f\| = 1$ and $f(u_0) = \text{dist}(u_0, L)$.

Recall that

$$\text{dist}(u_0, L) := \inf_{v \in L} \|u_0 - v\|. \quad (5)$$

If X is a real normed space, then this means that the closed hyperplane

$$H := \{u \in X : f(u) = 0\}$$

separates strictly the linear subspace L and the point u_0 , where $L \subseteq H$ (see Figure 1.3).

Proof. Set $L_0 := \text{span}\{u_0\} + L$. Then, $u \in L_0$ iff

$u = \lambda u_0 + v$, where $\lambda \in \mathbb{K}$ and $v \in L$.

This representation of u is unique. In fact, it follows from $u = \mu u_0 + w$ with $\mu \in \mathbb{K}$ and $w \in L$ that $(\mu - \lambda)u_0 = w - v$. Hence $\mu - \lambda = 0$ and $v - w = 0$ because $u_0 \notin L$ and $w - v \in L$. Define

$$F(u) := \lambda \operatorname{dist}(u_0, L) \quad \text{for all } u \in L_0.$$

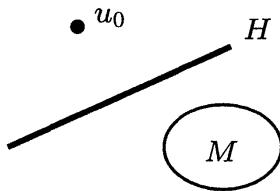


FIGURE 1.4.

Obviously, $F: L \rightarrow \mathbb{K}$ is linear. Furthermore,

$$|F(u)| \leq \|u\| \quad \text{for all } u \in L_0.$$

This follows from

$$|F(\lambda u_0 + v)| = |\lambda| \operatorname{dist}(u_0, L) \leq |\lambda| \|u_0 - (-\lambda^{-1}v)\| = \|\lambda u_0 + v\|,$$

for all $v \in L$ and $\lambda \in \mathbb{K}$ with $\lambda \neq 0$. According to Theorem 1.B, F can be extended to a linear continuous functional $f: X \rightarrow \mathbb{K}$ such that

$$|f(u)| \leq \|u\| \quad \text{for all } u \in X.$$

By (5), for each $\varepsilon > 0$, there is a $v \in L$ such that $\|u_0 - v\| < \operatorname{dist}(u_0, L) + \varepsilon$. Since $f = F$ on L , we get $f(u_0 - v) = \operatorname{dist}(u_0, L)$, and hence

$$\frac{f(u_0 - v)}{\|u_0 - v\|} > \frac{\operatorname{dist}(u_0, L)}{\operatorname{dist}(u_0, L) + \varepsilon}.$$

Letting $\varepsilon \rightarrow 0$, this implies that $\|f\| = 1$. □

Theorem 1.C. *Let M be a nonempty closed convex subset of a normed space X over \mathbb{K} , and let u_0 be a point of X with $u_0 \notin M$.*

Then, there exists a linear continuous functional $f: X \rightarrow \mathbb{K}$ such that

$$\operatorname{Re} f(u) \leq 1 \quad \text{for all } u \in M \quad \text{and} \quad \operatorname{Re} f(u_0) > 1.$$

In terms of geometry, this theorem tells us the following. Let X be a real normed space. Set

$$H := \{u \in X : f(u) = 1\}.$$

Then, the closed hyperplane H separates the set M and the point u_0 (see Figure 1.4).

Proof. Since $u_0 \notin M$ and M is closed, we get

$$d := \operatorname{dist}(u_0, M) > 0.$$

Otherwise, there exists a sequence (v_n) in M such that $\|u_0 - v_n\| \rightarrow 0$ as $n \rightarrow \infty$, and hence $u_0 \in M$. This is a contradiction. Define

$$M_d := \left\{ u \in X : \operatorname{dist}(u, M) < \frac{d}{2} \right\}.$$

Since M is convex, so is M_d . In fact, if $u_j \in M_d$, $j = 1, 2$, then there exist points $v_j \in M$ such that $\|u_j - v_j\| < \frac{d}{2}$ for $j = 1, 2$. For all $t \in [0, 1]$,

$$\|tu_1 + (1-t)u_2 - (tv_1 + (1-t)v_2)\| \leq t\|u_1 - v_1\| + (1-t)\|u_2 - v_2\| < \frac{d}{2},$$

and hence M_d is convex. Furthermore, let

$$\mathcal{M} := \text{closure of } M_d.$$

Using sequences, it follows easily that the closure of each convex set is again convex.

Summarizing, \mathcal{M} is a closed convex set such that $u_0 \notin \mathcal{M}$, and \mathcal{M} also has an interior point. In addition, $M \subseteq \mathcal{M}$. By Section 1.13 of AMS Vol. 108, the *Minkowski functional* p of \mathcal{M} is sublinear and we get

$$\mathcal{M} = \{u \in X : p(u) \leq 1\}, \quad (6)$$

along with

$$0 \leq p(u) \leq c\|u\| \quad \text{for all } u \in X \text{ and fixed } c > 0.$$

Step 1: Let $\mathbb{K} = \mathbb{R}$. Define $L := \text{span}\{u_0\}$ and

$$F(\lambda u_0) := \lambda p(u_0) \quad \text{for all } \lambda \in \mathbb{R}.$$

Then,

$$F(u) \leq p(u) \quad \text{for all } u \in L.$$

In fact, if $\lambda \geq 0$, then $F(\lambda u_0) = p(\lambda u_0)$, and if $\lambda < 0$, then $F(\lambda u_0) \leq 0$.

According to the Hahn–Banach theorem (Theorem 1.A), F can be extended to a linear functional $f: X \rightarrow \mathbb{R}$ such that

$$f(u) \leq p(u) \leq c\|u\| \quad \text{for all } u \in X.$$

Since $f(\pm u) = \pm f(u)$, we get $|f(u)| \leq c\|u\|$ for all $u \in X$, and hence f is continuous on X .

Finally, we obtain that

$$f(u) \leq p(u) \leq 1 \quad \text{for all } u \in \mathcal{M},$$

and $f(u_0) = F(u_0) = p(u_0) > 1$, by $u_0 \notin \mathcal{M}$ and (6). This proves the assertion.

Step 2: Let $\mathbb{K} = \mathbb{C}$. If we regard X as a real normed space, then we may construct the linear continuous functional $f: X \rightarrow \mathbb{R}$ as in Step 1. Then, the functional $h: X \rightarrow \mathbb{C}$ defined by

$$h(u) := f(u) - if(iu) \quad \text{for all } u \in X$$

has the desired properties. This follows by a similar argument as in the proof of Theorem 1.B. \square

1.3 The Dual Space $C[a, b]^*$

Proposition 1. Let $-\infty < a < b < \infty$. Then, $F \in C[a, b]^*$ iff there exists a function $\rho: [a, b] \rightarrow \mathbb{R}$ of bounded variation such that³

$$F(u) = \int_a^b u(x)d\rho(x) \quad \text{for all } u \in C[a, b]. \quad (7)$$

In addition, $\|F\| = V(\rho)$, where $V(\rho)$ denotes the total variation of ρ .

The integral (7) represents a *Stieltjes integral*. Such integrals along with functions of bounded variation are discussed in the appendix of AMS Vol. 108. The proof will be based on the Hahn–Banach theorem.

Proof. We set $X := C[a, b]$ and $\|u\| := \sup_{a \leq x \leq b} |u(x)|$.

Step 1: Let F be as given in (7). By the appendix of AMS Vol. 108,

$$|F(u)| \leq V(\rho)\|u\| \quad \text{for all } u \in X.$$

Hence $F \in C[a, b]^*$.

Step 2: Let $F \in C[a, b]^*$. We want to prove that F allows a representation of the form (7).

To this end, let Y denote the space of all *bounded functions* $u: [a, b] \rightarrow \mathbb{R}$. Then, Y becomes a normed space with respect to $\|u\|$.

Since X is a linear subspace of Y , it follows from the *Hahn–Banach theorem* (Theorem 1.B) that F can be extended to a linear continuous functional

$$f: Y \rightarrow \mathbb{R} \quad \text{with } \|F\| = \|f\|.$$

Set $\rho(t) := f(v_t)$ for all $t \in [a, b]$, where

$$v_t(x) := \begin{cases} 1 & \text{if } a \leq x \leq t \\ 0 & \text{if } t < x \leq b. \end{cases}$$

We will prove in Step 3 ahead that $\rho: [a, b] \rightarrow \mathbb{R}$ is of *bounded variation* and

$$V(\rho) \leq \|F\|. \quad (8)$$

³Recall from Chapter 1 of AMS Vol. 108 that the space $C[a, b]$ consists of all continuous functions $u: [a, b] \rightarrow \mathbb{R}$. The norm on $C[a, b]$ is given by

$$\|u\| := \max_{a \leq x \leq b} |u(x)|.$$

The dual space $C[a, b]^*$ consists of all linear continuous functionals on $C[a, b]$.

Now let $u \in X$ be given. Consider the partition $a = x_0 < x_1 < \dots < x_n = b$ of the interval $[a, b]$. Then, the continuous function $u: [a, b] \rightarrow \mathbb{R}$ can be approximated by the step function

$$u_n(x) := \sum_{j=1}^n u(x_j)(v_{x_j}(x) - v_{x_{j-1}}(x)).$$

Hence

$$f(u_n) = \sum_{j=1}^n u(x_j)(\rho(x_j) - \rho(x_{j-1})).$$

If the partition is made arbitrarily fine as $n \rightarrow \infty$, we have

$$\lim_{n \rightarrow \infty} f(u_n) = \int_a^b u(x)d\rho(x),$$

by the definition of the Stieltjes integral in the appendix of AMS Vol. 108. On the other hand,

$$u_n \rightarrow u \text{ in } Y \quad \text{as } n \rightarrow \infty.$$

Since f is continuous on Y , this implies $f(u_n) \rightarrow f(u)$ as $n \rightarrow \infty$. Hence

$$F(u) = f(u) = \int_a^b u(x)d\rho(x) \quad \text{for all } u \in X.$$

This is (7).

By Step 1, we get $\|F\| \leq V(\rho)$, and (8) yields $V(\rho) \leq \|F\|$. Thus, $\|F\| = V(\rho)$.

Step 3: Proof of (8). Using the partition $\{x_j\}$ of $[a, b]$ from Step 2 and letting $s_j := \operatorname{sgn}(\rho(x_j) - \rho(x_{j-1}))$, we have

$$\begin{aligned} \Delta := \sum_{j=1}^n |\rho(x_j) - \rho(x_{j-1})| &= \sum_{j=1}^n s_j (\rho(x_j) - \rho(x_{j-1})) \\ &= \sum_{j=1}^n s_j (f(v_{x_j}) - f(v_{x_{j-1}})) \\ &= f \left(\sum_{j=1}^n s_j (v_{x_j} - v_{x_{j-1}}) \right). \end{aligned}$$

Thus,

$$\Delta \leq \|f\| \left\| \sum_{j=1}^n s_j (v_{x_j} - v_{x_{j-1}}) \right\| = \|f\| = \|F\|,$$

and hence $V(\rho) \leq \|F\|$, by the definition of the total variation $V(\rho)$ of ρ in the appendix of AMS Vol. 108. \square

Example 2. Let $w: [a, b] \rightarrow \infty$ be a continuous function, where $-\infty < a < b < \infty$. Set

$$F(u) := \int_a^b u(x)w(x)dx \quad \text{for all } u \in C[a, b].$$

Then, $F \in C[a, b]^*$ and

$$\|F\| = \int_a^b |w(x)|dx.$$

Proof. Define

$$\rho(x) := \int_a^x w(t)dt \quad \text{for all } x \in [a, b].$$

Then

$$F(u) = \int_a^b u(x)d\rho(x) \quad \text{for all } u \in C[a, b],$$

and $\|F\| = V(\rho)$, by (5) in the appendix of AMS Vol. 108.

Let $a = x_0 < x_1 < \dots < x_n = b$ be a partition of the interval $[a, b]$. Then

$$\Delta := \sum_{j=1}^n |\rho(x_j) - \rho(x_{j-1})| \leq \sum_{j=1}^n \int_{x_{j-1}}^{x_j} |w(t)|dt = \int_a^b |w(t)|dt.$$

Hence

$$V(\rho) \leq \int_a^b |w(t)|dt.$$

By the mean value theorem,

$$\Delta = \sum_{j=1}^n |w(t_j)|(x_j - x_{j-1}), \quad \text{where } x_{j-1} \leq t_j \leq x_j.$$

Making the partition arbitrarily fine as $n \rightarrow \infty$, we get

$$\Delta \rightarrow \int_a^b |w(t)|dt \quad \text{as } n \rightarrow \infty,$$

and hence

$$V(\rho) = \int_a^b |w(t)|dt. \quad \square$$

1.4 Applications to the Moment Problem

The Finite Moment Problem. Let $-\infty < a < b < \infty$. We are given the real numbers $\mu_0, \mu_1, \dots, \mu_N$ for fixed $N \geq 0$. We are looking for a function $\rho: [a, b] \rightarrow \mathbb{R}$ of bounded variation such that

$$\int_a^b x^k d\rho(x) = \mu_k \quad \text{for all } k = 0, \dots, N. \quad (9)$$

In terms of physics, we are looking for a charge density ρ that has the prescribed moments μ_k , $k = 0, \dots, N$. In particular, μ_0 is equal to the total charge on the interval $[a, b]$.

Proposition 1. *The finite moment problem has always a solution.*

Proof. Let $X := C[a, b]$, and let $p_k(x) := x^k$, $k = 0, 1, \dots, N$. Set

$$L := \text{span}\{p_0, p_1, \dots, p_N\}.$$

Then, the $(N + 1)$ -dimensional linear subspace L of X consists of all the real polynomials of order $\leq N$. Let $u \in X$. Then

$$u(x) = a_0 + a_1x + \cdots + a_Nx^N \quad \text{for all } x \in [a, b].$$

Define

$$F(u) := a_0 + a_1\mu_1 + \cdots + a_N\mu_N \quad \text{for all } u \in L.$$

Obviously, the functional $F: L \rightarrow \mathbb{R}$ is linear. This functional is also *continuous*. To prove this, let

$$u_n \rightarrow u \quad \text{in } L \quad \text{as } n \rightarrow \infty.$$

This implies $u_n(x) \rightarrow u(x)$ as $n \rightarrow \infty$ uniformly on $[a, b]$. By the well-known Lagrangian interpolation formula, we get

$$u_n(x) = \sum_{j=0}^N u_n(x_j)\phi_j(x) \quad \text{on } [a, b] \text{ for all } n, \quad (10)$$

where $a = x_0 < x_1 < \cdots < x_N = b$ is a fixed given partition of $[a, b]$, and ϕ_0, \dots, ϕ_N are fixed N -th order polynomials with $\phi_j(x_i) = \delta_{ij}$ for all i, j . It follows from (10) that the coefficients of u_n converge to the corresponding coefficients of u as $n \rightarrow \infty$. Hence

$$F(u_n) \rightarrow F(u) \quad \text{as } n \rightarrow \infty.$$

Since $F: L \rightarrow \mathbb{R}$ is linear and continuous, we get

$$|F(u)| \leq \text{const}\|u\| \quad \text{for all } u \in L,$$

where $\|u\|$ denotes the norm on X , i.e., $\|u\| := \max_{a \leq x \leq b} |u(x)|$.

According to the *Hahn–Banach theorem*, F can be extended to a linear continuous functional $f: X \rightarrow \mathbb{R}$. By Section 1.3, there exists a function $\rho: [a, b] \rightarrow \mathbb{R}$ of bounded variation such that

$$f(u) = \int_a^b u(x)d\rho(x) \quad \text{for all } u \in C[a, b].$$

This implies (9). In fact, $f(p_j) = F(p_j) = \mu_j$. \square

The Moment Problem. Let $-\infty < a < b < \infty$. We are given the real numbers μ_0, μ_1, \dots . We are looking for a function $\rho: [a, b] \rightarrow \mathbb{R}$ of bounded variation such that

$$\int_a^b x^k d\rho(x) = \mu_k \quad \text{for all } k = 0, 1, \dots. \quad (11)$$

Proposition 2. The moment problem has a solution iff there is a constant $c > 0$ such that

$$\left| \sum_{k=0}^N a_k \mu_k \right| \leq c \max_{a \leq x \leq b} \left| \sum_{k=0}^N a_k x^k \right| \quad \text{for all real } a_k \text{ and all } N = 0, 1, 2, \dots \quad (12)$$

Proof. We use the same notation as in the proof of Proposition 1.

If the moment problem has a solution ρ , then the functional

$$F(u) := \int_a^b u(x)d\rho(x) \quad \text{for all } u \in X$$

is linear and continuous on X . Hence

$$|F(u)| \leq c\|u\| \quad \text{for all } u \in X.$$

Using $F(p_k) = \mu_k$ and $u = a_0 p_0 + \dots + a_N p_N$, we get (12).

Conversely, suppose that (12) holds true. Let $L := \text{span}\{p_0, p_1, \dots\}$. Define

$$F(p_k) := \mu_k \quad \text{for all } k = 0, 1, \dots.$$

This way, we obtain a linear functional $F: L \rightarrow \mathbb{R}$. By (12),

$$|F(u)| \leq c\|u\| \quad \text{for all } u \in L.$$

Since L is dense in X , the functional F can be extended to a linear continuous functional $F: X \rightarrow \mathbb{R}$, by the extension principle from Section 3.6 of AMS Vol. 108. According to Section 1.3, there exists a function $\rho: [a, b] \rightarrow \mathbb{R}$ of bounded variation such that

$$F(u) = \int_a^b u(x)d\rho(x) \quad \text{for all } u \in C[a, b].$$

This implies (11). \square

1.5 Minimum Norm Problems and Duality Theory

Along with the *primal problem*

$$\inf \|u - u_0\| = \alpha, \quad u \in L, \quad (13)$$

let us consider the *dual problem*

$$\sup \langle u^*, u_0 \rangle = \beta, \quad u^* \in L^\perp, \quad \|u^*\| \leq 1, \quad (13^*)$$

where $L^\perp := \{u \in X^* : \langle u^*, u \rangle = 0 \text{ for all } u \in L\}$.

Theorem 1.D (Minimum norm problem on the normed space X). *Let L be a linear subspace of the real normed space X . We are given $u_0 \in X$. Then the following conditions hold:*

- (i) Extremal values: $\alpha = \beta$.
- (ii) Dual problem: *The dual problem (13*) has a solution u^* .*
- (iii) Primal problem: *Let u^* be a fixed solution of the dual problem (13*). Then, the point $u \in L$ is a solution of the primal problem (13) iff*

$$\langle u^*, u_0 - u \rangle = \|u - u_0\|. \quad (14)$$

Corollary 1. *If $\dim L < \infty$, then the primal problem (13) always has a solution.*

Let $v \in L$ and $v^* \in L^\perp$ with $\|v^*\| \leq 1$. Then, from (i) we obtain the two-sided *error estimate* for the minimal value α :

$$\|v - u_0\| \geq \alpha \geq \langle v^*, u_0 \rangle.$$

Proof.⁴ Ad (i), (ii). For each $\varepsilon > 0$, there is a point $u \in L$ such that

$$\|u - u_0\| \leq \alpha + \varepsilon.$$

Thus, for all $u^* \in L^\perp$ with $\|u^*\| \leq 1$,

$$\langle u^*, u_0 \rangle = \langle u^*, u_0 - u \rangle \leq \|u^*\| \|u - u_0\| \leq \alpha + \varepsilon.$$

Hence $\beta \leq \alpha + \varepsilon$ for all $\varepsilon > 0$, that is, $\beta \leq \alpha$.

⁴The Latin notion “Ad (i)” stands for “proof of (i)”.

Let $\alpha > 0$. It follows from Proposition 3 in Section 1.2 that there is a functional $u^* \in L^\perp$ with $\|u^*\| = 1$ such that

$$\langle u^*, u_0 \rangle = \alpha. \quad (15)$$

Along with $\beta \leq \alpha$, this implies $\beta = \alpha$.

If $\alpha = 0$, then (15) holds with $u^* = 0$, and hence we again have $\alpha = \beta$.

Ad (iii). This follows from $\alpha = \beta$ and $\langle u^*, u \rangle = 0$. \square

Proof of Corollary 1. Since $0 \in L$, $\alpha \leq \|u_0\|$. Thus, problem (13) is equivalent to the finite-dimensional minimum problem

$$\|u - u_0\| = \min !, \quad u \in L_0,$$

where the set $L_0 := \{u \in L : \|u\| \leq \|u_0\|\}$ is compact. By the Weierstrass theorem (Proposition 8 in Section 1.11 of AMS Vol. 108), this problem has a solution. \square

Remark 2. Let $\dim L = \infty$, where L is a closed linear subspace of the real reflexive Banach space X (e.g., X is a real Hilbert space⁵), and let $u_0 \in X$ be given. Then, the primal problem (13) has a solution.

This will be proved in Section 2.9 (cf. Theorem 2.E). Note that this result will not be used in the present chapter.

In contrast to (13), we now consider the *modified* primal problem

$$\inf \|u^* - u_0^*\| = \alpha, \quad u^* \in L^\perp, \quad (16)$$

along with the dual problem

$$\sup \langle u_0^*, u \rangle = \beta, \quad u \in L, \quad \|u\| \leq 1. \quad (16^*)$$

Recall that $L^\perp := \{u^* \in X^* : \langle u^*, u \rangle = 0 \text{ for all } u \in L\}$. Thus, the primal problem (16) refers to the dual space X^* , whereas the dual problem (16^{*}) refers to the original space X .

Theorem 1.E (Minimum norm problem on the dual space X^*). *Let L be a linear subspace of the real normed space X . We are given $u_0^* \in X^*$. Then the following conditions hold:*

- (i) Extremal values: $\alpha = \beta$.
- (ii) Primal problem: *The primal problem (16) has a solution u^* .*

⁵The basic properties of Hilbert spaces can be found in Chapter 2 of AMS Vol. 108.

(iii) Dual problem: Let u^* be a fixed solution of the primal problem (16). Then, the point $u \in L$ with $\|u\| \leq 1$ is a solution of the dual problem (16*) iff

$$\langle u_0^* - u^*, u \rangle = \|u_0^* - u^*\|. \quad (17)$$

Proof. Ad (i), (ii). For all $u^* \in L^\perp$,

$$\begin{aligned} \|u^* - u_0^*\| &= \sup_{\|u\| \leq 1} (\langle u_0^*, u \rangle - \langle u^*, u \rangle) \\ &\geq \sup_{\|u\| \leq 1, u \in L} \langle u_0^*, u \rangle = \beta, \end{aligned}$$

since $\langle u^*, u \rangle = 0$ for all $u \in L$. Hence $\alpha \geq \beta$.

Let $u_r^*: L \rightarrow \mathbb{R}$ be the restriction of $u_0^*: X \rightarrow \mathbb{R}$ to L . Then

$$\|u_r^*\| = \sup_{\|u\| \leq 1, u \in L} \langle u_0^*, u \rangle = \beta.$$

By the *Hahn–Banach theorem* (Theorem 1.B), there exists an extension $U^*: X \rightarrow \mathbb{R}$ of u_r^* with $\|U^*\| = \|u_r^*\|$. This implies

$$v^* := u_0^* - U^* = 0 \quad \text{on } L,$$

that is, $v^* \in L^\perp$. Since $\alpha \geq \beta$ and

$$\|v^* - u_0^*\| = \|U^*\| = \|u_r^*\| = \beta, \quad v^* \in L^\perp,$$

we get $\alpha = \beta$.

Ad (iii). This follows from $\alpha = \beta$ with $\langle u^*, u \rangle = 0$. □

In Sections 1.6 and 1.7, Theorems 1.D and 1.E will be applied to Čebyšev approximation and the optimal control of rockets, respectively. In this connection, the following lemma will be used critically.

Let $-\infty < a \leq c \leq b < \infty$. Set

$$\delta_c(u) := u(c) \quad \text{for all } u \in C[a, b].$$

Obviously, $\delta_c \in C[a, b]^*$ and $\|\delta_c\| = 1$.

Lemma 3. Let $u^* \in C[a, b]^*$ be such that $\|u^*\| \neq 0$. Suppose that

$$\langle u^*, u \rangle = \|u^*\| \|u\| \quad \text{where } \|u\| := \max_{a \leq x \leq b} |u(x)|,$$

and $u: [a, b] \rightarrow \mathbb{R}$ is a continuous function such that $|u(x)|$ achieves its maximum at precisely N points of $[a, b]$ denoted by x_1, \dots, x_N .

Then, there exist real numbers $\alpha_1, \dots, \alpha_N$ such that

$$u^* = \alpha_1 \delta_{x_1} + \dots + \alpha_N \delta_{x_N},$$

and $|\alpha_1| + \cdots + |\alpha_N| = \|u^*\|$.

Proof. By Section 1.3, there exists a function $\rho: [a, b] \rightarrow \mathbb{R}$ of bounded variation such that

$$\langle u^*, u \rangle = \int_a^b u(x) d\rho(x) \quad \text{for all } u \in C[a, b],$$

and $V(\rho) = \|u^*\|$, where $V(\rho)$ denotes the total variation of ρ on the interval $[a, b]$ (cf. the appendix of AMS Vol. 108). We may assume that $\rho(a) = 0$. To explain the simple idea of the proof, assume that $N = 1$, $\pm u(x_1) = \|u\|$, and $a < x_1 < b$.

Let $J := [a, b] -]x_1 - \varepsilon, x_1 + \varepsilon[$ for fixed $\varepsilon > 0$, and let $V_J(\rho)$ denote the total variation of ρ on J . Then,

$$V_J(\rho) + |\rho(x_1 + \varepsilon) - \rho(x_1 - \varepsilon)| \leq V(\rho). \quad (18)$$

Case 1: Let $V_J(\rho) = 0$ for all $\varepsilon > 0$. Then, by (18), ρ is a step function of the following form:

$$\rho(x) := \begin{cases} 0 & \text{if } a \leq x < x_1 \\ \pm V(\rho) & \text{if } x_1 < x \leq b, \end{cases}$$

and, by the definition of the Stieltjes integral (cf. the appendix of AMS Vol. 108),

$$\langle u^*, u \rangle = \int_a^b u(x) d\rho(x) = \pm u(x_1)V(\rho) \quad \text{for all } u \in C[a, b].$$

Hence $u^* = \pm V(\rho)\delta_{x_1}$.

Case 2: Let $V_J(\rho) > 0$ for some $\varepsilon > 0$. We want to show that this is impossible. By the mean value theorem, there is a point $t \in [x - \varepsilon, x + \varepsilon]$ such that

$$\begin{aligned} \langle u^*, u \rangle &= \int_J u(x) d\rho(x) + \int_{x_1-\varepsilon}^{x_1+\varepsilon} u(x) d\rho(x) \\ &\leq \max_{x \in J} |u(x)| V_J(\rho) + |u(t)| |\rho(x + \varepsilon) - \rho(x - \varepsilon)|. \end{aligned}$$

Since $|u(x)|$ achieves its maximum exactly at the point x_1 , we get $\max_{x \in J} |u(x)| < \|u\|$. Thus, it follows from (18) that

$$\langle u^*, u \rangle < \|u\| V(\rho).$$

Hence $\langle u^*, u \rangle < \|u^*\| \|u\|$. This is a contradiction.

For $N > 1$, we use a similar argument. □

1.6 Applications to Čebyšev Approximation

For the given continuous function $u_0: [a, b] \rightarrow \mathbb{R}$ on the compact interval $[a, b]$, let us consider the following *approximation problem*:

$$\max_{a \leq x \leq b} |u_0(x) - u(x)| = \min !, \quad u \in L, \quad (19)$$

where L denotes the set of all real polynomials of degree $\leq N$, for fixed $N \geq 1$. Problem (19) corresponds to the so-called Čebyšev approximation of the function u_0 by polynomials.

Proposition 1. *Problem (19) has a solution. If u is a solution of (19), then*

$$|u_0(x) - u(x)|$$

achieves its maximum at at least $N + 2$ points of $[a, b]$.

Proof. Set $X := C[a, b]$ and $\|v\| := \max_{a \leq x \leq b} |v(x)|$. Then, the original problem (19) can be written in the form

$$\|u_0 - u\| = \min !, \quad u \in L. \quad (20)$$

Since $\dim L < \infty$, this problem has a solution, by Corollary 1 in Section 1.5.

We may assume that $u_0 \notin L$. Otherwise, the statement is trivial. Let u be a solution of (20). Then, $\|u_0 - u\| > 0$. By the duality theory from Theorem 1.D, there exists a functional $u^* \in C[a, b]^*$ such that

$$\langle u^*, u_0 - u \rangle = \|u_0 - u\| \quad (21)$$

along with $\|u^*\| = 1$ and

$$\langle u^*, p \rangle = 0 \quad \text{for all } p \in L. \quad (22)$$

Suppose that $|u_0(x) - u(x)|$ achieves its maximum on $[a, b]$ at precisely the points x_1, \dots, x_M , where $1 \leq M < N + 2$. It follows from (21) and Lemma 3 in Section 1.5 that there are real numbers $\alpha_1, \dots, \alpha_M$ with $|\alpha_1| + \dots + |\alpha_M| = 1$ such that

$$u^* = \alpha_1 \delta_{x_1} + \dots + \alpha_M \delta_{x_M}.$$

Assume that $\alpha_M \neq 0$. Choose a real polynomial p of degree N such that

$$p(x_1) = p(x_2) = \dots = p(x_{M-1}) = 0 \quad \text{and} \quad p(x_M) \neq 0.$$

This is possible, since $M - 1 \leq N$. Then, $p \in L$ and $\langle u^*, p \rangle \neq 0$, contradicting (22). \square

1.7 Applications to the Optimal Control of Rockets

We want to study the motion of a vertically ascending rocket that reaches a given altitude h with minimum fuel expenditure (see Figure 1.5).

The motion $x = x(t)$ of the rocket is governed by the equation

$$\begin{aligned} mx''(t) &= \mathcal{F}(t) - mg, & 0 < t < T, \\ x(0) &= x'(0) = 0, & x(T) = h, \end{aligned} \tag{23}$$

where m = mass of the rocket, mg = force of gravity, and $\mathcal{F}(t)$ = rocket force. We neglect the loss of mass by the burning of fuel. To simplify notation, we choose physical units with $m = g = 1$.

Let us measure the minimal fuel expenditure during the time interval $[0, T]$ through the integral

$$\int_0^T |\mathcal{F}(t)| dt$$

over the rocket force \mathcal{F} . First let $T > 0$ be fixed. Then, the minimal fuel expenditure $\alpha(T)$ during the time interval $[0, T]$ is given by a solution of the following *minimum problem*:

$$\min_{\mathcal{F}} \int_0^T |\mathcal{F}(t)| dt = \alpha(T), \tag{24}$$

where we vary over all integrable functions $\mathcal{F}: [0, T] \rightarrow \mathbb{R}$. We now choose the final time $\alpha(T)$ in such a way that $\alpha(T)$ becomes *minimal*, that is,

$$\alpha(T) = \min !. \tag{25}$$

Integration of (23) yields $x(t) = \int_0^t (t - \tau) \mathcal{F}(\tau) d\tau - \frac{t^2}{2}$, and hence

$$h = \int_0^T (T - \tau) \mathcal{F}(\tau) d\tau - \frac{T^2}{2}. \tag{26}$$

Summarizing, for a given altitude $h > 0$, we have to determine the optimal thrust program $\mathcal{F}(\cdot)$ and the final time T as a solution of problems (24) through (26).

This formulation has the following shortcoming. If we consider only classical force functions \mathcal{F} , then an impulse at time t of the form

$$\text{“}\mathcal{F} = \delta_t\text{”}$$

is excluded. However, we expect that such thrust programs may be of importance. For this reason, let us consider the following *generalized problem* for functionals:

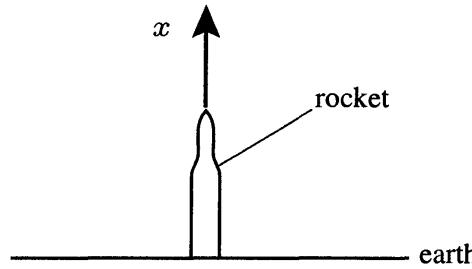


FIGURE 1.5.

(a) For a given altitude h and fixed final time $T > 0$, we are looking for a solution F of the following minimum problem:

$$\min \|F\| = \alpha(T), \quad F \in C[0, T]^*, \quad (27)$$

along with the side condition

$$h = F(w) - \frac{T^2}{2}, \quad \text{where we set } w(t) := T - t. \quad (28)$$

(b) We determine the final time T in such a way that

$$\alpha(T) = \min!. \quad (29)$$

Observe that condition (27) generalizes (24). In fact, if the functional $F \in C[0, T]^*$ has the following special form:

$$F(u) = \int_0^T u(t)\mathcal{F}(t)dt \quad \text{for all } u \in C[0, T],$$

where the fixed function $\mathcal{F}: [0, T] \rightarrow \mathbb{R}$ is continuous, then

$$\|F\| = \int_0^T |\mathcal{F}(t)|dt,$$

by Example 2 in Section 1.3.

Proposition 1. *Problem (a), (b) has the following solution:*

$$F = T\delta_0 \quad \text{and} \quad T = (2h)^{\frac{1}{2}},$$

with the minimal “fuel expenditure” $\|F\| = T$.

This solution corresponds to an impulse at the initial time $t = 0$. Proposition 1 shows that, in control theory, it is quite natural to use minimum problems with respect to functionals.

Proof. *Step 1:* Solution of problem (a). Let $X := C[a, b]$ and $L := \text{span}\{w\}$. By the Hahn–Banach theorem, there exists a functional $F_0 \in C[a, b]^*$ such that

$$F_0(w) = h + \frac{T^2}{2}.$$

Then, condition (28) says that $(F_0 - F)(w) = 0$, i.e., $(F_0 - F) \in L^\perp$. Consequently, problem (a) is equivalent to the *primal problem*:

$$\min \| (F_0 - F) - F_0 \| = \alpha(T), \quad (F_0 - F) \in L^\perp. \quad (30)$$

By Theorem 1.E in Section 1.5, the *dual problem* reads as follows:

$$\sup F_0(u) = \alpha(T), \quad u \in \text{span}\{w\}, \|u\| \leq 1. \quad (30^*)$$

Let us solve (30*) and (30). Observe that the dual problem (30*) is *one-dimensional*. Since $\|w\| = \max_{0 \leq t \leq T} |w(t)| = T$, (30*) has the solution

$$u = T^{-1}w.$$

Hence

$$\alpha(T) = F_0(T^{-1}w) = T^{-1}h + 2^{-1}T.$$

Explicitly,

$$u(t) = T^{-1}(T - t) \quad \text{for all } t \in [0, T].$$

By Theorem 1.E, the primal problem (30) has a solution $F_0 - F \in L^\perp$. Hence

$$\|F\| = \alpha(T) \quad \text{and} \quad F_0(w) = F(w).$$

Since $u = T^{-1}w$, the functional $F \in C[a, b]^*$ satisfies the equation $F(u) = F_0(u) = \alpha(T)$, i.e.,

$$F(u) = \|F\| \|u\|, \quad (31)$$

because $\|u\| = 1$. Since the functional $u(\cdot)$ achieves its maximum on $[0, T]$ precisely at the point $t = 0$, it follows from (31) and Lemma 3 in Section 1.5 that

$$F = \beta \delta_0$$

for some real number β with $|\beta| = \|F\|$. Since $\|\delta_0\| = 1$, this implies $F = \pm \|F\| \delta_0$. From $F(w) = F_0(w) > 0$ and $\delta_0(w) = w(0) > 0$, we get

$$F = \|F\| \delta_0, \quad \text{that is, } F = \alpha(T) \delta_0.$$

Step 2: Solution of problem (b). It follows from $\alpha'(T) = -T^{-2}h + 2^{-1} = 0$ that the problem $\alpha(T) = \min!$ has the solution $T = (2h)^{\frac{1}{2}}$. Hence $\alpha(T) = T^{-1}h + 2^{-1}T = (2h)^{\frac{1}{2}} = T$. \square

Problems

The concepts of “topological space” and “metric space” will be introduced in Problem 1.12ff. Important interrelations are pictured in Figure 1.6 ahead. Locally convex spaces will be defined in Problem 3.21 in connection with the weak topology (weak convergence) on Banach spaces. For example,

important spaces of generalized functions (distributions) are locally convex spaces, but not normed spaces.

1.1. Convex hull. Let M be a convex subset of the normed space X . Show that the closure \overline{M} is also convex.

1.2. The completion principle for Banach spaces. Two normed spaces X and Y over \mathbb{K} are called *normisomorphic* iff there exists a linear bijective operator $j: X \rightarrow Y$ such that j is *isometric*, i.e.,

$$\|j(u)\| = \|u\| \quad \text{for all } u \in X.$$

Let D be a normed space over \mathbb{K} . The Banach space X over \mathbb{K} is called a *completion* of D iff the set D is dense in X and the X -norm coincides with the D -norm on D .

1.2a. Uniqueness of completion. Show that two completions X and Y of D are normisomorphic.

1.2b. Existence of a completion. Show that there exists a Banach space X over \mathbb{K} that is a completion of D .

Hint: We will use the classic idea of Cantor and Méray who introduced real numbers in 1872 with the aid of equivalence classes of Cauchy sequences. Two Cauchy sequences (u_n) and (v_n) in D are called equivalent iff

$$\|u_n - v_n\| \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Let X be the set of the corresponding equivalence classes $[(u_n)]$. For $\alpha \in \mathbb{K}$, we define

$$[(u_n)] + [(v_n)] := [(u_n + v_n)], \quad \alpha[(u_n)] := [(\alpha u_n)],$$

and

$$\|[(u_n)]\| := \lim_{n \rightarrow \infty} \|u_n\|.$$

Prove that these operations make sense and that they are independent of the choice of the representatives. Cf. Zeidler (1986), Vol. 2A, p. 96.

1.3. The completion principle for Hilbert spaces. Two pre-Hilbert spaces⁶ X and Y are called *H-isomorphic* (or unitarily equivalent) iff there exists a unitary operator $j: X \rightarrow Y$. That is, j is linear, bijective, and

$$(j(u) | j(v)) = (u | v) \quad \text{for all } u, v \in X.$$

Let D be a pre-Hilbert space over \mathbb{K} . The Hilbert space X over \mathbb{K} is called a *completion* of D iff the set D is dense in X and the X -inner product coincides with the D -inner product on D .

⁶See Section 2.1 of AMS Vol. 108.

Show that there exists a completion X of D and that each completion of D is H -isomorphic to D .

Hint: Use Problem 1.2 and the fact that the inner product can be expressed by a sum of norms according to (99) in Chapter 2 of AMS Vol. 108. In particular, if $u := [(u_n)]$ and $v := [(v_n)]$, then

$$(u | v) := \lim_{n \rightarrow \infty} (u_n | v_n).$$

This limit exists and is independent of the choice of the representatives (u_n) and (v_n) of u and v , respectively.

Show that two pre-Hilbert spaces over \mathbb{K} are H -isomorphic iff they are normisomorphic.

1.4. The energetic space as a completion. Let $B: D(B) \subseteq X \rightarrow X$ be a linear, symmetric, and strongly monotone operator on the real Hilbert space X . As in Section 5.3 of AMS Vol. 108 we introduce the energetic inner product by setting

$$(u | v)_E := (Bu | v) \quad \text{for all } u, v \in D(B).$$

Show that the energetic space X_E from Section 5.3 of AMS Vol. 108 is just the completion of the domain of definition $D(B)$ with respect to $(\cdot | \cdot)_E$.

1.5. Separation of convex sets. Let A and B be nonempty convex sets in the real normed space X . Show that

- (i) A and B can be separated by a closed hyperplane provided

$$B \cap \text{int } A = \emptyset \quad \text{and} \quad \text{int } A \neq \emptyset.$$

- (ii) A and B can be strictly separated by a closed hyperplane provided $A \cap B = \emptyset$ and both A and B are open.

- (iii) A and B can be strictly separated by a closed hyperplane provided $A \cap B = \emptyset$, A is closed, and B is compact.

Hint: Use the Hahn–Banach theorem. Cf. Edwards (1994), Section 2.1.

1.6. Extension of linear positive functionals (the Krein theorem). Suppose that X is a real ordered normed space in the sense of Section 1.19 of AMS Vol. 108 with the order cone X_+ and that L is a linear subspace of X such that

$$L \cap \text{int } X_+ \neq \emptyset.$$

Let $F: L \rightarrow \mathbb{R}$ be a linear functional such that

$$F(u) \geq 0 \quad \text{for all } u \in L \text{ with } u \geq 0.$$

Show that F can be extended to a linear continuous functional $f: X \rightarrow \mathbb{R}$ such that $f(u) \geq 0$ for all $u \in X$ with $u \geq 0$.

Hint: Use the Hahn–Banach theorem along with $p(u) := \inf\{F(v): v \in L, v \geq u\}$. Cf. Edwards (1994), Section 2.5.2.

1.7*. Uniqueness of the Čebyšev approximation. Set $X := C[a, b]$, where $-\infty < a < b < \infty$ and

$$\|u - v\| := \max_{a \leq x \leq b} |u(x) - v(x)|.$$

Let L be a finite-dimensional linear subspace of X with $\dim L = N + 1$. By definition, L satisfies the *Haar condition* iff each nonzero function $v: [a, b] \rightarrow \mathbb{R}$ from L has at most N zeros. By Section 1.5, for given $u \in X$, the approximation problem

$$\|u - v\| = \min !, \quad v \in L \tag{32}$$

has a solution. In addition, the following can be shown.

- (i) If L satisfies the Haar condition, then the solution v of (32) is unique.
- (ii) Suppose that L satisfies the Haar condition. Let $u \notin L$, and let $v \in L$ be a given function. Suppose that there is a finite set of points $a \leq t_1 < t_2 < \dots < t_{N+1} \leq b$ such that

$$u(t_j) - v(t_j), \quad j = 1, \dots, N + 2,$$

attains alternately the values $\|u - v\|$ and $-\|u - v\|$ at consecutive points t_j .

Then, v is the unique solution of (32).

Study the proofs of (i) and (ii) in Kreyszig (1989), p. 340 and p. 345, respectively. It is shown in Zeidler (1986), Vol. 3, p. 181, that (i) and (ii) are special cases of a general functional analytic theorem.

In particular, (i) and (ii) apply to classic Čebyšev approximation where L is the space of all polynomials of degree $\leq N$ with real coefficients. In this case, the Haar condition is obviously satisfied.

Remark. We shall prove in Section 2.9 that the minimum problem (32) has a unique solution if X is strictly convex. Unfortunately, the space $C[a, b]$ is *not* strictly convex. Therefore, we need a more subtle uniqueness proof.

1.8.* A special case of the famous Pontrjagin maximum principle. This principle plays a fundamental role in the optimal control of time-dependent processes in technology and economics. Let us consider the following control problem with *fixed end time*. For a given time interval $[0, T]$ with $T > 0$, we are looking for a process $x: [0, T] \rightarrow \mathbb{R}$ and a piecewise-continuous control function $u: [0, T] \rightarrow \mathbb{R}$ such that the following hold:

(a) Control functional:

$$\int_0^T L(x(t), u(t)) dt = \min !.$$

(b) Control equation:

$$\begin{aligned} x'(t) &= v(x(t), u(t)) && \text{on } [0, T], \\ x(0) &= \text{fixed}. \end{aligned}$$

(c) Control restriction:

$$u(t) \in U \quad \text{on } [0, T].$$

Here, U is a prescribed subset of \mathbb{R}^m , and we assume that the prescribed functions $L, v: \mathbb{R} \times \mathbb{R}^m \rightarrow \mathbb{R}$ are C^1 . Following Pontrjagin, we introduce the generalized Hamiltonian

$$\mathcal{H}(x, u, p) := pv(x, u) - L(x, u)$$

along with the generalized canonical equation

$$\begin{aligned} p'(t) &= -\mathcal{H}_x(x(t), u(t), p(t)) && \text{on } [0, T], \\ p(T) &= 0. \end{aligned} \tag{33}$$

Suppose that $x = x(t), u = u(t)$ is a solution to the original problem (a) through (c) and let $p = p(t)$ be the solution to (33). Then, the following maximum principle holds:

$$\mathcal{H}(x(t), u(t), p(t)) = \max_{u \in U} \mathcal{H}(x(t), u, p(t)). \tag{34}$$

Study the proof of this theorem in Luenberger (1969), p. 263. The proof relies on the concept of the adjoint operator and the F -derivative from Chapter 4. The situation becomes much more complicated if the end time is *free*. A proof of the general Pontrjagin maximum principle can be found in Zeidler (1986), Vol. 3, p. 422. This proof is based on a general functional analytic theorem.

1.9. The relation of Pontrjagin's maximum principle to classical mechanics. In the special case where

$$v(x, u) := u, \quad U := \mathbb{R},$$

we get $x'(t) = u(t)$. It follows from (34) that $\mathcal{H}_u(x(t), u(t), p(t)) = 0$, and hence

$$p(t) = L_{x'}(x(t), x'(t)).$$

Then, problem 1.8(a) corresponds to the principle of least action in mechanics, where $x = x(t)$ describes the trajectory of a particle. The control function u coincides with the velocity of the particle, and p is called the momentum of the particle. In particular, equation (33) coincides with the equation of motion

$$\frac{d}{dt} L_{x'}(x(t), x'(t)) = L_x(x(t), x'(t)).$$

This is exactly the *Euler–Lagrange* equation (5) from Section 2.2, and it corresponds to the minimum problem 1.8(a).

1.10. *Application of Pontrjagin's maximum principle to the farmer's allocation problem* [from Luenberger (1969)]. A farmer produces a single crop such as wheat. After harvesting his crop, he may store it or sell and reinvest it by buying additional land and equipment to increase his production rate. The farmer wishes to *maximize* the total amount stored during the time interval $[0, T]$. Set

$$\begin{aligned} x(t) &:= \text{rate of production at time } t, \\ x_r(t) &:= \text{rate of reinvestment at time } t, \\ x_s(t) &:= \text{rate of storage at time } t. \end{aligned}$$

Then, the stored production during the time interval $[0, \tau]$ is equal to $\int_0^\tau x_s(t) dt$. Thus, we get the following maximum problem:

$$\int_0^T x_s(t) dt = \max !.$$

Obviously,

$$x(t) = x_r(t) + x_s(t) \quad \text{for all } t \in [0, T].$$

Moreover, we assume that

$$x_r(t) = cx'(t) \quad \text{for all } t \in [0, T] \text{ and fixed } c > 0.$$

Roughly speaking, this says that the reinvestment rate increases if the production accelerates. Define

$$u(t) := \begin{cases} \frac{x_r(t)}{x(t)} & \text{if } x(t) \neq 0 \\ 0 & \text{if } x(t) = 0. \end{cases}$$

Hence $x_r(t) = u(t)x(t)$ and $0 \leq u(t) \leq 1$. Since $x_s(t) = x(t) - x_r(t)$, we obtain the following control problem:

$$\int_0^T (1 - u(t))x(t) dt = \max !, \tag{35}$$

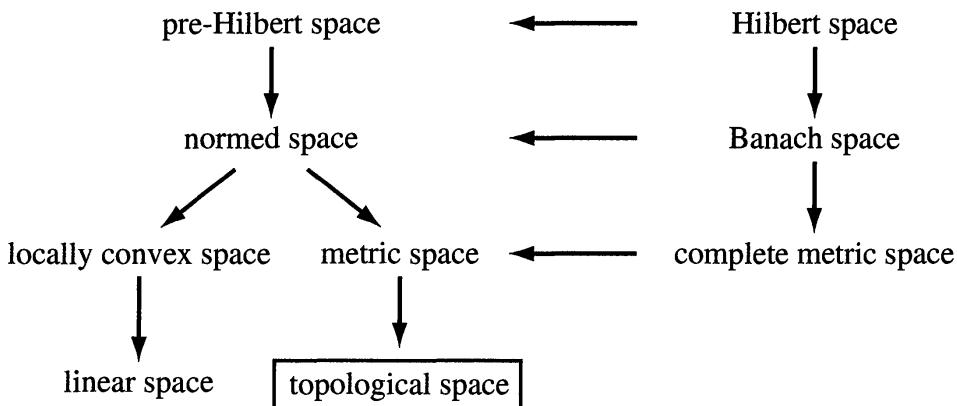


FIGURE 1.6.

$$\begin{aligned} cx'(t) &= u(t)x(t) \quad \text{on } [0, T], \quad x(0) = \text{fixed} > 0, \\ 0 &\leq u(t) \leq 1 \quad \text{on } [0, T]. \end{aligned}$$

Let $c := 1$ and assume that $T > 1$. Using the Pontrjagin maximum principle from Problem 1.8, show that this problem has the following quite natural solution:

$$u(t) = \begin{cases} 1 & \text{if } 0 \leq t \leq T - 1 \\ 0 & \text{if } T - 1 < t \leq T. \end{cases}$$

This means that the farmer stores nothing until time $T - 1$, at which point he stores all products. Such so-called bang–bang controls are typical for many control problems.

Hint: Cf. Luenberger (1969), p. 265.

1.11. Further optimization problems. Applications of separation theorems to general classes of optimization problems can be found in Zeidler (1986), Vol. 3, Chapters 47ff. As an elementary introduction to optimization theory, we recommend the monograph by Luenberger (1969).

1.12. Topological spaces.⁷ The most general class of “spaces” used in analysis is the class of topological spaces. Let us discuss the relation between normed spaces and topological spaces. Figure 1.6 tells us that each normed space is a metric space, and so forth.

1.12a. Definition. A set M is called a *topological space* iff there exists a system \mathcal{T} of subsets of M that has the following properties:

- (T1) $M \in \mathcal{T}$ and $\emptyset \in \mathcal{T}$.
- (T2) If $U_1, \dots, U_n \in \mathcal{T}$ for any natural number n , then $\bigcap_{j=1}^n U_j \in \mathcal{T}$.
- (T3) If $U_\alpha \in \mathcal{T}$ for all $\alpha \in A$, where A is an arbitrary index set, then $\bigcup_{\alpha \in A} U_\alpha \in \mathcal{T}$.

⁷The classic introduction to general topology is Kelley (1955).

The system \mathcal{T} is called a *topology*. A subset U of the topological space M is called *open* iff $U \in \mathcal{T}$.

A subset C of M is called *closed* iff the complement $M - C$ is open.

1.12b. Properties of closed sets. Let M be a topological space. Show that

- (i) M and the empty set \emptyset are closed.
- (ii) The union of a finite number of closed sets in M is again closed.
- (iii) The intersection of an arbitrary number of closed sets in M is again closed.

1.12c. Normed spaces. Let X be a normed space over \mathbb{K} . In Section 1.4 of AMS Vol. 108 we defined open sets in X . Show that these open sets of X form a topology of X (i.e., each normed space is also a topological space).

1.12d. Subsets of normed spaces. Let M be a subset of a normed space. A subset U of M is called *relatively open* iff there exists an open subset U_X of X such that

$$U = U_X \cap M.$$

Show that the relatively open sets of M form a topology. This way, M becomes a topological space.

1.12e. Neighborhoods. A subset $U(u)$ of the topological space M is called a *neighborhood* of the point u iff there exists an open set W such that

$$u \in W \subseteq U(u) \quad (\text{see Figure 1.7(a)}).$$

Show that, in a normed space, each ε -neighborhood $U_\varepsilon(u)$ of the point u (cf. Section 1.4 of AMS Vol. 108) is also a neighborhood of u in the general sense of topological spaces.

1.12f. Separation. A topological space M is called *separated* iff, for each pair u, v of different points in M , there are neighborhoods $U(u)$ and $U(v)$ such that

$$U(u) \cap U(v) = \emptyset \quad (\text{see Figure 1.7(b)}).$$

Show that each normed space is separated.

1.12g. Convergence. Let (u_n) be a sequence in a topological space. We write

$$u_n \rightarrow u \quad \text{as } n \rightarrow \infty$$

iff, for each neighborhood U of the point u , there exists a natural number n_U such that

$$u_n \in U \quad \text{for all } n \geq n_U.$$

Show that, in a normed space, this definition is equivalent to the definition given in Section 1.2 of AMS Vol. 108.

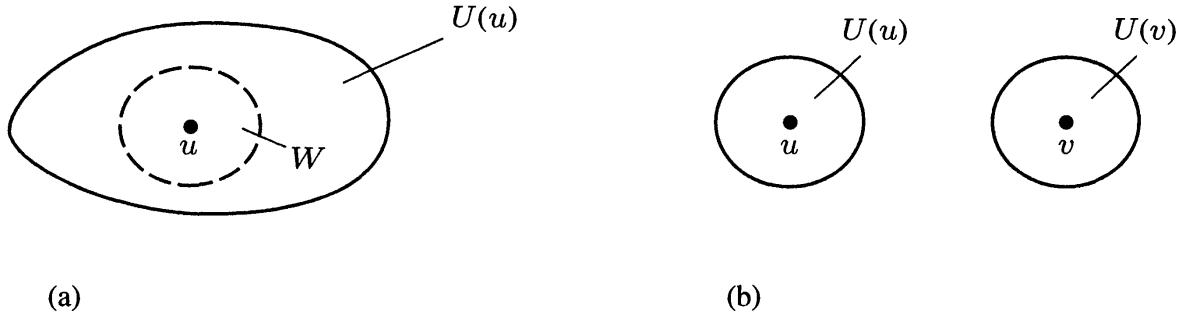


FIGURE 1.7.

Show that in a separated topological space the limit point of a convergent sequence is unique.

1.13. *Continuity in topological spaces.* Let M and Y be topological spaces. The map

$$f: M \rightarrow Y \quad (36)$$

is called *continuous* at the point $u \in M$ iff, for each neighborhood $U(f(u))$, there is a neighborhood $U(u)$ such that

$$f(U(u)) \subseteq U(f(u)).$$

Moreover, the map f from (36) is called *continuous* iff it is continuous at each point $u \in M$.

1.13a. *Preimages of continuous maps.* Show that the following three statements are mutually equivalent for the map f from (36):

- (i) f is continuous.
- (ii) The preimage $f^{-1}(W)$ of each open set W is again open.
- (iii) The preimage $f^{-1}(C)$ of each closed set C is again closed.

1.13b. *Continuity in normed spaces.* Let X and Y be normed spaces over \mathbb{K} , and let M be a subset of X . Show that the definition of continuity from Section 1.9 of AMS Vol. 108 coincides with the general definition in topological spaces.

1.14. *Compactness in topological spaces.* Let M be a subset of a topological space X (e.g., X is a normed space). The set M is called *compact* iff each open covering of M contains a *finite* subcovering. That is, each family $\{U_\alpha\}$ of open sets U_α with

$$M \subseteq \bigcup_{\alpha} U_\alpha$$

contains a finite subfamily $\{U_{\alpha_1}, \dots, U_{\alpha_n}\}$ such that

$$M \subseteq \bigcup_{j=1}^n U_{\alpha_j}.$$

M is called *relatively compact* iff the closure \overline{M} is compact.

1.14a. Compactness and continuity. Let X and Y be topological spaces, and let $f: M \subseteq X \rightarrow Y$ be a continuous map on the compact set M .

Show that $f(M)$ is also compact.

Hint: Use Problem 1.13a.

1.14b. The finite intersection property. A system \mathcal{S} of sets is called *centered* iff the intersection of finitely many sets in \mathcal{S} is never empty.

Show that a topological space M is compact iff every centered system of closed sets in M has a nonempty intersection.

1.15. Compactness in normed spaces. Let M be a subset of the normed space X over \mathbb{K} . Then, M is called *precompact* iff either $M = \emptyset$ or $M \neq \emptyset$ and M has a finite ε -net⁸ for each $\varepsilon > 0$. Moreover, M is called *complete* iff each Cauchy sequence in M converges to a point in M .

1.15a. The compactness theorem. Show that the following three statements are mutually equivalent:

- (i) M is compact.
- (ii) M is sequentially compact.
- (iii) M is precompact and complete.

The proof will be given ahead.

1.15b. The relative compactness theorem. Use Problem 1.15a in order to show that

- (a) M is relatively compact iff M is relatively sequentially compact.
- (b) If M is relatively compact, then M is precompact. The converse holds true if X is complete.

Solution: Ad (a). Let M be relatively compact, that is, the closure \overline{M} is compact. By Problem 1.15a, \overline{M} is sequentially compact. Thus, each sequence (u_n) in M has a convergent subsequence (i.e., $u_{n'} \rightarrow u$ as $n' \rightarrow \infty$ with $u \in \overline{M}$). Hence M is relatively sequentially compact.

Conversely, let M be relatively sequentially compact. If (u_n) is a sequence in the closure \overline{M} , then there exists a sequence (v_n) in M such that

$$\|u_n - v_n\| \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Since M is relatively sequentially compact, there exists a convergent subsequence $(v_{n'})$, that is,

$$v_{n'} \rightarrow v \quad \text{as } n \rightarrow \infty.$$

⁸The definition of a finite ε -net can be found in Section 1.11 of AMS Vol. 108.

Hence $\|u_{n'} - v\| \leq \|u_{n'} - v_{n'}\| + \|v_{n'} - v\| \rightarrow 0$ as $n' \rightarrow \infty$, that is,

$$u_{n'} \rightarrow v \quad \text{as } n' \rightarrow \infty \text{ and } v \in \overline{M}.$$

Thus, \overline{M} is sequentially compact. By Problem 1.15a, \overline{M} is compact, and hence M is relatively compact.

Ad (b). Let M be relatively compact. Then, \overline{M} is compact. By Problem 1.13a, \overline{M} is precompact. It follows easily that this implies the precompactness of M .

Conversely, let M be precompact and let X be complete. The proof of Proposition 10 in Section 1.11 of AMS Vol. 108 tells us that M is relatively sequentially compact. By (a), M is relatively compact.

1.15c. Proof of the compactness theorem from Problem 1.15a.

(i) \Rightarrow (ii). Let (u_n) be a sequence in M . Set

$$A_n := \{u_n, u_{n+1}, \dots\}.$$

We first show that there exists a point $u \in M$ such that

$$u \in \bigcap_{n=1}^{\infty} \bar{A}_n. \quad (37)$$

Otherwise, for each $u \in M$, there is an index m such that $u \notin \bar{A}_m$, i.e.,

$$M \subseteq \bigcup_{n=1}^{\infty} (X - \bar{A}_n).$$

Since $X - \bar{A}_n$ is open and M is compact, there exists a finite number of indices, say $n = 1, \dots, k$, such that

$$M \subseteq \bigcup_{n=1}^k (X - \bar{A}_n) \subseteq \bigcup_{n=1}^k (X - A_n).$$

This is a contradiction, since $u_k \in M$ and $u_k \in A_n$ for all $n = 1, \dots, k$.

(ii) \Rightarrow (iii). Let M be sequentially compact. If (u_n) is a Cauchy sequence in X , then there exists a convergent subsequence, that is, $u_{n'} \rightarrow u$ as $n' \rightarrow \infty$ with $u \in M$. By Proposition 7 in Section 1.3 of AMS Vol. 108, $u_n \rightarrow u$ as $n \rightarrow \infty$ (i.e., M is *complete*).

It follows as in the proof of Proposition 10 in Section 1.11 of AMS Vol. 108 that M is *precompact*.

(iii) \Rightarrow (i). Set $B_r(u) := \{v \in X : \|u - v\| \leq r\}$. Let $\{U_\alpha\}$ be an open covering of M . Suppose that there is no finite subfamily of $\{U_\alpha\}$ that covers M . We want to construct a sequence (u_n) in M such that, for all $n = 1, 2, \dots$,

$$B_{2^{-(n+1)}}(u_{n+1}) \cap B_{2^{-n}}(u_n) \neq \emptyset \quad (38)$$

and

$$B_{2^{-n}}(u_n) \text{ is not covered by a finite subfamily of } \{U_\alpha\}. \quad (39)$$

In fact, since M has a finite 2^{-1} -net, there exist points $v_1, \dots, v_k \in M$ such that the family

$$B_{2^{-1}}(v_1), \dots, B_{2^{-1}}(v_k)$$

covers M . Thus, there exists some point v_m ($1 \leq m \leq k$) such that $B_{2^{-1}}(v_m)$ is not covered by a finite subfamily of $\{U_\alpha\}$. Set $u_1 = v_m$. This is (39) for $n = 1$.

Since M has a finite 2^{-2} -net, there exists a ball $B_{2^{-2}}(u_2)$ with $u_2 \in M$ and

$$B_{2^{-1}}(u_1) \cap B_{2^{-2}}(u_2) \neq \emptyset$$

such that $B_{2^{-2}}(u_2)$ is not covered by a finite subfamily of $\{U_\alpha\}$. This is (38) and (39) for $n = 2$.

Now use an induction argument in order to prove (38) and (39) for $n > 2$.

According to (38),

$$\|u_{n+1} - u_n\| \leq 2^{-n-1} \quad \text{for all } n = 1, 2, \dots.$$

It follows from Corollary 8 in Section 1.3 of AMS Vol. 108 that (u_n) is Cauchy. Since M is complete, we get

$$u_n \rightarrow u \quad \text{as } n \rightarrow \infty \text{ and } u \in M.$$

Choose an index β such that $u \in U_\beta$. Since U_β is open, there is an $r > 0$ such that

$$B_{2r}(u) \subseteq U_\beta.$$

If m is sufficiently large, then $\|u - u_m\| \leq r$ with $2^{-m} \leq r$. By the triangle inequality,

$$B_{2^{-m}}(u_m) \subseteq B_{2r}(u) \subseteq U_\beta.$$

This contradicts (39).

1.16. The generalized Weierstrass theorem. Let $f: M \subseteq X \rightarrow \mathbb{R}$ be a continuous function on the nonempty compact subset M of the topological space X .

Show that f attains its maximum and minimum on M .

Solution: By Problem 1.14a, the set $f(M)$ is compact in \mathbb{R} and hence is closed and bounded. Consequently, the real numbers $\inf_{u \in M} f(u)$ and $\sup_{u \in M} f(u)$ are contained in M .

1.17. The Banach space $C(M, Y)$. Let M be a nonempty compact subset of a topological space, and let Y be a Banach space over \mathbb{K} . Let $C(M, Y)$ denote the set of all continuous functions $f: M \rightarrow Y$.

Show that $C(M, Y)$ is a Banach space over \mathbb{K} equipped with the norm

$$\|f\| := \max_{u \in M} \|f(u)\|.$$

1.18. Metric spaces. A nonempty set M is called a *metric space* iff there exists a function $d: M \rightarrow [0, \infty[$ such that, for all $u, v, w \in X$, the following hold:

- (i) $d(u, v) = 0$ iff $u = v$.
- (ii) $d(u, v) = d(v, u)$.
- (iii) $d(u, w) \leq d(u, v) + d(v, w)$ (triangle inequality).

The number $d(u, v)$ is called the *distance* between the two points u and v .

By convention, empty sets are also called metric spaces.

1.18a. The translation principle. Show that each subset M of a normed space X becomes a metric space by setting

$$d(u, v) := \|u - v\| \quad \text{for all } u, v \in M. \quad (40)$$

Using (40), we can directly translate many notions and propositions from normed spaces to metric spaces. For example, we say that a sequence (u_n) in the metric space M converges to the point $u \in M$ iff

$$\lim_{n \rightarrow \infty} d(u_n, u) = 0.$$

A sequence (u_n) in the metric space M is called *Cauchy* iff, for each $\varepsilon > 0$, there is a number $n_0(\varepsilon)$ such that

$$d(u_n, u_m) < \varepsilon \quad \text{for all } n, m \geq n_0(\varepsilon).$$

A metric space M is called *complete* iff each Cauchy sequence in M is convergent.

1.18b. Topology. A subset U of the metric space M is called *open* iff, for each $u \in U$, there is some $\varepsilon > 0$ such that the set

$$\{v \in X : d(u, v) < \varepsilon\}$$

is contained in U .

Show that the collection of all these open sets forms a topology on M and that this way each metric space becomes a separated topological space.

1.18c. Compactness in metric spaces. Show that all the compactness statements of Problem 1.15 remain valid if we replace normed spaces with

metric spaces. Convince yourself that the corresponding proofs for normed spaces can be directly translated to metric spaces.

1.19. *Some fundamental theorems.* Study the proofs of the following results.

1.19a.* *The Stone–Weierstrass approximation theorem.* Let M be a nonempty compact subset of a separated topological space. Let \mathcal{P} be a family of continuous functions $f: M \rightarrow \mathbb{K}$ such that the following hold:

- (i) \mathcal{P} is an algebra, i.e., if $f, g \in \mathcal{P}$, then $fg \in \mathcal{P}$ and $\alpha f + \beta g \in \mathcal{P}$ for all $\alpha, \beta \in \mathbb{K}$.
- (ii) \mathcal{P} contains the constant functions on M .
- (iii) \mathcal{P} separates the points of M , that is, if $u, v \in M$ and $u \neq v$, then there exists a function $p \in \mathcal{P}$ such that $p(u) \neq p(v)$.

Then, the set \mathcal{P} is *dense* in the Banach space $C(M, \mathbb{K})$.

Hint: Cf. Yosida (1988), introduction.

1.19b. *The Weierstrass approximation theorem in \mathbb{R}^n .* Let M be a nonempty compact set in \mathbb{R}^n . Use Problem 1.19a in order to show that the set \mathcal{P} of all polynomials $p: M \rightarrow \mathbb{R}$ in n variables with real coefficients is dense in the Banach space $C(M)$ of real continuous functions f on M .

Explicitly, this means that for each continuous function $f: M \rightarrow \mathbb{R}$ and each $\varepsilon > 0$ there is a real polynomial $p: M \rightarrow \mathbb{R}$ such that

$$|f(u) - p(u)| < \varepsilon \quad \text{for all } u \in M.$$

1.19c.* *The general Arzelà–Ascoli theorem.* Let M be a nonempty compact set of a metric space, and let Y be a Banach space over \mathbb{K} . Then, the family \mathcal{F} of continuous functions $f: M \rightarrow Y$ is a relatively compact subset of the Banach space $C(M, Y)$ iff the following two conditions are satisfied:

- (i) For each $u \in M$, the set $\{f(u): f \in \mathcal{F}\}$ is relatively compact in Y .
- (ii) \mathcal{F} is equicontinuous, that is, for each $u \in M$ and each $\varepsilon > 0$, there is a number $\delta(u, \varepsilon) > 0$ independent of f such that, for all $f \in \mathcal{F}$,

$$d(v, u) < \delta \quad \text{implies} \quad |f(v) - f(u)| < \varepsilon.$$

Hint: Cf. Dieudonné (1969), Section 7.5.

1.19d.* *The Tietze–Urysohn extension theorem.* Let $f: M \subseteq X \rightarrow \mathbb{R}$ be a continuous function on the nonempty closed subset of the metric space X . Then there exists a continuous extension $F: X \rightarrow \mathbb{R}$ of f such that

$$\inf_{u \in M} f(u) \leq F(v) \leq \sup_{u \in M} f(u) \quad \text{for all } v \in X.$$

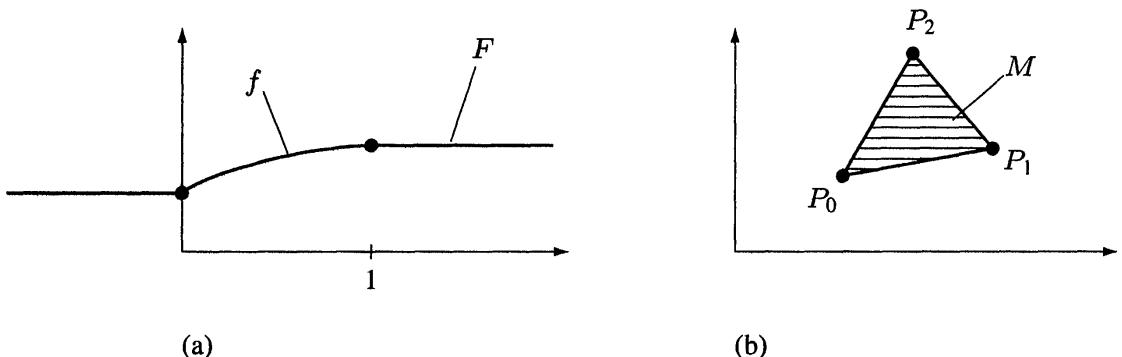


FIGURE 1.8.

Hint: Cf. Dieudonné (1969), Section 4.5. In the special case where $M := [0, 1]$ and $X := \mathbb{R}$, the intuitive meaning of this theorem is pictured in Figure 1.8(a).

1.19e.* The Krein–Milman convexity theorem. Let M be a nonempty convex compact subset of a real normed space X . Then

$$M = \overline{\text{co}} \mathcal{E}(M), \quad (41)$$

where $\mathcal{E}(M)$ denotes the set of extreme points of M . By definition, u is an extreme point of M iff

$$u = tv + (1 - t)w \quad \text{with} \quad v, w \in M \text{ and } 0 < t < 1$$

implies $v = w$.

Hint: Cf. Yosida (1988), Chapter 12. In the special case where M is a closed triangle in \mathbb{R}^2 , precisely the three vertices P_0 , P_1 , and P_2 are extreme points of M (see Figure 1.8(b)). Here, statement (41) says that a closed triangle is equal to the closed convex hull of its vertices.

1.19f. Application to linear optimization. The minimum problem

$$F(u) = \min !, \quad u \in M$$

has a solution u , where u is an extreme point of M , provided $F: M \subseteq X \rightarrow \mathbb{R}$ is a linear continuous functional on the nonempty convex compact subset M of the real normed space X .

1.20.* The structure of C^* -algebras.

1.20a.* The GNS-theorem (Gelfand–Naimark–Segal representation theorem). Let ω be a state⁹ of a C^* -algebra \mathfrak{A} . Then, there exist a complex Hilbert space X and a *-homomorphism $\phi: \mathfrak{A} \rightarrow L(X, X)$ such that

$$\omega(A) := (u | \phi(A)u) \quad \text{for all } A \in \mathfrak{A} \text{ and fixed } u \in X \text{ with } \|u\| = 1.$$

⁹See Section 5.18 of AMS Vol. 108.

In addition, u is *cyclic*, that is, by definition, the set $\{\phi(A)u: A \in \mathfrak{A}\}$ is dense in X .

Study the proof in Kadison and Ringrose (1983), Vol. 1, p. 278.

1.20b.* *The Gelfand–Naimark representation theorem.* Each C^* -algebra \mathfrak{A} is $*$ -isomorphic to a C^* -subalgebra of $L(X, X)$ for some complex Hilbert space X .

More precisely, there exists an injective $*$ -homomorphism $\phi: \mathfrak{A} \rightarrow L(X, X)$ such that each state ω of \mathfrak{A} has the form

$$\omega(A) = (u \mid \phi(A)u) \quad \text{for all } A \in \mathfrak{A} \text{ and fixed } u \in X \text{ with } \|u\| = 1.$$

Study the proofs in Kadison and Ringrose (1983), Vol. 1, p. 281.

1.20c.* *The Gelfand theorem.* Each commutative C^* -algebra is $*$ -isomorphic to $C(M, \mathbb{C})$ for some compact topological space M .

Recall that $C(M, \mathbb{C})$ consists of all the continuous functions $f: M \rightarrow \mathbb{C}$ with the $*$ -operation defined through

$$f^*(x) := \overline{f(x)} \quad \text{for all } x \in M.$$

Study the proof in Berberian (1974), p. 223.

1.21.* *Applications of C^* -algebras to spectral theory.* Study Rudin (1973), Chapters 12 and 13.

1.22.* *Applications of C^* -algebras and von Neumann algebras to quantum statistics.* Study Bratteli and Robinson (1979), Vol. 2 and Simon (1993).

1.23. Density and duality. Let X and Y be Banach spaces over \mathbb{K} such that the embedding

$$X \subseteq Y$$

is continuous, and X is dense in Y . Show that the following are met:

- (i) The embedding $Y^* \subseteq X^*$ is continuous.
- (ii) If X is reflexive, then Y^* is dense in X^* .

Hint: To prove (ii), use the Hahn–Banach theorem. Cf. Zeidler (1986), Vol. 2A, p. 98.

2

Variational Principles and Weak Convergence

Johann Bernoulli, professor of mathematics, greets the most sophisticated mathematicians in the world. Experience shows that noble intellectuals are driven to work for pursuit of knowledge by nothing more than being confronted with difficult and useful problems.

Six months ago, in the June edition of the *Leipzig Acta Eruditorum*, I presented such a problem. The allotted six-month deadline has now gone by, but no trace of a solution has appeared. Only the famous Leibniz informed me that he had unraveled the knot of this brilliant and outstanding problem, and he kindly asked me to extend the deadline until next Easter. I agreed to this honourable request.... I will repeat the problem here once more.

Two points, at different distances from the ground and not in a vertical line, should be connected by such a curve so that a body under the influence of gravitational forces passes in the shortest possible time from the upper to the lower point.¹

Johann Bernoulli, January 1697

How does one apply the methods of maxima and minima in the determination of unknown curves?

Leonhard Euler, 1744

The famous Euler succeeded in tracing back to a general method all investigations on variational problems. But however sophisticated

¹The solution of this classic problem will be given in Problem 2.1.

and fruitful his method may be, one has to admit that it is not simple. Here one finds a method which only uses simple principles of calculus.

Joseph Louis Lagrange, 1762

By generalizing Euler's method, Lagrange got the idea for his remarkable formulas, where in a single line there is contained the solution of all problems of analytic mechanics.

Carl Gustav Jakob Jacobi (1804–1851)

The Euler "Calculus of Variations" from 1744 is one of the most beautiful mathematical works that has ever been written.

Constantin Carathéodory (1873–1950)

Mathematics knows, besides the exclusive area of the Greeks, no luckier constellation than the one under which Leonhard Euler (1707–1783) was born. It was up to him to give mathematics a completely changed form and to shape it into the powerful edifice that it is today.²

Andreas Speiser (1885–1970)

The classical Weierstrass existence theorem from Section 1.11 in AMS Vol. 108 tells us the following:

(W) *The minimum problem*

$$F(u) = \min !, \quad u \in M, \tag{1}$$

has a solution provided the functional $F: M \rightarrow \mathbb{R}$ is continuous on the nonempty compact subset M of the Banach space X .

Unfortunately, this result is useless for many variational problems because of the following crucial fact:

In infinite-dimensional Banach spaces, closed balls are not compact.

This is the decisive difficulty in the calculus of variations. To overcome this difficulty, we shall introduce the notion of *weak convergence*. The basic result reads as follows:

(C) *In a reflexive Banach space, each bounded sequence has a weakly convergent subsequence.*

²Seen statistically, Euler must have made a discovery every week. He wrote nearly 900 research papers and 5,000 letters. His *Collected Papers* comprise 72 volumes.

In particular, every Hilbert space is a reflexive Banach space. For Hilbert spaces, the convergence principle (C) is a consequence of the *Riesz theorem* from Section 2.10 in AMS Vol. 108.

In the case of reflexive Banach spaces, we need some results about linear continuous functionals that are consequences of the *Hahn–Banach theorem*.

The reflexivity of the Banach space X implies

$$X = X^{**},$$

that is, the bidual space $X^{**} = (X^*)^*$ can be identified with the original space X . Consequently, reflexive Banach spaces are closely related to the concept of *duality*. Roughly speaking, in the case of a Hilbert space X , we get

$$X = X^*;$$

in other words, the dual space X^* can be identified with the original space X by means of the Riesz theorem. This implies the reflexivity $X = X^{**}$.

In finite-dimensional Banach spaces, the weak convergence coincides with the usual convergence. The fundamental notion of weak convergence in Hilbert spaces was introduced by Hilbert in 1906.

The convergence principle (C) implies the following fundamental generalization of the classical Weierstrass theorem (W).

(W*) *The minimum problem (1) has a solution provided the functional $F: M \rightarrow \mathbb{R}$ is weakly sequentially lower semicontinuous on the closed ball M of the reflexive Banach space X .*

More generally, this remains true if M is a nonempty bounded closed convex set in the reflexive Banach space X . In particular, we shall show that the following result is an easy consequence of (W*).

The minimum problem (1) with $M = X$ has a solution provided the functional $F: X \rightarrow \mathbb{R}$ is convex and continuous on the reflexive Banach space X and $F(u) \rightarrow +\infty$ as $\|u\| \rightarrow \infty$.

It turns out that the theory of infinite-dimensional minimum problems allows a simple formulation if *convexity* is involved (i.e., both the set M and the functional F are convex).

We now want to discuss in which way minimum problems are closely related to operator equations. Suppose that the original minimum problem (1) has a solution $u_0 \in \text{int } M$, that is, u_0 is an *inner* point of M . Then

$$F'(u_0) = 0 \quad (\text{Euler equation}), \tag{1*}$$

where $F'(u_0)$ denotes the Gâteaux derivative, which will be introduced in Section 2.1. Formally, equation (1*) looks like the equation in classical analysis where F is a real function. Here, condition (1*) means that the tangent line is *horizontal* at the minimal point u_0 (cf. Figure 2.1(a)).

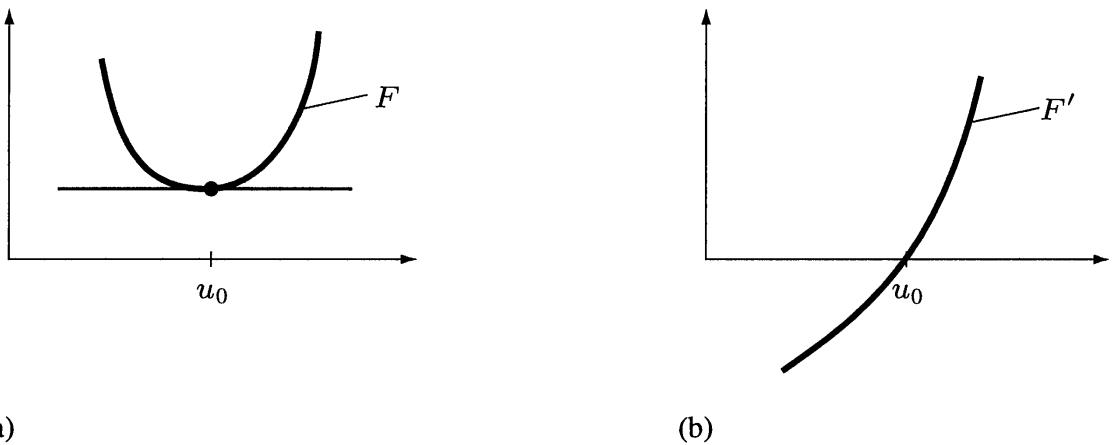


FIGURE 2.1.

In the case where $F: M \subseteq X \rightarrow \mathbb{R}$ is a functional, equation (1 *) represents an *operator equation* for the operator $F': M \subseteq X \rightarrow X^*$. This way it is possible to solve operator equations of the form (1 *) by considering the corresponding minimum problem (1).

If the solution u_0 of the minimum problem (1) is *not* an inner point of the convex set M , then we get

$$\langle F'(u_0), v - u_0 \rangle \geq 0 \quad \text{for all } v \in M. \quad (1^{**})$$

This is called *variational inequality*.

Finally, let us explain why our considerations about *convex minimum problems* are closely related to the theory of *monotone operators*. The operator $A: X \rightarrow X^*$ on the reflexive Banach space X is called monotone iff

$$\langle Au - Av, u - v \rangle \geq 0 \quad \text{for all } u, v \in X.$$

If the functional $F: X \rightarrow \mathbb{R}$ is *convex*, then its Gâteaux derivative $F': X \rightarrow X^*$ is *monotone*, that is, each solution u_0 of the convex minimum problem (1) is also a solution of the monotone operator equation (1 *). In Section 2.18, we will use the *Galerkin method* in order to solve the more general operator equation

$$Au_0 = 0, \quad (1^{***})$$

where $A: X \rightarrow X^*$ is a monotone operator that is not necessarily the Gâteaux derivative F' of a functional F .

The relation between convex functionals and monotone operators generalizes the following well-known fact from classical analysis:

If the real function F is convex, then the derivative F' is monotone (cf. Figure 2.1).

Convexity plays a fundamental role in the mathematical description of nature. For example, by the first and second laws of thermodynamics, all the processes in nature are governed by energy E and entropy S . Observe

that the negative entropy $-S$ is a convex functional, and the energy E is frequently a convex functional.

The following maximum problem

$$F(u) = \max!, \quad u \in M,$$

can always be reduced to a minimum problem by replacing the functional F with $-F$.

For the convenience of the reader, we first present an elementary approach to variational principles, using only very simple facts about Banach spaces and Hilbert spaces. This can be found in Sections 2.1 through 2.7. The generalizations to reflexive Banach spaces via the Hahn–Banach theorem will be considered in Section 2.8.

The applications in this chapter concern the calculus of variations, nonlinear eigenvalue problems, variational inequalities, duality theory, game theory, and nonlinear monotone operators.

2.1 The n th Variation

Recall that each map $F: M \rightarrow \mathbb{K}$ with values in $\mathbb{K} = \mathbb{R}, \mathbb{C}$ is called a *functional*. By an open neighborhood $U(u_0)$ of the point u_0 in the normed space X , we understand an open set in X with $u_0 \in U(u_0)$.

Definition 1. Let $F: U(u_0) \subseteq X \rightarrow \mathbb{R}$ be a functional on the open neighborhood $U(u_0)$ of the point u_0 in the real normed space X . For fixed $h \in X$, set

$$\phi(t) := F(u_0 + th),$$

where the real parameter t lives in an open neighborhood of the point $t = 0$.

By the *n th variation* $\delta^n F(u_0; h)$ of the functional F at the point u_0 in the direction h , we understand

$$\delta^n F(u_0; h) := \phi^{(n)}(0), \quad n = 1, 2, \dots.$$

In particular, the first variation is given through $\delta F(u_0; h) := \phi'(0)$.

The functional F has a *Gâteaux derivative* $F'(u_0)$ at the point u_0 iff the first variation $\delta F(u_0; h)$ exists for each $h \in X$ and there exists a linear continuous functional $F'(u_0)$ on X such that

$$\delta F(u_0; h) = F'(u_0)(h) \quad \text{for all } h \in X.$$

The Gâteaux derivative $F'(u_0)$ is called a *Fréchet derivative* iff

$$F(u_0 + h) - F(u_0) = F'(u_0)(h) + \|h\|\varepsilon(h)$$

for all $h \in X$ in an open neighborhood of $h = 0$, where $\varepsilon(h) \rightarrow 0$ as $h \rightarrow 0$. Obviously, the following condition holds:

If the Fréchet derivative $F'(u_0)$ exists, then F is continuous at the point u_0 .

The functional $F: U \subseteq X \rightarrow \mathbb{R}$ on the open set U of the normed space X is called a C^1 -functional iff the Fréchet derivative $F'(u)$ exists for all $u \in U$ and the operator $F': U \rightarrow X^*$ is continuous.

Definition 2. Let the functional $F: U(u_0) \subseteq X \rightarrow \mathbb{R}$ be as given in Definition 1. Then, F has a *local minimum* (resp., local maximum) at the point u_0 iff there is an open neighborhood $V(u_0)$ of the point u_0 such that $V(u_0) \subseteq U(u_0)$ and

$$F(u) \geq F(u_0) \quad \text{for all } u \in V(u_0)$$

(resp., $F(u) \leq F(u_0)$ for all $u \in V(u_0)$).

The functional F has the *critical point* u_0 iff

$$\delta F(u_0; h) = 0 \quad \text{for all } h \in X. \quad (2)$$

If the Gâteaux derivative $F'(u_0)$ exists, then condition (2) is equivalent to

$$F'(u_0) = 0. \quad (2^*)$$

Example 3. Let the C^1 -function $F: U(u_0) \subseteq \mathbb{R}^N \rightarrow \mathbb{R}$ be given on the open neighborhood $U(u_0)$ of the point u_0 , where $N = 1, 2, \dots$. Then,

$$\delta F(u_0; h) = \sum_{j=1}^N h_j \partial_j F(u_0) \quad \text{for all } h \in \mathbb{R}^N,$$

where $h = (h_1, \dots, h_N)$. Thus, u_0 is a critical point of F iff

$$\partial_j F(u_0) = 0 \quad \text{for all } j = 1, \dots, N.$$

Standard Example 4. Let X be a real Hilbert space. Set

$$F(u) := 2^{-1}(u | u) - (v | u) \quad \text{for all } u \in X$$

and fixed $v \in X$. Then, for all $u, h \in X$,

$$\delta F(u; h) = (u | h) - (v | h), \quad \delta^2 F(u; h) = (h | h),$$

and $\delta^n F(u; h) = 0$ if $n \geq 3$.

Proof. Set $\phi(t) := F(u + th)$ for all $t \in \mathbb{R}$ and fixed $u, h \in X$. Then

$$\phi(t) = 2^{-1}(u | u) + t(u | h) + 2^{-1}t^2(h | h) - (v | u) - t(v | h).$$

By definition, $\delta^k F(u; h) = \phi^{(k)}(0)$. □

2.2 Necessary and Sufficient Conditions for Local Extrema and the Classical Calculus of Variations

Theorem 2.A. Let the functional $F: U(u_0) \subseteq X \rightarrow \mathbb{R}$ be given on the open neighborhood $U(u_0)$ of the point u_0 in the real normed space X . Then the following are true:

(i) Necessary condition. If F has a local minimum or a local maximum at the point u_0 , then u_0 is a critical point of F , that is,

$$\delta F(u_0; h) = 0 \quad \text{for all } h \in X, \quad (3)$$

provided the first variation $\delta F(u_0; h)$ exists for each $h \in X$.

If the Gâteaux derivative $F'(u_0)$ exists, then condition (3) is equivalent to

$$F'(u_0) = 0 \quad (\text{Euler equation}).$$

(ii) Sufficient condition. The functional F has a local minimum at the point u_0 provided the following hold true:

(α) Condition (3) is satisfied.

(β) The second variation $\delta^2 F(u; h)$ exists for all u in an open neighborhood of u_0 and for all $h \in X$. There is a constant $c > 0$ such that

$$\delta^2 F(u_0; h) \geq c\|h\|^2 \quad \text{for all } h \in X.$$

(γ) For each given $\varepsilon > 0$, there is an $\eta(\varepsilon) > 0$ such that

$$|\delta^2 F(u; h) - \delta^2 F(u_0; h)| \leq \varepsilon\|h\|^2$$

for all $u, h \in X$ with $\|u - u_0\| < \eta(\varepsilon)$.

Proof. Ad (i). Set $\phi(t) := F(u_0 + th)$, where the real parameter t lives in a neighborhood of $t = 0$. The real function ϕ has a local minimum or local maximum at $t = 0$. Hence

$$\phi'(0) = 0.$$

This is condition (3).

Ad (ii). Since $\phi'(0) = 0$, the classical Taylor theorem yields

$$\begin{aligned} F(u_0 + h) - F(u_0) &= \phi(1) - \phi(0) = 2^{-1}\phi''(\theta) \\ &= 2^{-1}\delta^2 F(u_0 + \theta h; h) \quad \text{for all } h \in X, \end{aligned}$$

where $0 < \theta < 1$. Using $\delta^2 F(u_0 + \theta h; h) = \delta^2 F(u_0; h) + [\delta^2 F(u_0 + \theta h; h) - \delta^2 F(u_0; h)]$, we get

$$F(u_0 + h) - F(u_0) \geq \frac{1}{2} \left(c - \frac{c}{2} \right) \|h\|^2 \geq \frac{c}{4} \|h\|^2,$$

for all $h \in X$ with $\|h\| < \frac{\eta c}{2}$. □

The same argument yields the following, slightly more general, result.

Corollary 1. *Let the function $F: U(u_0) \subseteq X \rightarrow \mathbb{R}$ be given on the open neighborhood $U(u_0)$ of the point u_0 in the normed space X . Let Y be a linear subspace of X .*

Suppose that F has a local minimum at the point u_0 with respect to the plane $u_0 + Y$, that is, there is some $r > 0$ such that

$$F(u) \geq F(u_0) \quad \text{for all } u \in X \text{ with } u - u_0 \in Y \text{ and } \|u - u_0\| < r.$$

Then,

$$\delta F(u_0; h) = 0 \quad \text{for all } h \in Y,$$

provided the first variation $\delta F(u_0; h)$ exists for all $h \in Y$.

Standard Example 2. Let us study the following classical *variational problem*³:

$$F(u) := \int_G L(x, u(x), \partial_1 u(x), \dots, \partial_N u(x)) dx = \min!, \quad (4)$$

$$u = g \quad \text{on } \partial G, \quad u \in C^1(\bar{G}),$$

where G is a nonempty bounded open set in \mathbb{R}^N , $N \geq 1$. We are given the function $g \in C^1(\bar{G})$. Let the *Lagrangian* $L: \bar{G} \times \mathbb{R}^{N+1} \rightarrow \mathbb{R}$ be C^1 .

We set $X := C^1(\bar{G})$, where X is equipped with the maximum norm $\|u\| := \max_{x \in \bar{G}} |u(x)|$. Furthermore, set

$$Y := \{u \in X : u = 0 \text{ on } \partial G\}.$$

Suppose that $u_0 \in X$ is a local C^1 -minimal point of the original problem (4). That is, by definition, there is a number $r > 0$ such that

$$F(u) \geq F(u_0) \quad \text{for all } u \in X \text{ with } u - u_0 \in Y \text{ and } \|u - u_0\| < r.$$

Then, u_0 is a solution the following *Euler–Lagrange equation*:

$$\sum_{j=1}^N \partial_j L_{\partial_j u}(p(x)) = L_u(p(x)) \quad \text{on } G, \quad (5)$$

where $p(x) := (x, u_0(x), \partial_1 u_0(x), \dots, \partial_N u_0(x))$, and $L_{\partial_j u}$ and L_u denote the partial derivatives of L with respect to $\partial_j u$ and u , respectively.

³Spaces of smooth functions like $C^1(\bar{G})$ were introduced in Section 2.2.3 of AMS Vol. 108.

The following elegant proof dates back to Lagrange. This proof corresponds to the proof of Corollary 1.

Proof. We set

$$\phi(t) := F(u_0 + th) \quad \text{for fixed } h \in Y,$$

where the real parameter t lives in a neighborhood of $t = 0$. Then the real function ϕ has a local minimum at the point $t = 0$. Hence

$$\phi'(0) = 0.$$

Recall that $\delta F(u_0; h) = \phi'(0)$. Since

$$\phi(t) = \int_G L(x, u_0(x) + th(x), \partial_1 u_0(x) + t\partial_1 h(x), \dots, \partial_N u_0(x) + t\partial_N h(x)) dx,$$

we get

$$0 = \phi'(0) = \int_G \sum_{j=1}^N L_{\partial_j u}(p(x)) \partial_j h(x) + L_u(p(x)) h(x) dx.$$

In particular, choose $h \in C_0^\infty(G)$. Integration by parts⁴ yields

$$0 = \delta F(u_0; h) = \int_G \left[\sum_{j=1}^N -\partial_j L_{\partial_j u}(p(x)) + L_u(p(x)) \right] h(x) dx$$

for all $h \in C_0^\infty(G)$.

By the variational lemma from Section 2.2.3 in AMS Vol. 108, this implies (4). \square

Remark 3 (Critical points). Instead of the minimum problem (4), let us consider the following more general problem:

$$\begin{aligned} F(u) &:= \int_G L(x, u(x), \partial_1 u(x), \dots, \partial_N u(x)) dx = \text{stationary!}, \\ u &= g \quad \text{on } \partial G. \end{aligned} \tag{4*}$$

Here we are looking for a critical point u_0 of F . By definition, u_0 is a solution of (4*) iff $u_0 \in C^1(\bar{G})$ and the first variation vanishes, that is,

$$\delta F(u_0; h) = 0 \quad \text{for all } h \in Y.$$

The proof of Standard Example 2 immediately shows the following:

⁴See Section 2.2.5 in AMS Vol. 108.

If u_0 is a solution of (4*), then u_0 is a solution of the Euler–Lagrange equation (5).

Most variational problems in physics are not minimum problems, but they are of the type that (4*) is (the *principle of stationary action*).

Remark 4 (Systems of Euler–Lagrange equations). Consider problem (4*), where $u = (u_1, \dots, u_M)$. Applying the proof of Standard Example 2 to each fixed component u_j of u , we obtain the following:

Let $u_0 = (u_{10}, \dots, u_{M0})$ be a solution of (4*). Then, u_0 is a solution of the Euler–Lagrange system

$$\sum_{j=1}^N \partial_j L_{\partial_j u_m}(p(x)) = L_{u_m}(p(x)) \quad \text{on } G, \quad m = 1, \dots, M, \quad (5^*)$$

where $p(x) := (x, u_0(x), \partial_1 u_0(x), \dots, \partial_M u_0(x))$.

2.3 The Lack of Compactness in Infinite-Dimensional Banach Spaces

Theorem 2.B. A Banach space X is finite-dimensional iff the closed unit ball is compact.

Proof. Let $B := \{u \in X : \|u\| \leq 1\}$. If $\dim X$ is finite, then B is compact, by Corollary 8 in Section 1.12 of AMS Vol. 108.

Conversely, let $\dim X = \infty$. We have to show that B is not compact. Suppose first that X is a separable Hilbert space. Then there exists a countable orthonormal system (u_n) in X . By the Pythagorean theorem,

$$\|u_n - u_m\|^2 = \|u_n\|^2 + \|u_m\|^2 = 2 \quad \text{for all } n \neq m.$$

Thus, the sequence (u_n) in B has no convergent subsequence, and hence B is not compact.

Suppose now that X is a Banach space with $\dim X = \infty$.

Step 1: Almost orthogonal elements. Let W be a closed linear subspace of X with $W \neq X$. Then, for each $\varepsilon \in]0, 1[$, there exists a point $u_\varepsilon \in X$ such that

$$\|u_\varepsilon\| = 1 \quad \text{and} \quad \text{dist}(u_\varepsilon, W) \geq 1 - \varepsilon. \quad (6)$$

Recall that $\text{dist}(v, W) = \inf_{w \in W} \|v - w\|$.

To prove (6), let $v \in X - W$. Then,

$$\text{dist}(v, W) > 0.$$

Otherwise, there would exist a sequence (w_n) in W such that $\|v - w_n\| \rightarrow 0$ as $n \rightarrow \infty$, and hence $v \in W$, since W is closed. But this contradicts $v \notin W$.

We choose a point $w_\varepsilon \in W$ with $0 < \|v - w_\varepsilon\| \leq (1 - \varepsilon)^{-1} \operatorname{dist}(v, W)$. Set $u_\varepsilon := \frac{(v - w_\varepsilon)}{\|v - w_\varepsilon\|}$. Then, u_ε is the desired element. In fact,

$$\begin{aligned}\|u_\varepsilon - w\| &= \|v - w_\varepsilon\|^{-1} \|v - w_\varepsilon - \|v - w_\varepsilon\| \cdot w\| \\ &\geq \|v - w_\varepsilon\|^{-1} \operatorname{dist}(v, W) \geq 1 - \varepsilon \quad \text{for all } w \in W.\end{aligned}$$

Step 2: We want to show that B is not compact. To this end, we choose a point $w_1 \in X$ with $\|w_1\| = 1$. Let $W := \operatorname{span}\{w_1\}$. By Step 1, there exists a point $w_2 \in X$ with $\|w_2\| = 1$ and $\|w_2 - w_1\| \geq 2^{-1}$. Continuing this construction, we get a sequence (w_n) with $\|w_n\| = 1$ for all n and

$$\|w_n - w_m\| \geq 2^{-1} \quad \text{for all } n \neq m.$$

Thus, the sequence (w_n) in B has no convergent subsequence, i.e., B is not compact. \square

2.4 Weak Convergence

Recall that we introduced the following notation in Section 1.21 of AMS Vol. 108:

$$\langle f, u \rangle := f(u) \quad \text{for all } f \in X^*, u \in X.$$

Definition 1. Let (u_n) be a sequence in the normed space X over \mathbb{K} . We write

$$u_n \rightharpoonup u \quad \text{as } n \rightarrow \infty \tag{7}$$

iff

$$\langle f, u_n \rangle \rightarrow \langle f, u \rangle \quad \text{as } n \rightarrow \infty \quad \text{for all } f \in X^*. \tag{7^*}$$

We say that the sequence (u_n) converges weakly to u in X as $n \rightarrow \infty$.

The weak limit u is uniquely determined. In fact, if $u_n \rightharpoonup u$ and $u_n \rightharpoonup v$ as $n \rightarrow \infty$, then $f(u - v) = 0$ for all $f \in X^*$, and hence $u = v$, by Corollary 3 in Section 1.1.

The norm convergence $u_n \rightarrow u$ as $n \rightarrow \infty$ (i.e., $\|u_n - u\| \rightarrow 0$ as $n \rightarrow \infty$) is also called the *strong convergence*.

Standard Example 2. Let X be a Hilbert space over \mathbb{K} . Then

(i) $u_n \rightharpoonup u$ in X as $n \rightarrow \infty$ iff

$$(v | u_n) \rightarrow (v | u) \quad \text{as } n \rightarrow \infty \quad \text{for all } v \in X.$$

- (ii) Suppose that $\dim X = \infty$, and let (u_n) be a countable *orthonormal system* in X . Then, (u_n) has *no* convergent subsequence, but

$$u_n \rightharpoonup 0 \quad \text{as } n \rightarrow \infty.$$

In particular, observe that $\|u_n\| = 1$ for all n , but the weak limit of (u_n) does not belong to the boundary of the unit ball. We will show in Corollary 4 ahead that the weak limit of (u_n) always belongs to the *closed convex hull* of the set $\{u_1, u_2, \dots\}$.

Proof. Ad (i). This follows from the *Riesz theorem* in Section 2.10 of AMS Vol. 108.

Ad (ii). By the proof of Theorem 2.B, the sequence (u_n) has no convergent subsequence. Moreover, for each $v \in X$, the *Bessel inequality* from Section 3.1 of AMS Vol. 108 yields

$$\sum_{n=1}^{\infty} |(v | u_n)|^2 \leq \|v\|^2,$$

and hence $(v | u_n) \rightarrow 0$ as $n \rightarrow \infty$ for all $v \in X$. \square

Proposition 3. *Let X be a normed space over \mathbb{K} . Then*

- (i) $u_n \rightarrow u$ in X as $n \rightarrow \infty$ implies $u_n \rightharpoonup u$ in X as $n \rightarrow \infty$.
- (ii) The converse is true if X is finite-dimensional.

Proof. Ad (i). This follows from the continuity of the functional f in (7*).

Ad (ii). If $X = \{0\}$, then the statement is trivial. Let $\dim X = n$, where $n = 1, 2, \dots$. Choose a basis $\{e_1, \dots, e_n\}$ of X . Then, each functional $f \in X^*$ has the form

$$f(u) = \sum_{k=1}^n \alpha_k \beta_k, \quad \text{where } u = \sum_{k=1}^n \alpha_k e_k,$$

and $\alpha_k, \beta_k \in \mathbb{K}$ for all k . Letting $\beta_k = 1$ for fixed k and $\beta_m = 0$ for all $m \neq k$, we obtain that $u_n \rightharpoonup u$ as $n \rightarrow \infty$ is equivalent to the convergence of the corresponding components. In turn, this is equivalent to $u_n \rightarrow u$ as $n \rightarrow \infty$. \square

Theorem 2.C. *Each bounded sequence (u_n) in a Hilbert space X over \mathbb{K} has a weakly convergent subsequence.*

Corollary 4. *The limit point of each weakly convergent subsequence of (u_n) belongs to the closed convex hull of the set $\{u_1, u_2, \dots\}$.*

In the following proof we will critically use the *Riesz theorem* from Section 2.10 of AMS Vol. 108.

Proof of Theorem 2.C. For $X = \{0\}$, the statement is trivial. Let $X \neq \{0\}$.

Step 1: Suppose first that X is *separable*. We choose a countable set $\{v_k\}$, which is dense in X , and we use the following *diagonal procedure*:

$$\begin{aligned} (u_{11} | v_1), (u_{12} | v_1), (u_{13} | v_1), \dots &\rightarrow a_1, \\ (u_{21} | v_2), (u_{22} | v_2), (u_{23} | v_2), \dots &\rightarrow a_2, \\ \dots \end{aligned}$$

To be precise, since $|(u_n | v_1)| \leq \|u_n\| \|v_1\|$ for all n , the sequence of the numbers $(u_n | v_1)$ is bounded in \mathbb{K} . Thus, there exists a subsequence of (u_n) , denoted by (u_{1n}) , such that $(u_{1n} | v_1) \rightarrow a_1$ as $n \rightarrow \infty$. Furthermore, there exists a subsequence (u_{2n}) of (u_{1n}) such that $(u_{2n} | v_2) \rightarrow a_2$ as $n \rightarrow \infty$, and so on.

The diagonal sequence (w_n) defined by $w_n := u_{nn}$ has the crucial property that

$$(w_n | v_k) \rightarrow a_k \quad \text{as } n \rightarrow \infty \quad \text{for all } k.$$

Moreover, there exist numbers $a(v)$ such that

$$(w_n | v) \rightarrow a(v) \quad \text{as } n \rightarrow \infty \quad \text{for each } v \in X. \quad (8)$$

This follows from

$$\begin{aligned} |(w_n - w_m | v)| &= |(w_n - w_m | v - v_k) + (w_n - w_m | v_k)| \\ &\leq \|w_n - w_m\| \|v - v_k\| + |(w_n - w_m | v_k)| < \varepsilon \end{aligned}$$

for suitable v_k and all $n, m \geq n_0(\varepsilon)$. Note that the sequence (w_n) is bounded and the set $\{v_k\}$ is dense in X .

Obviously, the map $v \mapsto a(v)$ is linear and, by (8),

$$|a(v)| \leq \|v\| \sup_n \|w_n\| \quad \text{for all } v \in X.$$

According to the *Riesz theorem* from Section 2.10 of AMS Vol. 108, there exists a $w \in X$ such that

$$a(v) = (w | v) \quad \text{for all } v \in X.$$

By (8), $(v | w_n) \rightarrow (v | w)$ as $n \rightarrow \infty$ for all $v \in X$. Hence

$$w_n \rightarrow w \quad \text{as } n \rightarrow \infty.$$

Step 2: If the Hilbert space X is *not* separable, then let Y be the closure of $\text{span}\{u_1, u_2, \dots\}$. Y is thus separable. In fact, for each $\alpha \in \mathbb{K}$ and each $\varepsilon > 0$, there are rational numbers β and γ such that

$$|\alpha - (\beta + \gamma i)| < \varepsilon.$$

Consequently, the countable set of all the finite linear combinations

$$(\beta_1 + \gamma_1 i)u_1 + \cdots + (\beta_n + \gamma_n i), \quad n = 1, 2, \dots,$$

with rational coefficients β_j, γ_j is dense in Y .

Applying Step 1 to the space Y , there exists a subsequence (w_n) of (u_n) such that

$$(v | w_n) \rightarrow (v | w) \quad \text{as } n \rightarrow \infty \quad \text{for all } v \in Y \text{ and fixed } w \in Y. \quad (9)$$

Let $z \in X$. According to Section 2.9 in AMS Vol. 108, we get the decomposition $z = v + v^\perp$, where $v \in Y$ and $v^\perp \in Y^\perp$. Since $(v^\perp | y) = 0$ for all $y \in Y$, it follows from (9) that

$$(z | w_n) \rightarrow (z | w) \quad \text{as } n \rightarrow \infty \quad \text{for all } z \in X.$$

Hence $w_n \rightharpoonup w$ as $n \rightarrow \infty$. \square

Proof of Corollary 4. Let (w_n) be a bounded sequence in the Hilbert space X such that $w_n \rightharpoonup w$ as $n \rightarrow \infty$, that is,

$$(v | w_n) \rightarrow (v | w) \quad \text{as } n \rightarrow \infty \quad \text{for all } v \in X. \quad (10)$$

It is sufficient to prove that there are indices $n_1 < n_2 < \dots$ such that

$$k^{-1}(w_{n_1} + \cdots + w_{n_k}) \rightarrow w \quad \text{as } n \rightarrow \infty.$$

Replacing w_n by $w_n - w$, we can assume that $w = 0$.

First let $n_1 := 1$. By (10),

$$|(w_{n_1} | w_n)| \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Thus, there exists an index $n_2 > n_1$ such that $|(w_{n_1} | w_{n_2})| \leq 2^{-1}$. Continuing this procedure, we get indices such that

$$|(w_{n_1} | w_{n_k})| \leq (k-1)^{-1}, \dots, |(w_{n_{k-1}} | w_{n_k})| \leq (k-1)^{-1},$$

for all $k = 3, 4, \dots$. Since $\|w_n\| \leq \text{const} = C$ for all n , we obtain

$$\begin{aligned} \|k^{-1}(w_{n_1} + \cdots + w_{n_k})\|^2 &\leq k^{-2} \left\{ \sum_{j=1}^k \|w_{n_j}\|^2 + 2 \sum_{j=1}^{k-1} |(w_{n_j} | w_{n_k})| \right. \\ &\quad \left. + 2 \sum_{j=1}^{k-2} |(w_{n_j} | w_{n_{k-1}})| + \dots \right\} \\ &\leq k^{-2}(kC^2 + 2(k-1)(k-1)^{-1} \\ &\quad + 2(k-2)(k-2)^{-1} + \dots + 2) \\ &\leq k^{-1}(C^2 + 2) \rightarrow 0 \quad \text{as } k \rightarrow \infty. \end{aligned} \quad \square$$

2.5 The Generalized Weierstrass Existence Theorem

Definition 1. Let $F: M \subseteq X \rightarrow \mathbb{R}$ be a functional on the subset M of the real normed space X . Then

- (i) F is called *weakly sequentially continuous* iff, for each $u \in M$ and each sequence (u_n) in M ,

$$u_n \rightharpoonup u \quad \text{implies} \quad F(u_n) \rightarrow F(u) \quad \text{as } n \rightarrow \infty.$$

- (ii) F is called *weakly sequentially lower semicontinuous*⁵ iff

$$F(u) \leq \liminf_{n \rightarrow \infty} F(u_n) \tag{11}$$

for each $u \in M$ and each sequence (u_n) in M with $u_n \rightharpoonup u$ as $n \rightarrow \infty$.

- (iii) F is called *coercive* iff $\frac{F(u)}{\|u\|} \rightarrow +\infty$ as $\|u\| \rightarrow \infty$ on M .

- (iv) F is called *weakly coercive* iff $F(u) \rightarrow +\infty$ as $\|u\| \rightarrow \infty$ on M .

- (v) F is called *strictly convex* iff the set M is convex and

$$F(\alpha u + (1 - \alpha)v) < \alpha F(u) + (1 - \alpha)F(v),$$

for all $\alpha \in]0, 1[$ and all $u, v \in M$ with $u \neq v$.

Recall that $F: M \rightarrow \mathbb{R}$ is convex iff “ $<$ ” in (v) is replaced with “ \leq .”

In the case where F is a real function, the strict convexity of F means that the graph of F lies properly under the chord. The functions F pictured in Figures 2.2(a) and 2.2(b) are convex and strictly convex, respectively.

Intuitively, strict convexity of F ensures the *uniqueness* of the minimal point u_0 (Figure 2.2(b)). This will be proved rigorously in Corollary 2 just ahead.

Moreover, it follows intuitively from Figure 2.2 that each local minimum of a convex function is also a *global minimum*. This will be proved in Section 2.9.

Theorem 2.D. Suppose that the functional $F: M \rightarrow \mathbb{R}$ has the following three properties:

- (i) M is a nonempty closed convex subset of the real Hilbert space X .

⁵The definition and properties of the classical symbols “ \lim ” and “ $\overline{\lim}$ ” will be recalled in Problem 2.7b.

The intuitive meaning of (11) will be discussed in Problem 2.7b.

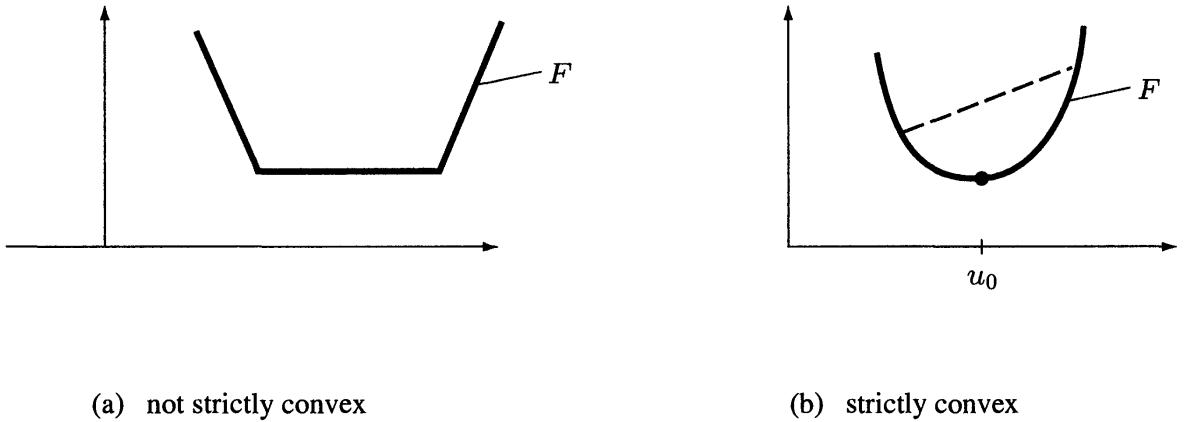


FIGURE 2.2.

- (ii) F is weakly sequentially lower semicontinuous.
 - (iii) If the set M is unbounded, then F is weakly coercive.

Then the minimum problem

$$F(u) = \min !, \quad u \in M, \quad (12)$$

has a solution.

Corollary 2. If, in addition, F is strictly convex, then problem (12) has a unique solution u .

Proof of Theorem 2.D. *Step 1:* Suppose that M is bounded. Set

$$\gamma := \inf_{u \in M} F(u).$$

Hence $-\infty \leq \gamma < \infty$. Then there exists a sequence (u_n) in M such that

$$F(u_n) \rightarrow \gamma \quad \text{as } n \rightarrow \infty. \quad (13)$$

Since M is bounded, the sequence (u_n) is bounded. By Theorem 2.C, there exists a convergent subsequence, again denoted by (u_n) , such that $u_n \rightarrow u$ as $n \rightarrow \infty$. Corollary 4 in Section 2.4 tells us that u lies in the closed convex hull of $\{u_1, u_2, \dots\}$. Hence $u \in M$.

Since F is weakly sequentially lower semicontinuous,

$$F(u) \leq \liminf_{n \rightarrow \infty} F(u_n) = \gamma.$$

This implies $F(u) = \gamma$.

Step 2: Suppose that the set M is unbounded. Fix $v \in M$. Since $F(u) \rightarrow +\infty$ as $\|u\| \rightarrow \infty$, there exists an $r > 0$ such that

$$F(u) > F(v) \quad \text{for all } u \in M \text{ with } \|u\| > r, \quad (14)$$

and the set $M_r := \{u \in M : \|u\| \leq r\}$ is not empty. By (14), each solution u of the modified problem

$$F(u) = \min !, \quad u \in M_r, \quad (15)$$

is also a solution of the original problem (12).

Since the set M_r is closed, convex, and bounded, problem (15) has a solution, by Step 1. \square

Proof of Corollary 2. Suppose that problem (12) has two different solutions, u and v . Then, $\frac{1}{2}(u + v) \in M$, and hence

$$F\left(\frac{1}{2}(u + v)\right) < \frac{1}{2}F(u) + \frac{1}{2}F(v) = F(u).$$

This contradicts the fact that $F(u)$ is the minimal value of F on M . \square

Definition 3. Let $F: M \rightarrow \mathbb{R}$ be a functional on the subset M of the real normed space X . For each $r \in \mathbb{R}$, set

$$\mathcal{M}_r := \{u \in M : F(u) \leq r\}.$$

- (a) F is called *lower semicontinuous* on the closed set M iff the set \mathcal{M}_r is closed for all $r \in \mathbb{R}$.
- (b) F is called *quasi-convex* on the convex set M iff the set \mathcal{M}_r is convex for all $r \in \mathbb{R}$.

The following hold:

$$\begin{aligned} F \text{ is convex} &\Rightarrow F \text{ is quasi-convex;} \\ F \text{ is continuous} &\Rightarrow F \text{ is lower semicontinuous.} \end{aligned}$$

In fact, let F be convex on the convex set M . If $u, v \in \mathcal{M}_r$, then

$$F(\alpha u + (1 - \alpha)v) \leq \alpha F(u) + (1 - \alpha)F(v) \leq \alpha r + (1 - \alpha)r \leq r \text{ for all } \alpha \in [0, 1],$$

and hence $\alpha u + (1 - \alpha)v \in \mathcal{M}_r$. Furthermore, if F is continuous on the closed set M , then it follows from $u_n \in \mathcal{M}_r$ for all n and $u_n \rightarrow u$ as $n \rightarrow \infty$ that $F(u_n) \leq r$, and hence $F(u) \leq r$ (i.e., $u \in \mathcal{M}_r$).

Proposition 4. Suppose that the functional $F: M \subseteq X \rightarrow \mathbb{R}$ has the following three properties:

- (i) M is a nonempty, closed, convex subset of the real Hilbert space X .
- (ii) F is quasi-convex and lower semicontinuous.

(iii) If M is unbounded, then F is weakly coercive.

Then, the minimum problem $F(u) = \min !, u \in M$, has a solution. This solution is unique provided that F is strictly convex.

This follows from Theorem 2.D and Corollary 2 along with the following result.

Lemma 5. Let the functional $F: M \subseteq X \rightarrow \mathbb{R}$ be lower semicontinuous and quasi-convex on the nonempty, closed, convex set M of the real normed space X .

Then, F is weakly sequentially lower semicontinuous on M .

Proof. If the assertion is not true, then there exist a point $u \in M$ and a sequence (u_n) in M such that $u_n \rightharpoonup u$ as $n \rightarrow \infty$ and

$$F(u) > \lim_{n \rightarrow \infty} F(u_n).$$

Consequently, there is a real number r so that $r < F(u)$ and $u_n \in \mathcal{M}_r$ for all $n \geq n_0(\varepsilon)$. Since \mathcal{M}_r is closed and convex, it follows from $u_n \rightharpoonup u$ as $n \rightarrow \infty$ that $u \in \mathcal{M}_r$, and hence $F(u) \leq r$. This is a contradiction. \square

Remark 6 (Generalization to reflexive Banach spaces). All the results of this section remain valid if X is not a real Hilbert space but rather a real **reflexive** Banach space.

The proofs given above remain unchanged. However, instead of the properties of weak convergence in Hilbert spaces (Theorem 2.C and Corollary 4 in Section 2.4), we have to use the corresponding properties in reflexive Banach spaces that will be proved in Section 2.8.

2.6 Applications to the Calculus of Variations

Instead of the classical variational problem

$$F(u) := \int_G L(x, u(x), \partial_1 u(x), \dots, \partial_N u(x)) dx = \min !, \quad u \in C^1(\bar{G}),$$

$u = g \text{ on } \partial G \quad (\text{boundary condition}),$

let us consider the following *generalized problem* on a Sobolev space⁶:

$$\begin{aligned} F(u) &= \min !, & u &\in W_2^1(G), \\ u - g &\in \overset{\circ}{W}_2^1(G) & (\text{generalized boundary condition}). \end{aligned} \tag{16}$$

We assume that

⁶The Sobolev spaces $W_2^1(G)$ and $\overset{\circ}{W}_2^1(G)$ were introduced in Section 2.5 of AMS Vol. 108.

- (H1) G is a nonempty, bounded, open set in \mathbb{R}^N , $N \geq 1$.
- (H2) The function $L: \bar{G} \times \mathbb{R}^{N+1} \rightarrow \mathbb{R}$ is continuous.
- (H3) (Convexity) For each $x \in \bar{G}$, the function $L(x, \dots)$ is convex on \mathbb{R}^N with respect to the variables $u, \partial_1 u, \dots, \partial_N u$.
- (H4) (Growth condition) For all $(x, u, \partial_1 u, \dots, \partial_N u) \in \bar{G} \times \mathbb{R}^{N+1}$,

$$|L(x, u, \partial_1 u, \dots, \partial_N u)| \leq \text{const} \left(|u|^2 + \sum_{j=1}^N |\partial_j u|^2 \right).$$

- (H5) (Coerciveness condition) For all $(x, u, \partial_1 u, \dots, \partial_N u) \in \bar{G} \times \mathbb{R}^{N+1}$,

$$c \sum_{j=1}^N |\partial_j u|^2 - d \leq L(x, u, \partial_1 u, \dots, \partial_N u),$$

where $c > 0$ and $d \geq 0$ are constants.

Proposition 1. *For each given function $g \in W_2^1(G)$, the variational problem (16) has a solution.*

The proof will be based on the following result.

Lemma 2. *Let $u_n \rightarrow u$ in $L_2(G)$ as $n \rightarrow \infty$, where G is a nonempty open set in \mathbb{R}^N , $N \geq 1$. Then, there exist a subsequence $(u_{n'})$ and a function $w \in L_2(G)$ such that*

$$u_{n'}(x) \rightarrow u(x) \quad \text{for almost all } x \in G,$$

and $|u_{n'}(x)| \leq w(x)$ for all n' and all $x \in G$.

Proof. First let G be an open interval $]a, b[$ in \mathbb{R} . Then, the result follows from Step 4 of the proof of Standard Example 4 in Section 2.2.1 of AMS Vol. 108 with $w = |v_1| + s$. For general open sets in \mathbb{R}^N , the proof proceeds completely analogously. \square

Proof of Proposition 1. We want to use Proposition 4 from Section 2.5. To simplify notation, we consider the case where $N = 1$. The general case proceeds completely analogously.

Set $Y := W_2^1(G)$. Recall that

$$\|u\| = \left(\int_G (|u(x)|^2 + |\partial u(x)|^2) dx \right)^{\frac{1}{2}}.$$

Step 1: We show that the functional $F: Y \rightarrow \mathbb{R}$ is convex, weakly coercive, and continuous.

It follows from the growth condition (H4) that

$$F(u) := \int_G L(x, u(x), \partial u(x)) dx \leq \text{const} \|u\|_Y^2.$$

Thus, the functional $F: Y \rightarrow \mathbb{R}$ is well defined. By (H3), F is convex. The coerciveness condition (H5) yields

$$c \int_G |\partial u(x)|^2 dx - d \int_G dx \leq \int_G L(x, u(x), \partial u(x)) dx.$$

By the *Poincaré–Friedrichs inequality* from Section 2.5.6 of AMS Vol. 108, there is a constant $C > 0$ such that

$$C\|u\|^2 \leq \int_G |\partial u(x)|^2 dx \quad \text{for all } u \in Y.$$

Thus, $\|u\| \rightarrow \infty$ on Y implies $F(u) \rightarrow +\infty$. Hence F is weakly coercive on Y .

To prove that $F: Y \rightarrow \mathbb{R}$ is continuous, let

$$u_n \rightarrow u \quad \text{in } Y \quad \text{as } n \rightarrow \infty.$$

Hence

$$u_n \rightarrow u \quad \text{and} \quad \partial u_n \rightarrow \partial u \quad \text{in } L_2(G) \quad \text{as } n \rightarrow \infty.$$

By Lemma 2, there are a subsequence $(u_{n'})$ and functions $v, w \in L_2(G)$ such that

$$u_{n'}(x) \rightarrow u(x) \quad \text{and} \quad \partial u_{n'}(x) \rightarrow \partial u(x) \quad \text{as } n' \rightarrow \infty,$$

for almost all $x \in G$, and

$$|u_{n'}(x)| \leq v(x) \quad \text{and} \quad |\partial u_{n'}(x)| \leq w(x),$$

for all n' and all $x \in G$. The growth condition (H4) tells us that

$$|L(x, u_{n'}(x), \partial u_{n'}(x))| \leq \text{const}(|v(x)|^2 + |w(x)|^2),$$

for all n' and all $x \in G$. By the continuity of L ,

$$L(x, u_{n'}(x), \partial u_{n'}(x)) \rightarrow L(x, u(x), \partial u(x)) \quad \text{as } n' \rightarrow \infty,$$

for almost all $x \in G$. Thus, the *dominated convergence theorem* (cf. the appendix of AMS Vol. 108) yields

$$\int_G L(x, u_{n'}(x), \partial u_{n'}(x)) dx \rightarrow \int_G L(x, u(x), \partial u(x)) dx \quad \text{as } n' \rightarrow \infty.$$

That is,

$$F(u_{n'}) \rightarrow F(u) \quad \text{as } n' \rightarrow \infty.$$

The same argument tells us that each convergent subsequence of $(F(u_n))$ has the limit $F(u)$. Hence the total sequence is convergent. That is,

$$F(u_n) \rightarrow F(u) \quad \text{as } n \rightarrow \infty.$$

Step 2: Let $X := \overset{\circ}{W}_2^1(G)$, and let $H(v) := F(g + v)$ for all $v \in X$. By Step 1, the functional $H: X \rightarrow \mathbb{R}$ is convex, weakly coercive, and continuous. Proposition 3 in Section 2.5 with $M = X$ tells us that the minimum problem

$$H(v) = \min !, \quad v \in X, \tag{16*}$$

has a solution. Problem (16*) is equivalent to the original problem (16). \square

2.7 Applications to Nonlinear Eigenvalue Problems

Let us consider the following eigenvalue problem:

$$Au = \lambda u, \quad u \in X, \quad \lambda \in \mathbb{R}, \quad \|u\| = r, \tag{17}$$

for the nonlinear operator $A: X \rightarrow X$. We assume that

- (H1) The functional $F: X \rightarrow \mathbb{R}$ is *weakly sequentially continuous* on the real Hilbert space X , where $X \neq \{0\}$.
- (H2) The operator $A: X \rightarrow X$ corresponds to the Fréchet derivative of the functional F ; in other words, for each given $u \in X$, we have

$$F(u + h) - F(u) = (Au | h) + \|h\|\varepsilon(h; u) \quad \text{for all } h \in X, \tag{18}$$

where $\varepsilon(h; u) \rightarrow 0$ as $h \rightarrow 0$.

- (H3) $Au = 0$ implies $u = 0$, and $F(0) = 0$.
- (H4) There is a point $v \in X$ with $F(v) > 0$. We choose $r > 0$ in such a way that $\|v\| \leq r$.

Proposition 1. *The eigenvalue problem (17) has a solution.*

We will show that each solution w of the maximum problem

$$F(u) = \max !, \quad \|u\| = r, \tag{19}$$

is a solution of the original eigenvalue problem (17).

Proof. *Step 1:* Let $B := \{u \in X: \|u\| \leq r\}$. We first show that the *modified maximum problem*

$$F(u) = \max !, \quad u \in B, \quad (19^*)$$

has a solution w . However, this follows from Theorem 2.D applied to $-F$.

Step 2: We show that $\|w\| = r$. In fact, if $\|w\| < r$, then w is an inner point of B , and hence Theorem 2.A yields

$$\delta F(w; h) = 0 \quad \text{for all } h \in X.$$

By (18), $\delta F(w; h) = (Aw \mid h)$. Thus, $Aw = 0$, and hence $w = 0$, by (H3). This implies

$$F(0) = 0.$$

Hence $F(u) \leq 0$ for all $u \in B$. By (H4), there is some $v \in B$ with $F(v) > 0$. This is a contradiction.

Summarizing, we obtain w as a solution to the maximum problem (19).

Step 3: Finally, we prove that w is a solution to the original eigenvalue problem (17). Let $Y := \text{span}\{w\}$. Set

$$\psi(t) := F((\cos t)w + (\sin t)k) \quad \text{for all } t \in \mathbb{R} \text{ and fixed } k \in Y^\perp.$$

If we let $u := w$ and $h := (\cos t - 1)w + (\sin t)k$, it follows from (18) that

$$\psi(t) - \psi(0) = t(Aw \mid k) + t\eta(t), \quad \text{where } \eta(t) \rightarrow 0 \text{ as } t \rightarrow 0.$$

Hence

$$\psi'(0) = (Aw \mid k).$$

Observe that, for all $k \in Y^\perp$ with $\|k\| = r$,

$$\begin{aligned} \|(\cos t)w + (\sin t)k\|^2 &= \cos^2 t\|w\|^2 + \sin^2 t\|k\|^2 \\ &= (\cos^2 t + \sin^2 t)r^2 = r^2. \end{aligned}$$

Since w is a solution of (19), the real function ψ has a *local maximum* at the point $t = 0$. Hence $\psi'(0) = 0$, that is,

$$(Aw \mid k) = 0 \quad \text{for all } k \in Y^\perp.$$

Recall that $Y := \{\lambda w: \lambda \in \mathbb{R}\}$. Using the orthogonal decomposition

$$Aw = y + k, \quad y \in Y, \quad k \in Y^\perp,$$

we get $y = \lambda w$ and $k = 0$ because $0 = (Aw \mid k) = (k \mid k)$.

Therefore, $Aw = \lambda w$, that is, w is a solution of (17). \square

2.8 Reflexive Banach Spaces

Let X be a normed space over \mathbb{K} . Recall from Section 1.21 of AMS Vol. 108 that, by definition, the *dual space* X^* consists of all linear continuous functionals $f: X \rightarrow \mathbb{K}$. We set

$$X^{**} := (X^*)^*,$$

that is, the *bidual* space X^{**} consists of all linear continuous functionals $F: X^* \rightarrow \mathbb{K}$. Recall also that we have introduced the following notation:

$$\langle f, u \rangle := f(u) \quad \text{for all } f \in X^*, u \in X.$$

The following definition is crucial for the general theory of variational problems in terms of functional analysis.

Definition 1. The normed space X over \mathbb{K} is called *reflexive* iff each $F \in X^{**}$ allows the following representation:

$$F(f) = \langle f, u \rangle \quad \text{for all } f \in X^* \text{ and some fixed } u \in X.$$

Standard Example 2. Each Hilbert space X over \mathbb{K} is reflexive.

We will show that this is a consequence of the *Riesz theorem* from Section 2.10 in AMS Vol. 108.

Proof. By Section 2.11 in AMS Vol. 108, it follows from the Riesz theorem that there exists a bijective map $J: X \rightarrow X^*$ such that

$$\langle Ju, v \rangle = (u | v) \quad \text{for all } u, v \in X,$$

and J is *antilinear*, meaning that

$$J(\alpha u + \beta w) = \bar{\alpha}Ju + \bar{\beta}Jw \quad \text{for all } u, w \in X, \alpha, \beta \in \mathbb{K},$$

where the bar denotes the conjugate complex number. Moreover, we have $\|Ju\| = \|u\|$ for all $u \in X$. Let $F \in X^{**}$. We set

$$G(u) := \overline{F(Ju)} \quad \text{for all } u \in X.$$

Then, $G: X \rightarrow \mathbb{K}$ is linear. For all $u \in X$,

$$|G(u)| \leq \|F\| \|Ju\| = \|F\| \|u\|.$$

Hence $G \in X^*$. By the Riesz theorem in Section 2.10 of AMS Vol. 108, there exists a $v \in X$ such that

$$G(u) = (v | u) \quad \text{for all } u \in X.$$

This implies

$$F(w) = \overline{G(J^{-1}w)} = (J^{-1}w \mid v) = \langle w, v \rangle \quad \text{for all } w \in X^*. \quad \square$$

Proposition 3. *Let X be a normed space over \mathbb{K} . Define the map $j: X \rightarrow X^{**}$ through*

$$j(u)(f) := \langle f, u \rangle \quad \text{for all } u \in X, f \in X^*.$$

Then the following are true:

- (i) *The map j is linear and*

$$\|j(u)\| = \|u\| \quad \text{for all } u \in X.$$

- (ii) *The space X is reflexive iff $j: X \rightarrow X^{**}$ is bijective.*

Proof. Ad (i). Obviously, $j(\alpha u + \beta v) = \alpha j(u) + \beta j(v)$ for all $u, v \in X$ and $\alpha, \beta \in \mathbb{K}$. Moreover,

$$\|j(u)\| = \sup_{f \in X^*, \|f\| \leq 1} |\langle f, u \rangle| = \|u\|,$$

by Corollary 2 in Section 1.1.

Ad (ii). It follows from (i) that $j: X \rightarrow X^{**}$ is injective, since $j(u) = 0$ implies $u = 0$. By Definition 1, X is reflexive iff j is surjective. \square

Proposition 4. *Every closed linear subspace Y of a reflexive Banach space X over \mathbb{K} is again a reflexive Banach space.*

Proof. Obviously, Y is a Banach space. The following simple arguments will be based on restrictions and extensions of functionals. In particular, we will use the Hahn–Banach theorem and the separation of convex sets.

For each given $x^* \in X^*$, let $x_r^*: Y \rightarrow \mathbb{K}$ denote the restriction of $x^*: X \rightarrow \mathbb{K}$ to the subspace Y . Clearly, $x_r^* \in Y^*$. In this sense, $Y \subseteq X$ implies

$$X^* \subseteq Y^*. \tag{20}$$

Hence we obtain $(Y^*)^* \subseteq (X^*)^*$, that is,

$$Y^{**} \subseteq X^{**}. \tag{21}$$

Let $y^{**} \in Y^{**}$. We have to show that there exists a $y \in Y$ such that

$$y^{**}(y^*) = y^*(y) \quad \text{for all } y^* \in Y^*. \tag{22}$$

In fact, it follows from (20) and (21) that

$$y^{**}(x_r^*) = y^{**}(x^*) \quad \text{for all } x^* \in X^*. \quad (23)$$

Since X is *reflexive*, there exists a $y \in X$ such that

$$y^{**}(x^*) = x^*(y) \quad \text{for all } x^* \in X^*. \quad (24)$$

We claim that $y \in Y$. Otherwise, we would have $\text{dist}(y, Y) > 0$, since Y is closed in X . By Proposition 3 in Section 1.2 (separation of convex sets), there exists an $x^* \in X^*$ such that $x_r^* = 0$ and $x^*(y) \neq 0$. By (23) and (24),

$$0 = y^{**}(x_r^*) = x^*(y),$$

contradicting $x^*(y) \neq 0$.

Let $y^* \in Y^*$. By the *Hahn–Banach extension theorem* (Theorem 1.B), there exists an $x^* \in X^*$ such that $x_r^* = y^*$. Therefore, by (23) and (24),

$$y^{**}(y^*) = y^{**}(x_r^*) = y^{**}(x^*) = x^*(y) = y^*(y).$$

This is (22). \square

Proposition 5. *Let X be a Banach space over \mathbb{K} .*

- (i) *If X^* is separable, X is also.*
- (ii) *Conversely, if X is separable and reflexive, then X^* is separable.*

In the following proof we will use Proposition 3 from Section 1.2 on the separation of convex sets. If $X = \{0\}$, then the statements are trivial. Let $X \neq \{0\}$.

Proof. Ad (i). Suppose that the set $\{f_1, f_2, \dots\}$ is dense in X^* . Since

$$\|f_n\| = \sup_{\|u\|=1} |\langle f_n, u \rangle|,$$

there is a $u_n \in X$ such that $\|u_n\| = 1$ and

$$|\langle f_n, u_n \rangle| \geq 2^{-1} \|f_n\|.$$

Let Y be the closure of $\text{span}\{u_1, u_2, \dots\}$. Assume that $Y \neq X$. By Proposition 3 in Section 1.2, there exists a functional $f \in X^*$ such that $f(u) = 0$ on Y and $f \neq 0$. Thus, for all n ,

$$\begin{aligned} \|f - f_n\| &\geq |\langle f - f_n, u_n \rangle| = |\langle f_n, u_n \rangle| \geq 2^{-1} \|f_n\| \\ &\geq 2^{-1} (\|f\| - \|f - f_n\|). \end{aligned}$$

Since $\{f_1, f_2, \dots\}$ is dense in X^* , this implies

$$\|f\| \leq 3 \inf_n \|f - f_n\| = 0,$$

contradicting $f \neq 0$.

Ad (ii). By Proposition 3, there is a bijective map $j: X \rightarrow X^{**}$ such that $\|j(u)\| = \|u\|$ for all $u \in X$. Thus, the separability of X implies the separability of X^{**} . Since $X^{**} = (X^*)^*$, the space X^* is also separable, by (i). \square

The following two crucial results generalize the corresponding properties of weak convergence for Hilbert spaces, which we proved in Section 2.4.

Proposition 6. *Each bounded sequence (u_n) in a reflexive Banach space X over \mathbb{K} has a weakly convergent subsequence.*

Corollary 7. *The limit point of each weakly convergent subsequence of (u_n) belongs to the closed convex hull of the set $\{u_1, u_2, \dots\}$.*

The proof of Corollary 7 will be based on the *separation of convex sets* by closed hyperplanes (Theorem 1.C).

Proof of Proposition 6. The proof proceeds similarly to the proof of Theorem 2.C. However, instead of the Riesz theorem on Hilbert spaces we will use the *reflexivity* of the Banach space X .

For $X = \{0\}$, the statement is trivial. Let $X \neq \{0\}$.

Step 1: Suppose first that X^* is *separable*. Let $\{v_k\}$ be a countable set in X^* . Then

$$|\langle v_k, u_n \rangle| \leq \|v_k\| \|u_n\| \quad \text{for all } n, k.$$

As in the proof of Theorem 2.C, we obtain a subsequence (w_n) of (u_n) such that, for each v_k ,

$$\langle v_k, w_n \rangle \rightarrow a_k \quad \text{as } n \rightarrow \infty.$$

Moreover, there exist numbers $a(v)$ such that

$$\langle v, w_n \rangle \rightarrow a(v) \quad \text{as } n \rightarrow \infty \quad \text{for each } v \in X^*. \quad (25)$$

This follows from

$$\begin{aligned} |\langle v, w_n - w_m \rangle| &= |\langle v - v_k, w_n - w_m \rangle + \langle v_k, w_n - w_m \rangle| \\ &\leq \|v - v_k\| \|w_n - w_m\| + |\langle v_k, w_n - w_m \rangle| < \varepsilon, \end{aligned}$$

for suitable v_k and all $n, m \geq n_0(\varepsilon)$. Note that the sequence (w_n) is bounded and the set $\{v_k\}$ is dense in X^* .

Obviously, the map $v \mapsto a(v)$ is linear and, by (25),

$$|a(v)| \leq \|v\| \sup_n \|w_n\| \quad \text{for all } v \in X^*.$$

Hence $a \in X^{**}$. Since the Banach space X is *reflexive*, there exists a $w \in X$ such that

$$a(v) = \langle v, w \rangle \quad \text{for all } v \in X^*.$$

Thus, it follows from (25) that

$$w_n \rightharpoonup w \quad \text{as } n \rightarrow \infty.$$

Step 2: If X^* is *not* separable, then let Y be the closure of $\text{span}\{u_1, u_2, \dots\}$. It follows as in the proof of Theorem 2.C that Y is a separable closed linear subspace of X . By Proposition 4, Y is reflexive. Proposition 5 tells us that Y^* is *separable*.

Applying Step 1 to the space Y , we see that a subsequence (w_n) of (u_n) and a $w \in Y$ exist such that

$$\langle y^*, w_n \rangle \rightarrow \langle y^*, w \rangle \quad \text{as } n \rightarrow \infty \quad \text{for all } y^* \in Y^*.$$

Since $X^* \subseteq Y^*$, this implies $w_n \rightharpoonup w$ as $n \rightarrow \infty$. \square

Proof of Corollary 7. Let M be a closed convex set such that $u_n \in M$ for all n and $u_{n'} \rightharpoonup u$ as $n' \rightarrow \infty$. We have to show that $u \in M$.

Suppose that $u \notin M$. Then, by the *separation theorem for convex sets* from Section 1.2 (Theorem 1.C), there exists a functional $v^* \in X^*$ such that

$$\operatorname{Re} \langle v^*, v \rangle \leq 1 \quad \text{for all } v \in M$$

and $\operatorname{Re} \langle v^*, u \rangle > 1$. Letting $n' \rightarrow \infty$, it follows from $\operatorname{Re} \langle v^*, u_{n'} \rangle \leq 1$ for all n' that

$$\operatorname{Re} \langle v^*, u \rangle \leq 1.$$

This is a contradiction. \square

The following classic result will be used in the next section.

Lemma 8. Let $\phi: J \subseteq \mathbb{R} \rightarrow \mathbb{R}$ be a differentiable function on the interval J . Then, ϕ is convex iff the derivative $\phi': J \rightarrow \mathbb{R}$ is monotone, that is, $t \leq s$ implies $\phi'(t) \leq \phi'(s)$.

Proof. Suppose that ϕ is convex. Let $t < \tau < s$. It follows from

$$\tau = \frac{\tau - t}{s - t}s + \frac{s - \tau}{s - t}t$$

that

$$\phi(\tau) \leq \frac{\tau - t}{s - t}\phi(s) + \frac{s - \tau}{s - t}\phi(t), \tag{26}$$

and hence

$$\frac{\phi(\tau) - \phi(t)}{\tau - t} \leq \frac{\phi(s) - \phi(\tau)}{s - \tau}. \tag{27}$$

Letting $\tau \rightarrow t$ or $\tau \rightarrow s$, we get

$$\phi'(t) \leq \frac{\phi(s) - \phi(t)}{s - t} \leq \phi'(s).$$

Conversely, if ϕ' is monotone, then the mean value theorem yields

$$\frac{\phi(\tau) - \phi(t)}{\tau - t} = \phi'(\alpha) \quad \text{and} \quad \frac{\phi(s) - \phi(\tau)}{s - \tau} = \phi'(\beta),$$

where $t < \alpha < \tau$ and $\tau < \beta < s$. This implies (27) and hence we get (26), which says that ϕ is convex. \square

2.9 Applications to Convex Minimum Problems and Variational Inequalities

Let us consider the *minimum problem*

$$F(u) = \min !, \quad u \in M, \tag{28}$$

along with the *variational inequality*

$$\delta F(u; v - u) \geq 0 \quad \text{for all } v \in M \text{ and fixed } u \in M, \tag{28*}$$

which corresponds to

$$\langle F'(u), v - u \rangle \geq 0 \quad \text{for all } v \in M \text{ and fixed } u \in M. \tag{28**}$$

We assume that

- (H1) M is a nonempty, closed, convex subset of the real reflexive Banach space X .
- (H2) The functional $F: M \subseteq X \rightarrow \mathbb{R}$ is continuous and convex.
- (H3) If M is unbounded, then F is weakly coercive.

For example, assumptions (H2) and (H3) are satisfied if

$$F(u) := \|u - u_0\| \quad \text{for all } u \in X \text{ and fixed } u_0 \in X.$$

In the following, the postulated existence of the first variation $\delta F(u; h)$ or of the Gâteaux derivative $F'(u)$ includes tacitly that the functional F is defined on an open neighborhood of the point u .

Theorem 2.E. *Assume (H1) through (H3). Then the following are true:*

- (i) *The minimum problem (28) has a solution u . If F is strictly convex, then this solution is unique.*
- (ii) *If the first variation $\delta F(v; h)$ exists for all $v \in M$ and all $h \in X$, then the minimum problem (28) is equivalent to the variational inequality (28*).*
- (iii) *If the Gâteaux derivative $F'(v)$ exists for all $v \in M$, then the minimum problem (28) is equivalent to the variational inequality (28**).*

Proof. Ad (i). This follows from Proposition 4 and Remark 6 in Section 2.5.

Ad (ii). Let $u, v \in M$. Since M is convex, we get $u + t(v - u) \in M$ for all $t \in [0, 1]$. Define

$$\phi(t) := F(u + t(v - u)) \quad \text{for all } t \in [0, 1].$$

If u is a solution of (28), then

$$\phi(t) \geq \phi(0) \quad \text{for all } t \in [0, 1].$$

Hence

$$\delta F(u; v - u) = \phi'(0) = \lim_{t \rightarrow +0} t^{-1}(\phi(t) - \phi(0)) \geq 0.$$

Conversely, let u be a solution of (28*). Then, $\phi'(0) \geq 0$. Since ϕ is convex on $[0, 1]$, the derivative ϕ' is monotone on $[0, 1]$. By the mean value theorem, there is a number $0 < \theta < 1$ such that

$$\phi(1) - \phi(0) = \phi'(\theta) \geq \phi(0) \geq 0.$$

Hence

$$F(v) - F(u) \geq 0 \quad \text{for all } v \in M,$$

that is, u is a solution of (28).

Ad (iii). Observe that $\delta F(u; h) = \langle F'(u), h \rangle$. □

Proposition 1. *Let $F: X \rightarrow \mathbb{R}$ be a continuous, convex, weakly coercive functional on the real reflexive Banach space X . Suppose that the Gâteaux derivative $F'(v)$ exists for each $v \in X$. Then the following are true:*

- (i) *The minimum problem*

$$F(u) = \min !, \quad u \in X, \tag{29}$$

has a solution u . This solution is unique if F is strictly convex.

- (ii) *The minimum problem (29) is equivalent to the operator equation*

$$F'(u) = 0 \quad (\text{Euler equation}). \tag{29*}$$

Proof. This is a special case of Theorem 2.E. Observe that the variational inequality

$$\langle F'(u), v - u \rangle \geq 0 \quad \text{for all } v \in X$$

is equivalent to $\langle F'(u), \pm h \rangle \geq 0$ for all $h \in X$. In turn, this is equivalent to $F'(u) = 0$. \square

Proposition 2. Let $F: M \subseteq X \rightarrow \mathbb{R}$ be a convex functional on the convex subset M of the real normed space X .

Then, each local minimum u_0 of F on M is also a global minimum of F on M .

Proof. There is a number $r > 0$ such that

$$F(u_0) \leq F(v) \quad \text{for all } v \in M \text{ with } \|v - u_0\| < r.$$

Let $u \in M$ be given. Set $v := u_0 + \alpha(u - u_0)$. If $\alpha > 0$ is sufficiently small, then $\|v - u_0\| < r$. By the convexity of F ,

$$F(u_0) \leq F(v) \leq \alpha F(u) + (1 - \alpha)F(u_0).$$

Hence $F(u_0) \leq F(u)$ for all $u \in M$. \square

The following two corollaries characterize convex functionals in terms of the Gâteaux derivative and the second variation, respectively.

Corollary 3 (Convex functionals and monotone operators). Let $F: X \rightarrow \mathbb{R}$ be a functional on the real normed space X such that the Gâteaux derivative $F'(u)$ exists for all $u \in X$.

Then F is convex on X iff the operator $F': X \rightarrow X^*$ is monotone, that is,

$$\langle F'(v) - F'(u), v - u \rangle \geq 0 \quad \text{for all } u, v \in X. \quad (30)$$

Proof. Let $u, v \in X$. Set

$$\phi(t) := F(u + t(v - u)) \quad \text{for all } t \in \mathbb{R}. \quad (31)$$

Then,

$$\phi'(t) = \delta F(u + t(v - u); v - u) = \langle F'(u + t(v - u)), v - u \rangle \quad \text{for all } t \in \mathbb{R}. \quad (32)$$

If F is convex on X , then so is ϕ on \mathbb{R} . By Lemma 8 in Section 2.8, the derivative ϕ' is monotone on \mathbb{R} . Hence

$$\phi'(1) \geq \phi'(0).$$

This is (30).

Conversely, if F' is monotone on X , then it follows from (30) that

$$\langle F'(u + s(v - u)) - F'(u + t(v - u)), (s - t)(v - u) \rangle \geq 0,$$

for all $u, v \in X$ and $t, s \in \mathbb{R}$. By (32), this implies

$$\phi'(t) \leq \phi'(s) \quad \text{for all } t \leq s.$$

Thus, ϕ' is monotone on \mathbb{R} . By Lemma 8 in Section 2.8, ϕ is convex on \mathbb{R} . Hence F is convex on \mathbb{R} . \square

Corollary 4 (Convex functionals and the definiteness of the second variation). *Let $F: X \rightarrow \mathbb{R}$ be a functional on the real normed space X such that the second variation $\delta^2 F(u; h)$ exists for all $u, h \in X$. Then*

(i) *F is convex on X iff*

$$\delta^2 F(u; h) \geq 0 \quad \text{for all } u, h \in X. \quad (33)$$

(ii) *If*

$$\delta^2 F(u; h) > 0 \quad \text{for all } u, h \in X \text{ with } h \neq 0, \quad (33^*)$$

then F is strictly convex on X .

Proof. Ad (i). Let us use the function ϕ introduced in (31). Observe that

$$\phi''(t) = \delta^2 F(u + t(v - u); v - u) \quad \text{for all } t \in \mathbb{R}.$$

If F is convex on X , then ϕ is convex on \mathbb{R} . By Lemma 8 in Section 2.8, ϕ' is monotone on \mathbb{R} . Hence $\phi''(t) \geq 0$ for all $t \in \mathbb{R}$. Letting $t = 0$, we get (33).

Conversely, it follows from (33) that $\phi''(t) \geq 0$ for all $t \in \mathbb{R}$. Thus, ϕ' is monotone on \mathbb{R} , and hence ϕ is convex on \mathbb{R} . This implies the convexity of F on X .

Ad (ii). Let $v \neq u$. It follows from (33*) that $\phi''(t) > 0$ for all $t \in \mathbb{R}$. Hence ϕ' is strictly monotone on \mathbb{R} . The proof of Lemma 8 in Section 2.8 tells us that ϕ is strictly monotone on \mathbb{R} . Thus, F is strictly convex on X . \square

Example 5. A real Banach space is called *strictly convex* iff the norm function

$$u \mapsto \|u\|$$

is strictly convex. Each real Hilbert space X is strictly convex.

Proof. Set $F(u) := (u | u)$ for all $u \in X$. Then,

$$\delta^2 F(u; h) = (h | h) > 0 \quad \text{for all } h \in X, h \neq 0,$$

by Example 4 in Section 2.1. Hence F is strictly convex.

Set $G(u) := \|u\| = \psi(F(u))$, where $\psi(x) := x^{\frac{1}{2}}$ for all $x \geq 0$. Since the real function $\psi: [0, \infty[\rightarrow \mathbb{R}$ is strictly increasing and convex, the functional $G: X \rightarrow \mathbb{R}$ is strictly convex. In fact, let $u, v \in X$ be given such that $u \neq v$, and let $t \in]0, 1[$. Then

$$\begin{aligned} G(tu + (1-t)v) &= \psi(F(tu + (1-t)v)) < \psi(tF(u) + (1-t)F(v)) \\ &\leq t\psi(F(u)) + (1-t)\psi(F(v)) = tG(u) + (1-t)G(v). \quad \square \end{aligned}$$

The following result generalizes the main theorem on quadratic variational problems from Section 2.4 of AMS Vol. 108. An application to elasticity will be considered in the next section.

Proposition 6 (Quadratic variational inequalities). *Suppose that*

- (a) *$a: X \times X \rightarrow \mathbb{R}$ is a symmetric, bounded, strongly monotone, bilinear form, where X is a real Hilbert space.*
- (b) *$b: X \rightarrow \mathbb{R}$ is a linear continuous functional.*
- (c) *M is a nonempty, closed, convex subset of X .*

Then, the variational problem

$$2^{-1}a(u, u) - b(u) = \min !, \quad u \in M,$$

has a unique solution u , which is also the unique solution of the variational inequality

$$a(u, v - u) - b(v - u) \geq 0 \quad \text{for all } v \in M \text{ and fixed } u \in M.$$

Proof. Set

$$F(u) := 2^{-1}a(u, u) - b(u) \quad \text{for all } u \in X.$$

Introducing the real function $\phi(t) := F(u + th)$ for all $t \in \mathbb{R}$ and fixed $u, h \in X$, we get

$$\phi(t) = 2^{-1}a(u, u) + ta(u, h) + 2^{-1}a(h, h) - b(u) - tb(h).$$

By definition, $\delta^n F(u; h) := \phi^{(n)}(0)$. Thus, for all $u, h \in X$, $\delta F(u; h) = a(u, h) - b(h)$, $\delta^2 F(u; h) = 2^{-1}a(h, h)$, and $\delta^n F(u; h) = 0$ if $n \geq 3$.

Since $a(\cdot, \cdot)$ is strongly monotone, there exists a constant $c > 0$ such that $a(u, u) \geq c\|u\|^2$ for all $u \in X$, and hence

$$\delta^2 F(u; h) \geq 2^{-1}c\|h\|^2 \quad \text{for all } h \in X.$$

By Corollary 4, $F: X \rightarrow \mathbb{R}$ is *strictly convex*.

Moreover, it follows from Proposition 2 in Section 2.3 of AMS Vol. 108 that, as $n \rightarrow \infty$,

$$u_n \rightarrow u \quad \text{implies} \quad F(u_n) \rightarrow F(u),$$

that is, $F: X \rightarrow \mathbb{R}$ is *continuous*.

Finally, since $|b(u)| \leq \|b\| \|u\|$, we get

$$F(u) \geq 2^{-1}c\|u\|^2 - \|b\| \|u\| \quad \text{for all } u \in X \text{ and fixed } c > 0.$$

Hence $F(u) \rightarrow +\infty$ as $\|u\| \rightarrow \infty$, so that $F: X \rightarrow \mathbb{R}$ is *weakly coercive*.

The assertion now follows from Theorem 2.E. \square

2.10 Applications to Obstacle Problems in Elasticity

Let us consider the following *minimum problem*:

$$F(u) := 2^{-1} \int_G \sum_{j=1}^N (\partial_j u)^2 dx - \int_G f u dx = \min !, \quad u \in M, \quad (34)$$

where

$$M := \{u \in \overset{\circ}{W}_2^1(G) : u(x) \geq 0 \text{ for almost all } x \in G\},$$

along with the following *variational inequality*:

$$\int_G \sum_{j=1}^N \partial_j u (\partial_j v - \partial_j u) dx - \int_G f(v - u) dx \geq 0 \quad \text{for all } v \in M \quad (34^*)$$

and fixed $u \in M$.

Proposition 1. *We are given the function $f \in L_2(G)$, where G is a nonempty, bounded, open set in \mathbb{R}^N , $N \geq 1$.*

Then the minimum problem (34) has a unique solution u , which is also the unique solution of the variational inequality (34).*

In the special case where $N = 1$ and $G =]a, b[$, problem (34) allows the following *physical interpretation*. As in Section 2.7 of AMS Vol. 108, we use

$$\begin{aligned} u(x) &= \text{deflection of a string at the point } x; \\ f(x) &= \text{force density at the point } x. \end{aligned}$$

Then, problem (34) corresponds to the principle of *minimal potential energy*. The side condition $u \in M$ postulates that

$$u(x) \geq 0 \quad \text{for almost all } x \in]a, b[\quad (\text{obstacle condition})$$

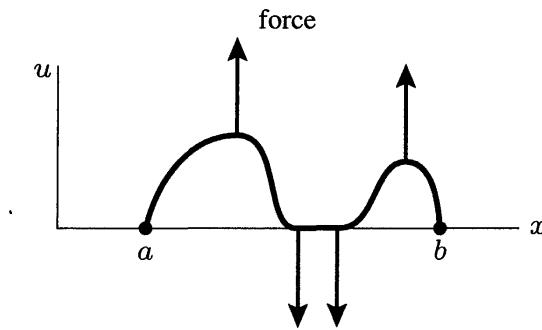


FIGURE 2.3.

and

$$u(a) = u(b) = 0 \quad (\text{boundary condition}).$$

A possible physical situation is pictured in Figure 2.3.

Proof. Set $X := \overset{\circ}{W}_2^1(G)$ and

$$\begin{aligned} a(u, v) &:= \int_G \sum_{j=1}^N \partial_j u \partial_j v \, dx, \\ b(u) &:= \int_G f u \, dx, \quad \text{for all } u, v \in X. \end{aligned}$$

By the proof of Theorem 2.B (Dirichlet principle) in AMS Vol. 108, $a: X \times X \rightarrow \mathbb{R}$ is a symmetric, bounded, strongly monotone, bilinear form, and $b: X \rightarrow \mathbb{R}$ is linear and continuous.

Obviously, the set M is *convex*. Moreover, M is also *closed*. In fact, if

$$u_n \rightarrow u \text{ in } M \quad \text{as } n \rightarrow \infty,$$

then there exists a subsequence $(u_{n'})$ such that

$$u_{n'}(x) \rightarrow u(x) \quad \text{as } n' \rightarrow \infty \quad \text{for almost all } x \in G,$$

by Lemma 2 in Section 2.6. Thus, $u_n(x) \geq 0$ for almost all $x \in G$ and all n implies $u(x) \geq 0$ for almost all $x \in G$, and hence $u \in M$.

The assertion follows now from Proposition 6 in Section 2.9. \square

2.11 Saddle Points

Definition 1. Let the function $L: A \times B \rightarrow \mathbb{R}$ be given, where A and B are arbitrary nonempty sets. The point $(u_0, p_0) \in A \times B$ is called a *saddle point* of L with respect to $A \times B$ iff

$$L(u_0, p) \leq L(u_0, p_0) \leq L(u, p_0) \quad \text{for all } u \in A, p \in B. \quad (35)$$

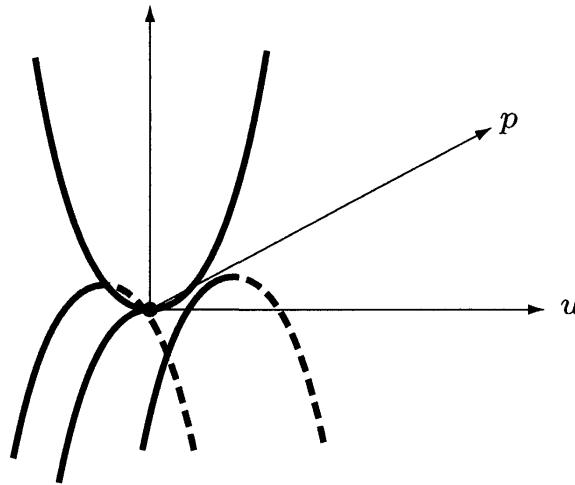


FIGURE 2.4.

This is equivalent to

$$\max_{p \in B} L(u_0, p) = L(u_0, p_0) = \min_{u \in A} L(u, p_0). \quad (35^*)$$

Example 2. Let $L(u, p) := u^2 - p^2$. Then, $(0, 0)$ is a saddle point of L with respect to $\mathbb{R} \times \mathbb{R}$ (cf. Figure 2.4).

In the following three sections we want to show that saddle points play an important role in duality theory and in game theory.

2.12 Applications to Duality Theory

The point of departure for the formulation of dual problems is the following symmetric pair of formulas:

$$\inf_{u \in A} \left(\sup_{p \in B} L(u, p) \right) = \alpha, \quad (36)$$

$$\sup_{p \in B} \left(\inf_{u \in A} L(u, p) \right) = \beta. \quad (36^*)$$

Theorem 2.F. Let $L: A \times B \rightarrow \mathbb{R}$ be a function on the product $A \times B$ of the nonempty sets A and B . Then, the following two statements are equivalent:

- (i) (u_0, p_0) is a saddle point of L with respect to $A \times B$.
- (ii) u_0 is a solution of the primal problem (36), p_0 is a solution of the dual problem (36*), and $\alpha = \beta$.

Corollary 1. If (ii) holds, then $\alpha = \beta = L(u_0, p_0)$.

Suppose that we are given the primal problem

$$\inf_{u \in A} F(u) = \alpha. \quad (37)$$

In order to find the corresponding dual problem, we are looking for a function L such that

$$F(u) = \sup_{p \in B} L(u, p).$$

Thus, problem (37) is identical to (36). Letting

$$G(p) := \inf_{u \in A} L(u, p),$$

the dual problem corresponding to (36*) reads as follows:

$$\sup_{p \in B} G(p) = \beta. \quad (37^*)$$

Corollary 2. Let $L: A \times B \rightarrow \mathbb{R}$, where A and B are nonempty sets.

- (i) Then $-\infty \leq \beta \leq \alpha \leq \infty$.
- (ii) Let $u \in A$ and $p \in B$ be given. Then we get the following error estimates for the extremal values α and β :

$$G(p) \leq \beta \leq \alpha \leq F(u).$$

- (iii) Suppose that we know two points $u_0 \in A$ and $p_0 \in B$ such that

$$\overbrace{G(p_0)} \geq F(u_0). \quad (38)$$

Then u_0 is a solution of the primal problem (37), p_0 is a solution of the dual problem (37*), and $\alpha = \beta$.

Proof of Corollary 2. Ad (i). Obviously,

$$\inf_{u \in A} L(u, p) \leq L(v, p) \quad \text{for all } v \in A,$$

and hence

$$\sup_{p \in B} \inf_{u \in A} L(u, p) \leq \sup_{p \in B} L(v, p) \quad \text{for all } v \in A.$$

Therefore,

$$\beta = \sup_{p \in B} \inf_{u \in A} L(u, p) \leq \inf_{v \in A} \sup_{p \in B} L(v, p) = \alpha.$$

Ad (ii). This follows from (i).

Ad (iii). By (ii),

$$G(p_0) \leq \beta \leq \alpha \leq F(u_0).$$

Thus, it follows from (38) that $G(p_0) = \beta = \alpha = F(u_0)$. \square

Proof of Theorem 2.F. (i) \Rightarrow (ii). Since (u_0, p_0) is a saddle point of L ,

$$\sup_{p \in B} L(u_0, p) = L(u_0, p_0) = \inf_{u \in A} L(u, p_0). \quad (39)$$

Therefore,

$$\alpha = \inf_{u \in A} \sup_{p \in B} L(u, p) \leq L(u_0, p_0) \leq \sup_{p \in B} \inf_{u \in A} L(u, p) = \beta.$$

Since $\beta \leq \alpha$, by Corollary 2, the equality sign appears everywhere, that is,

$$\alpha = L(u_0, p_0) = \beta,$$

and (39) tells us that

$$F(u_0) = L(p_0, u_0) = G(p_0).$$

This proves statement (ii). \square

(ii) \Rightarrow (i). From (ii) it follows that $\alpha = F(u_0)$ and $\beta = G(p_0)$. Hence

$$\beta = G(p_0) \equiv \inf_{u \in A} L(u, p_0) \leq L(u_0, p_0) \leq \sup_{p \in B} L(u_0, p) \equiv F(u_0) = \alpha.$$

Since $\alpha = \beta$, the equality sign appears everywhere. Hence (u_0, p_0) is a saddle point of L with respect to $A \times B$.

This also proves Corollary 1. \square

Application of this general duality theory to numerous optimization problems can be found in Zeidler (1986), Vol. 3, Chapters 49 through 53.

In the following section we represent fundamental results on the existence of saddle points.

2.13 The von Neumann Minimax Theorem on the Existence of Saddle Points

Our goal is the relation

$$L(u_0, p_0) = \min_{u \in A} \max_{p \in B} L(u, p) = \max_{p \in B} \min_{u \in A} L(u, p). \quad (40)$$

By definition, a functional f is called *concave* (resp., upper semicontinuous) iff $-f$ is *convex* (resp., lower semicontinuous). We assume that

- (H1) The functional $L: A \times B \rightarrow \mathbb{R}$ is given, where A and B are nonempty, closed, convex sets in the real, strictly convex, reflexive Banach space X (e.g., X is a real Hilbert space).
- (H2) The functional $u \mapsto L(u, p)$ is *convex* and lower semicontinuous (e.g., continuous) on A , for each $p \in B$.
- (H3) The functional $p \mapsto L(u, p)$ is *concave* and upper lower semicontinuous (e.g., continuous) on B , for each $u \in A$.
- (H4) The sets A and B are *bounded*.

Theorem 2.G. *The functional L has a saddle point (u_0, p_0) with respect to $A \times B$ and the relation (40) holds true.*

Moreover, relation (40) is valid for each saddle point (u_0, p_0) of L .

For the special case where $X = \mathbb{R}^N$, John von Neumann proved this theorem in 1928. His paper marked the birth of mathematical game theory. A more general result than Theorem 2.G can be found in Zeidler (1986), Vol. 1, Section 9.6. There the proof is based on a fixed-point theorem for multivalued mappings. Other proofs make use of the lemma of Knaster, Kuratowski, and Mazurkiewicz or of the Hahn–Banach theorem (separation of convex sets). The following proof uses only *elementary arguments* based on weak convergence. \checkmark

Proof. By (H2), the functional $u \mapsto L(u, p)$ is convex and lower semicontinuous on A . Hence $u_n \rightharpoonup u$ as $n \rightarrow \infty$ on A implies

$$L(u, p) \leq \varliminf_{n \rightarrow \infty} L(u_n, p) \quad \text{for all } p \in B,$$

by Lemma 5 and Remark 6 in Section 2.5.

Furthermore, by (H3), the functional $p \mapsto -L(u, p)$ is convex and lower semicontinuous on B . Hence $p_n \rightharpoonup p$ as $n \rightarrow \infty$ on B implies

$$-L(u, p) \leq \varliminf_{n \rightarrow \infty} -L(u, p_n) \quad \text{for all } u \in A.$$

This is equivalent to

$$L(u, p) \geq \overline{\lim}_{n \rightarrow \infty} L(u, p_n).$$

Step 1: We set

$$G(p) := \min_{u \in A} L(u, p) \quad \text{for all } p \in B, \tag{41}$$

and

$$F(u) := \max_{p \in B} L(u, p) \quad \text{for all } u \in A. \tag{41*}$$

These definitions make sense. In fact, since the functional $u \mapsto L(u, p)$ is convex and lower semicontinuous, the minimum problem (41) has a solution, by Proposition 4 and Remark 6 in Section 2.5.

Replacing L with $-L$, the same argument tells us that the maximum problem (41*) has a solution.

Step 2: We show that the functional $F: A \rightarrow \mathbb{R}$ is *quasi-convex* and *lower semicontinuous*. To this end, put

$$A_r := \{u \in A : F(u) \leq r\} \quad \text{for each } r \in \mathbb{R}.$$

Then, it follows from $v, w \in A_r$ and $z := \alpha v + (1 - \alpha)w$ with $\alpha \in [0, 1]$ that

$$L(v, p), L(w, p) \leq r \quad \text{for all } p \in B.$$

By the convexity of $u \mapsto L(u, p)$, this implies

$$L(z, p) \leq \alpha L(v, p) + (1 - \alpha)L(w, p) \leq r \quad \text{for all } p \in B,$$

and hence $F(z) \leq r$ (i.e., $z \in A_r$).

Moreover, if $u_n \in A_r$ for all n and $u_n \rightarrow u$ as $n \rightarrow \infty$, then

$$L(u_n, p) \leq r \quad \text{for all } n \text{ and all } p \in B.$$

Since $u \mapsto L(u, p)$ is semicontinuous, we get $\limsup_{n \rightarrow \infty} L(u_n, p) \leq r$ for all $p \in B$. Hence $F(u) \leq r$, that is, $u \in A_r$.

Replacing L with $-L$, we see that the function $-G$ is convex and lower semicontinuous on B . By Proposition 4 and Remark 6 in Section 2.5, the minimum problem

$$F(u_*) = \min_{u \in A} F(u)$$

and the maximum problem

$$G(p_0) = \max_{p \in B} G(p)$$

have solutions u_* and p_0 , respectively.

Step 3: For the moment, let us assume that

(H) The functional $u \mapsto L(u, p)$ is *strictly convex* on A .

Then, for each $p \in B$, the minimum problem

$$G(p) = \min_{u \in A} L(u, p) \tag{42}$$

has a *unique* solution called $u := \phi(p)$. Hence

$$G(p) = \min_{u \in A} L(u, p) = L(\phi(p), p) \quad \text{for all } p \in B. \tag{43}$$

Set

$$u_0 := \phi(p_0).$$

By (43),

$$G(p_0) \leq L(u, p_0) \quad \text{for all } u \in A. \quad (44)$$

We shall show in Step 4 that the following *key inequality* is valid:

$$G(p_0) \geq L(u_0, p) \quad \text{for all } p \in B. \quad (44^*)$$

It follows from (44) and (44^{*}) that $G(p_0) = L(u_0, p_0)$, and hence

$$L(u_0, p) \leq L(u_0, p_0) \leq L(u, p_0) \quad \text{for all } u \in A, p \in B.$$

Consequently, (u_0, p_0) is a *saddle point* of L with respect to $A \times B$. By Theorem 2.F, we get $\alpha = \beta$, i.e.,

$$\sup_{p \in B} \inf_{u \in A} L(u, p) = \inf_{u \in A} \sup_{p \in B} L(u, p) = L(u_0, p_0).$$

According to Steps 1 and 2, we may replace “sup” and “inf” with “max” and “min.” Hence

$$\max_{p \in B} \min_{u \in A} L(u, p) = \min_{u \in A} \max_{p \in B} L(u, p) = L(u_0, p_0).$$

Step 4: We prove the decisive inequality (44^{*}). Let $p \in B$. Put

$$p_n := (1 - n^{-1})p_0 + n^{-1}p \quad \text{and} \quad u_n := \phi(p_n) \quad \text{for } n = 1, 2, \dots.$$

By (43),

$$G(p_0) \geq G(p_n) = L(u_n, p_n) \quad \text{for } n = 1, 2, \dots.$$

Since $p \mapsto L(u, p)$ is *concave*,

$$G(p_0) \geq L(u_n, p_n) \geq (1 - n^{-1})L(u_n, p_0) + n^{-1}L(u_n, p).$$

By (43), $G(p_0) \leq L(u_n, p_0)$. Hence $G(p_0) \geq (1 - n^{-1})G(p_0) + n^{-1}L(u_n, p)$, that is,

$$G(p_0) \geq L(u_n, p) \quad \text{for all } n = 1, 2, \dots \text{ and all } p \in B. \quad (45)$$

Since $u_n \in A$ for all n , the sequence (u_n) is bounded. Thus, there exists a subsequence, again denoted by (u_n) , such that

$$u_n \rightharpoonup w \quad \text{as } n \rightarrow \infty.$$

By (45),

$$G(p_0) \geq \lim_{n \rightarrow \infty} L(u_n, p) \geq L(w, p) \quad \text{for all } p \in B. \quad (46)$$

It remains to show that $w = u_0$. By (43) and the definition of u_n ,

$$L(u_n, p_n) \leq L(u, p_n) \quad \text{for all } u \in A, n = 1, 2, \dots.$$

Since the functional $p \mapsto L(u, p)$ is concave, this implies

$$(1 - n^{-1})L(u_n, p_0) + n^{-1}L(u_n, p) \leq L(u, p_n) \quad \text{for all } p \in B, n = 1, 2, \dots. \quad (47)$$

By (43), $L(u_n, p) \geq G(p)$. Hence

$$(1 - n^{-1})L(u_n, p_0) + n^{-1}G(p) \leq L(u, p_n) \quad \text{for all } p \in B, n = 1, 2, \dots.$$

If we let $n \rightarrow \infty$, it follows from (46) and (47) that

$$L(w, p_0) \leq \underline{\lim}_{n \rightarrow \infty} L(u, p_n) \quad \text{for all } u \in A.$$

Since $p_n \rightarrow p_0$ as $n \rightarrow \infty$, $\overline{\lim}_{n \rightarrow \infty} L(u, p_n) \leq L(u, p_0)$, and hence

$$L(w, p_0) \leq L(u, p_0) \quad \text{for all } u \in A.$$

Thus, it follows from (43) that $w = \phi(p_0)$, and hence $w = u_0$.

Under the additional hypothesis (H), the proof of Theorem 2.G has been finished.

Step 5: If condition (H) is not satisfied, then we use the *regularized* functions

$$L_n(u, p) := L(u, p) + n^{-1}\|u\| \quad \text{for } n = 1, 2, \dots. \quad (47)$$

Since the Banach space X is strictly convex, the function $u \mapsto \|u\|$ is strictly convex. Consequently, the function L_n satisfies condition (H) together with all the other assumptions, (H1) through (H4) (cf. Problem 2.7). By the preceding proof, there exists a saddle point (u_n, p_n) of L_n with respect to $A \times B$, and hence

$$L(u_n, p) + n^{-1}\|u_n\| \leq L(u_n, p_n) + n^{-1}\|u_n\| \leq L(u, p_n) + n^{-1}\|u\|, \quad (48)$$

for all $u \in A$ and $p \in B$.

Since (u_n) and (p_n) are bounded, there exist subsequences, again denoted by (u_n) and (p_n) , such that

$$u_n \rightharpoonup u \quad \text{and} \quad p_n \rightharpoonup p_0 \quad \text{as } n \rightarrow \infty.$$

The sets A and B are closed and convex. Hence $u_0 \in A$ and $p_0 \in B$. Letting $n \rightarrow \infty$ in (48), we obtain

$$\begin{aligned} L(u_0, p) &\leq \underline{\lim}_{n \rightarrow \infty} L(u_n, p) + n^{-1}\|u_n\| \leq \overline{\lim}_{n \rightarrow \infty} L(u, p_n) + n^{-1}\|u\| \\ &\leq L(u, p_0) \quad \text{for all } u \in A, p \in B. \end{aligned}$$

Hence

$$L(u_0, p) \leq L(u_0, p_0) \leq L(u, p_0) \quad \text{for all } u \in A, p \in B.$$

Thus, (u_0, p_0) is a saddle point of L with respect to $A \times B$. As in Step 3, this implies (40). \square

We now replace the boundedness of the sets A and B by the following more general condition:

H4*) If A is *not* bounded, then there exists a point $q \in B$ such that $L(u, q) \rightarrow +\infty$ as $\|u\| \rightarrow \infty$, $u \in A$.

If B is *not* bounded, then there exists a point $v \in A$ such that $L(v, p) \rightarrow -\infty$ as $\|p\| \rightarrow \infty$, $p \in B$.

Proposition 1. *Assume (H1) through (H3) and (H4*). Then, L has a saddle point with respect to $A \times B$.*

Proof. Set

$$A_n := \{u \in A : \|u\| \leq n\} \quad \text{and} \quad B_n := \{p \in B : \|p\| \leq n\} \quad n = 1, 2, \dots$$

For sufficiently large n , the sets A_n and B_n are not empty. By Theorem 2.G, the functional L has a saddle point (u_n, p_n) with respect to $A_n \times B_n$, that is,

$$L(u_n, p) \leq L(u_n, p_n) \leq L(u, p_n) \quad \text{for all } n, u \in A_n, p \in B_n. \quad (49)$$

Letting $u := v$ and $p := q$, this implies that the sequences (u_n) and (p_n) are *bounded*, by (H4*).

In fact, (49) yields

$$L(u_n, q) \leq L(v, p_n) \quad \text{for all } n. \quad (50)$$

It follows from (50) along with (H4*) that both the sequences (u_n) and (p_n) cannot be bounded. If (u_n) is bounded, then the sequence $((L(u_n, q))$ is bounded below.⁷ Thus, (p_n) must be bounded, by (50) and (H4*). Similarly,

⁷By Lemma 5 in Section 2.5, $u \mapsto L(u, p)$ is weakly sequentially lower semi-continuous. Suppose that there exists a subsequence, again denoted by (u_n) , such that

$$L(u_n, q) \rightarrow -\infty \quad \text{as } n \rightarrow \infty.$$

Since (u_n) is bounded, there is a subsequence, again denoted by (u_n) , such that $u_n \rightharpoonup u$ as $n \rightarrow \infty$. Hence

$$L(u, q) \leq \liminf_{n \rightarrow \infty} L(u_n, q).$$

This is a contradiction.

we obtain that the boundedness of (p_n) implies the boundedness of (u_n) . Thus, both the sequences (u_n) and (p_n) are bounded.

Passing over to subsequences, if necessary, we get

$$u_n \rightarrow u_0 \quad \text{and} \quad p_n \rightarrow p_0 \quad \text{as } n \rightarrow \infty.$$

By (49), for all $u \in A$ and $p \in B$,

$$L(u_0, p) \leq \underline{\lim}_{n \rightarrow \infty} L(u_n, p) \leq \overline{\lim}_{n \rightarrow \infty} L(u, p_n) \leq L(u, p_0).$$

Consequently, (u_0, p_0) is a saddle point of L . \square

2.14 Applications to Game Theory

Game theory is a mathematical search for the optimal balance of competing interests, such as between two partners. To explain this, let us consider two players \mathcal{P} and \mathcal{Q} having the strategy set A and B available, respectively. If \mathcal{P} chooses the strategy $u \in A$ and \mathcal{Q} having the strategy $p \in B$, then let

$$L(u, p) := \text{gain by } \mathcal{Q} = \text{loss by } \mathcal{P}$$

(e.g., in dollars). We allow $L(u, p)$ to be negative, and if this is the case, player \mathcal{Q} has a negative gain, that is, a loss of $|L(u, p)|$ dollars.

Definition 1. The pair (u_0, p_0) in $A \times B$ is called an *optimal strategy pair* iff (u_0, p_0) is a saddle point of the *gain function* L , with respect to $A \times B$, that is,

$$L(u_0, p) \leq L(u_0, p_0) \leq L(u, p_0) \quad \text{for all } u \in A \text{ and } p \in B.$$

Here, $L(u_0, p_0)$ is called the *value* of the game.

This definition reflects the fact that each player will play so as to maximize his or her interests.

- (i) If $L(u_0, p_0) = 0$, then the game ends undecided if both players play optimally. Neither player gains or loses anything. This means that the game is *fair*.
- (ii) If $L(u_0, p_0) > 0$ or < 0 , then \mathcal{Q} or \mathcal{P} , respectively, wins by an amount of $|L(u_0, p_0)|$ dollars if both players play optimally.

Example 2 (Coin game). Players \mathcal{P} and \mathcal{Q} each simultaneously displays a coin showing heads or tails. They agree that \mathcal{Q} will win one dollar if both

show the same. Otherwise, \mathcal{P} wins a dollar. In this case the strategy sets are

$$A = B = \{H, T\},$$

where H and T correspond to heads and tails, respectively. Here, the gain function L is given by

$$1 = L(H, H) = L(T, T) = -L(H, T) = -L(T, H).$$

Obviously,

$$\min_{u \in A} \max_{p \in B} L(u, p) = 1 \quad \text{and} \quad \max_{p \in B} \min_{u \in A} L(u, p) = -1.$$

Thus, it follows from Theorem 2.F that L has no saddle point—there can be *no* optimal strategy pair. This corresponds to our intuition in the matter.

In order to get a nice result, which also applies to the coin game, let us pass to a *probabilistic approach*. We assume that \mathcal{P} has the strategies E_1, \dots, E_N available, while \mathcal{Q} has F_1, \dots, F_N . We let

$$L(E_i, F_j) := \text{win for } \mathcal{Q} \text{ under strategy pair } (E_i, F_j).$$

Furthermore, we assume that

/

player \mathcal{P} chooses strategy E_i with probability p_i ,

while \mathcal{Q} chooses F_j with probability q_j . Set $p := (p_1, \dots, p_N)$ and $q := (q_1, \dots, q_N)$. Since p_i and q_j are probabilities, we get $p \in A$ and $p \in B$, where

$$A := \left\{ p \in \mathbb{R}^N : 0 \leq p_i \leq 1 \text{ for all } i, \sum_{i=1}^N p_i = 1 \right\},$$

$$B := \left\{ q \in \mathbb{R}^N : 0 \leq q_j \leq 1 \text{ for all } j, \sum_{j=1}^N q_j = 1 \right\}.$$

Definition 3. Each pair (p, q) in $A \times B$ is called a *mixed strategy*. The quantity

$$\mathcal{L}(p, q) := \sum_{i=1}^N \sum_{j=1}^N L(E_i, F_j) p_i q_j$$

is called the *expected win value* for player \mathcal{Q} .

The pair (p_0, q_0) in $A \times B$ is called an *optimal mixed strategy pair* iff (p_0, q_0) is a saddle point of \mathcal{L} .

The following main theorem of game theory was proved by John von Neumann in 1928.

Proposition 4. *Under the above hypotheses, there always exists an optimal mixed strategy pair, and*

$$\min_{p \in A} \max_{q \in B} \mathcal{L}(p, q) = \max_{q \in B} \min_{p \in A} \mathcal{L}(p, q). \quad (51)$$

For any optimally mixed strategy pair (p_0, q_0) , the common value of the two expressions in (51) is equal to $\mathcal{L}(p_0, q_0)$.

This is a special case of Theorem 2.G, since A and B are compact convex sets.

Example 5 (Probabilistic coin game). Parallel to Example 2, we assume that

Player \mathcal{P} chooses H (heads) and T (tails) with probability p_1 and p_2 , respectively.

Analogously, \mathcal{Q} chooses H and T with probability q_1 and q_2 , respectively. If we set $E_1 = F_1 = H$ and $E_2 = F_2 = T$, then the expected win value for player \mathcal{Q} is given by

$$\mathcal{L}(p, q) = \sum_{i=1}^2 \sum_{j=1}^2 L(E_i, F_j) = (p_1 - p_2)(q_1 - q_2).$$

We also set

$$A := \{(p_1, p_2) \in \mathbb{R}^2 : 0 \leq p_1 \leq 1, 0 \leq p_2 \leq 1, \text{ and } p_1 + p_2 = 1\},$$

and $B := A$. Thus, A is the line segment connecting the points $(0, 1)$ and $(1, 0)$ in \mathbb{R}^2 .

An elementary argument shows that the point

$$(p_0, q_0) \quad \text{with } p_0 = q_0 := \left(\frac{1}{2}, \frac{1}{2}\right)$$

is the only saddle point of \mathcal{L} with respect to $A \times B$ and $\mathcal{L}(p_0, q_0) = 0$.

As expected, the optimal mixed strategy for each player consists in choosing heads and tails with equal probability. Therefore, this game is fair.

2.15 The Ekeland Principle about Quasi-Minimal Points

Our point of departure is the following minimum problem:

$$F(u) = \min !, \quad u \in M. \quad (52)$$

We want to prove the existence of quasi-solutions. To this end, we consider the following *regularized problem*:

$$F(w) + \varepsilon \lambda^{-1} \|u - w\| = \min !, \quad w \in M, \quad (53)$$

for fixed numbers $\varepsilon > 0$ and $\lambda > 0$.

Definition 1. Each solution $u \in M$ of (53) is called an ε -quasi-solution of (52), that is,

$$F(u) < F(w) + \varepsilon \lambda^{-1} \|u - w\| \quad \text{for all } w \in M \text{ with } w \neq u. \quad (54)$$

Obviously, each solution u of the original problem (52) is also an ε -quasi-solution.

We assume that

- (H1) M is a nonempty closed subset of the real Banach space X .
- (H2) The functional $F: M \rightarrow \mathbb{R}$ is lower semicontinuous and $\inf_{w \in M} F(w) > -\infty$.
- (H3) For given numbers $\varepsilon > 0$ and $\lambda > 0$, we choose a point $v \in M$ such that

$$F(v) \leq \inf_{w \in M} F(w) + \varepsilon.$$

Theorem 2.H. *There exists an ε -quasi-solution $u \in M$ of (52) such that*

$$\|u - v\| \leq \lambda \quad \text{and} \quad F(u) \leq F(v).$$

This theorem was proved by Ekeland in 1974.

Corollary 2. *Suppose that the functional $F: X \rightarrow \mathbb{R}$ is lower semicontinuous on the real Banach space X . Moreover, suppose that the Gâteaux derivative $F'(u)$ exists for each $u \in X$ and $\inf_{u \in X} F(u) > -\infty$.*

Then, for each $\varepsilon > 0$, there exists a point $u \in X$ such that

$$F(u) \leq \inf_{w \in X} F(w) + \varepsilon \quad \text{and} \quad \|F'(u)\| \leq \varepsilon.$$

In the next two sections (2.16 and 2.17) we will show that the existence of quasi-solutions implies the *existence* of solutions provided the *Palais–Smale condition*, which represents some compactness condition, is satisfied.

Proof of Corollary 2. By Theorem 2.H with $\lambda = 1$, there exists a $u \in X$ such that

$$F(u) \leq F(w) + \varepsilon \|u - w\| \quad \text{for all } w \in X.$$

We choose $w = u + tv$, where $t \in \mathbb{R}$ with $t \neq 0$ and $v \in X$. Then

$$t^{-1}(F(u + tv) - F(u)) \geq -\varepsilon \|v\|.$$

As $t \rightarrow 0$, we obtain $\langle F'(u), v \rangle \geq -\varepsilon \|v\|$. That is, $\langle F'(u), z \rangle \leq \pm\varepsilon \|z\|$ for all $z \in X$. Thus $\|F'(u)\| \leq \varepsilon$. \square

Proof of Theorem 2.H. It suffices to assume that $\lambda = 1$, since we can pass from $\|\cdot\|$ to $\lambda^{-1}\|\cdot\|$. We inductively define a sequence (u_n) for $n = 0, 1, \dots$. Let $u_0 := v$. If we know $u_n \in M$, then we construct $u_{n+1} \in M$ as follows.

Case 1: $F(w) > F(u_n) - \varepsilon \|u_n - w\|$ for all $w \in M$. Then, let $u_{n+1} := u_n$. Hence

$$F(u_{n+1}) = F(u_n).$$

Case 2: $F(w) \leq F(u_n) - \varepsilon \|u_n - w\|$ for some $w \in M$. Let S_n be the set of all these points w , and let $\alpha_n := \inf_{w \in S_n} F(w)$. We then choose a point $u_{n+1} \in S_n$ with

$$F(u_{n+1}) \leq \alpha_n + 2^{-1}(F(u_n) - \alpha_n).$$

This is possible since $F(u_n) \geq \alpha_n + \varepsilon \|u_n - w\|$. It follows from $u_{n+1} \in S_n$ that

$$F(u_{n+1}) \leq F(u_n) - \varepsilon \|u_n - u_{n+1}\| \leq F(u_n).$$

Our construction is so constituted that all the $F(u_n)$ form a nonincreasing sequence, which by (H2) is bounded below and hence *convergent*. Let $F(u_n) \rightarrow \beta$ as $n \rightarrow \infty$. We will show that the sequence (u_n) is convergent. In fact, by construction,

$$\varepsilon \|u_n - u_{n+1}\| \leq F(u_n) - F(u_{n+1}) \quad \text{for all } n.$$

Using the triangle inequality, addition yields

$$\varepsilon \|u_n - u_m\| \leq F(u_n) - F(u_m) \quad \text{for all } m > n. \quad (55)$$

Therefore, (u_n) is a *Cauchy sequence* and thus a convergent sequence. Let $u_n \rightarrow u$ as $n \rightarrow \infty$. Hence $u \in M$ and

$$F(u) \leq \lim_{n \rightarrow \infty} F(u_n), \quad (56)$$

since F is weakly sequentially lower semicontinuous, by Lemma 5 in Section 2.5. We shall show that the point u has all the desired properties.

Proof of $F(u) \leq F(v)$. From (56) and $F(u_n) \leq F(u_0)$ for all n , it follows that $F(u) \leq F(u_0)$. Note that $u_0 = v$.

Proof of $\|v - u\| \leq 1$. For $n = 0$ and $m \rightarrow \infty$, it follows from (55) and (56) that

$$\varepsilon \|v - u\| \leq F(v) - F(u) \leq F(v) - \inf_{w \in M} F(w) \leq \varepsilon.$$

Proof of (54). On the contrary, suppose that (54) is false. Then, there exists a point $w \in M$ with $w \neq u$ such that

$$\varepsilon \|w - u\| + F(w) \leq F(u). \quad (57)$$

As $m \rightarrow \infty$, from (55) and (56) we obtain

$$\varepsilon \|u_n - u\| + F(u) \leq F(u_n) \quad \text{for all } n.$$

The triangle inequality yields

$$\varepsilon \|w - u_n\| + F(w) \leq F(u_n) \quad \text{for all } n.$$

By Case 2, $w \in S_n$ for all n . Hence

$$2F(u_{n+1}) - F(u_n) \leq \alpha_n \leq F(w) \quad \text{for all } n.$$

This implies $\beta \leq F(w)$, since $F(u_n) \rightarrow \beta$ as $n \rightarrow \infty$. From (56) it follows that $F(u) \leq \beta$. Thus, $F(u) \leq F(w)$. By (57), $F(w) < F(u)$. This is a contradiction. \square

2.16 Applications to a General Minimum Principle via the Palais–Smale Condition

Together with the minimum problem

$$F(u) = \min !, \quad u \in X, \quad (58)$$

we consider the operator equation

$$F'(u) = 0, \quad u \in X. \quad (59)$$

The following existence theorem is based on an important *compactness property* of functionals, which we shall first define.

Definition 1. Suppose that the functional $F: X \rightarrow \mathbb{R}$ has a Gâteaux derivative $F'(u)$ for each point $u \in X$, where X is a Banach space. Then, F satisfies the *Palais–Smale condition* (PS) iff the following holds:

If u_n is a sequence in X with these two properties:

- (i) $(F(u_n))$ is bounded, and
- (ii) $\|F'(u_n)\| \rightarrow 0$ as $n \rightarrow \infty$,

then (u_n) has a convergent subsequence.

Theorem 2.I. Let $F: X \rightarrow \mathbb{R}$ be a functional on the real Banach space X such that the following hold:

- (i) *The Gâteaux derivative $F'(u)$ exists for each $u \in X$,*
- (ii) *The functional F is lower semicontinuous (e.g., continuous), is bounded below on X , and satisfies the Palais–Smale condition (PS).*

Then, the minimum problem (58) has a solution u , which also satisfies the operator equation (59).

Proof. Let $\alpha := \inf_{u \in X} F(u)$. Then $\alpha > -\infty$. According to Corollary 1 in Section 2.15, there exists a sequence (u_n) in X such that

$$F(u_n) \rightarrow \alpha \quad \text{and} \quad \|F'(u_n)\| \rightarrow 0.$$

The condition (PS) ensures the existence of a subsequence, again denoted by (u_n) , such that $u_n \rightarrow u$. By Lemma 5 in Section 2.5, F is weakly sequentially lower semicontinuous on X . Hence

$$F(u) \leq \lim_{n \rightarrow \infty} F(u_n).$$

Thus, we get $F(u) = \alpha$, so that u is a solution of (58). By Theorem 2.E, $F'(u) = 0$. \square

2.17 Applications to the Mountain Pass Theorem

We assume that

- (H1) The functional $H: Y \rightarrow \mathbb{R}$ is C^1 on the real Banach space Y , and H satisfies the Palais–Smale condition (PS).
- (H2) There exist positive constants R and α such that

$$H(y) \geq \alpha \quad \text{for all } y \in Y \text{ with } \|y\| = R.$$

- (H3) $H(0) < \alpha$.
- (H4) There exists a point $y_1 \in Y$ such that

$$\|y_1\| > R \quad \text{and} \quad H(y_1) < \alpha.$$

- (H5) We denote by M the set of all continuous functions $p: [0, 1] \rightarrow Y$ with $p(0) = 0$ and $p(1) = y_1$.
- Furthermore, we set

$$c := \inf_{p \in M} \sup_{0 \leq t \leq 1} H(p(t)). \tag{60}$$

Let us discuss the intuitive meaning of this situation. If $Y := \mathbb{R}^2$, then we can think of $H(y)$ as the *height* of a mountain landscape at the point y . We shall designate the points y with $\|y\| = R$ as a *mountain chain* \mathcal{C} . Then, by (H3) and (H4), valleys occur at the points $y = 0$ and $y = y_1$.

To each $p(\cdot)$ there corresponds a *path* that connects the two valleys over the mountain chain \mathcal{C} . Intuitively, one now expects that there exists a *saddle point* of our landscape at height c . The following theorem justifies this expectation.

Theorem 2.J. *If (H1) through (H5) hold, then the functional H possesses a critical point $y \in Y$, that is,*

$$H'(y) = 0.$$

In addition, $H(y) = c$ and $c \geq \alpha$.

This theorem was proved by Ambrosetti and Rabinowitz in 1973. Numerous interesting applications of this important theorem to periodic solutions of Hamiltonian systems and to nonlinear partial differential equations can be found in Rabinowitz (1986), Mawhin and Willem (1987), Ekeland (1990), and Struwe (1990). In particular, applications to the famous N -body problem in celestial mechanics are contained in Ambrosetti and Coti-Celati (1993) (cf. Problem 2.9).

Proof. We want to use Theorem 2.H. To this end, let X denote the set of all continuous functions $p: [0, 1] \rightarrow Y$. Then, X becomes a real Banach space equipped with the norm

$$\|p\|_X := \max_{0 \leq t \leq 1} \|p(t)\|_Y$$

(use the same argument as in the proof of Standard Example 6 in Section 1.3 of AMS Vol. 108). Define the functional $F: X \rightarrow \mathbb{R}$ through

$$F(p) := \max_{0 \leq t \leq 1} H(p(t)).$$

This definition makes sense, since the functional $H: Y \rightarrow \mathbb{R}$ is *continuous*, by (H1). Thus, the continuous function $t \mapsto H(p(t))$ attains its maximum on the compact set $[0, 1]$.

Our goal is the investigation of the following minimum problem:

$$F(p) = \min !, \quad p \in M. \tag{61}$$

By (60), $c = \inf_{p \in M} F(p)$.

We also set $d := \max\{H(0), H(y_1)\}$. From (H2) we get

$$c \geq \alpha > d. \tag{62}$$

Step 1: Existence of a quasi-solution via Theorem 2.H. Choose $\varepsilon > 0$ such that

$$0 < \varepsilon < c - d. \quad (63)$$

Since c denotes the infimum of F on the set M , there exists a $v \in M$ such that

$$c \leq F(v) \leq c + \varepsilon.$$

Letting $\lambda := \varepsilon^{\frac{1}{2}}$, it follows from Theorem 2.H in Section 2.14 that there exists a point $u \in M$ such that

$$F(u) \leq F(v) \quad \text{and} \quad \|u - v\|_X \leq \varepsilon^{\frac{1}{2}}, \quad (64)$$

as well as

$$F(u) < F(w) + \varepsilon^{\frac{1}{2}} \|u - w\|_X, \quad (65)$$

for all $w \in M$ and $w \neq u$.

Step 2: We shall show ahead that (64) and (65) imply the existence of a number $s \in [0, 1]$ such that

$$/ \quad \|H'(u(s))\| \leq \varepsilon^{\frac{1}{2}} \quad (66)$$

and

$$c - \varepsilon \leq H(u(s)). \quad (67)$$

It follows from $H(u(s)) \leq F(u) \leq F(v) \leq c + \varepsilon$ that

$$c - \varepsilon \leq H(u(s)) \leq c + \varepsilon. \quad (67^*)$$

Step 3: Existence of a solution via the *Palais–Smale condition* (PS). Set $\varepsilon := \frac{1}{n}$ and $y_n := u(s)$. Then, it follows from (66) and (67*) that

$$\|H'(y_n)\| \leq n^{-\frac{1}{2}}, \quad c - n^{-1} \leq H(y_n) \leq c + n^{-1}, \quad (68)$$

for sufficiently large $n \geq n_0$. By (PS), there exists a subsequence, again denoted by (y_n) , such that $y_n \rightarrow y$ in Y as $n \rightarrow \infty$. Thus, it follows from (68) that

$$\|H'(y)\| = 0 \quad \text{and} \quad H(y) = c.$$

This is the desired result.

Step 4: It remains to prove the assertion from Step 2. Let $u \in M$ be given as in Step 1. Recall that, by the definition of M , the path $u: [0, 1] \rightarrow Y$ is continuous along with $u(0) = 0$ and $u(1) = y_1$. Observe that

$$\|H'(u(s))\| = \sup_{\|y\|_Y=1} \langle H'(u(s)), y \rangle.$$

Furthermore, let us introduce the set

$$S := \{s \in [0, 1]: c - \varepsilon \leq H(u(s))\}.$$

By (63), $d < c - \varepsilon$. Since $u(0) = 0$ and $H(0) \leq d$, we get $0 \notin S$. Moreover, since the functional H is continuous on Y , the set S is closed, and hence S is *compact*.

Suppose that the assertion from Step 2 is *not* true. Then,

$$\|H'(u(s))\| > \varepsilon^{\frac{1}{2}} \quad \text{for all } s \in S.$$

Thus, for each $s \in S$, there exists a point $y_s \in Y$ with $\|y_s\|_Y = 1$ such that

$$\langle H'(u(s)), -y_s \rangle > \varepsilon^{\frac{1}{2}}. \quad (69)$$

We want to construct a special path $p: [0, 1] \rightarrow Y$ with $p \in M$ such that $p \neq u$ and

$$F(u) \geq F(p) + \varepsilon^{\frac{1}{2}} \|u - p\|_X. \quad (70)$$

This is the desired *contradiction* to (65).

Step 5: Construction of p . Since H' is continuous on Y , it follows from (69) that for each $s \in S$, there exist a number $\beta_s > 0$ and an open interval J_s in \mathbb{R} such that⁸

$$\langle H'(u(t) + h), -y_s \rangle > \varepsilon^{\frac{1}{2}}, \quad (71)$$

for all $t \in J_s$ and all $h \in Y$ with $\|h\| \leq \beta_s$.

Obviously, the family $\{J_s\}_{s \in S}$ of open intervals J_s covers the *compact* set S . Therefore, a *finite* subfamily $\{J_{s_1}, \dots, J_{s_m}\}$ already covers the set S . For brevity of notation, we set

$$J_k := J_{s_k}.$$

Since $0 \notin S$, we may assume that $0 \notin J_k$, and hence

})

$[0, 1] - J_k$ is closed and not empty for all $k = 1, \dots, m$.

Therefore, if $t \in \bigcup_{k=1}^m J_k$, then

$$\sum_{k=1}^m \text{dist}(t, [0, 1] - J_k) > 0.$$

Define

$$\psi(t) := \begin{cases} 1 & \text{if } c \leq H(u(t)) \\ 0 & \text{if } H(u(t)) \leq c - \varepsilon. \end{cases}$$

Since the function $t \mapsto H(u(t))$ is continuous on $[0, 1]$, the function ψ is defined on two disjoint closed subsets of $[0, 1]$. Hence, it can be extended

⁸Observe that

$$\Delta := |\langle H'(y) - H'(z), y_s \rangle| \leq \|H'(y) - H'(z)\| \|y_s\| \quad \text{for all } y, z \in Y.$$

Consequently, Δ is sufficiently small provided $\|y - z\|_Y$ is sufficiently small.

to a continuous function $\psi: [0, 1] \rightarrow [0, 1]$, by the Tietze–Urysohn extension theorem. Furthermore, for $j = 1, \dots, m$, we define the function $\psi_j: [0, 1] \rightarrow \mathbb{R}$ through

$$\psi_j(t) := \begin{cases} \frac{\text{dist}(t, [0, 1] - J_j)}{\sum_{k=1}^m \text{dist}(t, [0, 1] - J_k)} & \text{if } t \in \bigcup_{k=1}^m J_k \\ 0 & \text{otherwise.} \end{cases}$$

One checks easily that ψ_j is *continuous* on $[0, 1]$. Furthermore, we get

$$\sum_{j=1}^m \psi_j(t) \leq 1 \quad \text{for all } t \in [0, 1]$$

and $\psi_j(t) = 0$ if $t \notin J_j$.

Finally, we introduce the *continuous* function $p: [0, 1] \rightarrow \mathbb{R}$ through

$$| \quad p(t) := u(t) + \beta \psi(t) \sum_{j=1}^m \psi_j(t) y_{s_j}, \quad (72)$$

where $\beta := \min\{\beta_{s_1}, \dots, \beta_{s_m}\}$.

Let us investigate the properties of the *special path* p from (72). We first prove that $p \in M$. In fact, since $H(0) \leq d$, $H(y_1) \leq d$, and $d < c - \varepsilon$, we get $\psi(0) = \psi(1) = 0$. Hence

$$p(0) = u(0) = 0 \quad \text{and} \quad p(1) = u(1) = y_1,$$

that is, $p \in M$.

Next let us prove inequality (70). To this end, set

$$\phi(\tau) := H(u(t) + \tau[p(t) - u(t)]), \quad \tau \in \mathbb{R}.$$

By the classical mean value theorem, there is some $\tau_0 \in]0, 1[$ such that

$$\phi(1) - \phi(0) = \phi'(\tau_0).$$

This is identical to

$$H(p(t)) - H(u(t)) = \langle H'(u(t) + \tau_0[p(t) - u(t)]), p(t) - u(t) \rangle. \quad (73)$$

Note that

$$p(t) - u(t) = \beta \psi(t) \sum_{j=1}^m \psi_j(t) y_{s_j}, \quad \text{for all } t \in [0, 1].$$

Because $\|y_{s_j}\| = 1$, we have

$$\|p(t) - u(t)\|_Y \leq \beta \sum_{j=1}^m \psi_j(t) \|y_{s_j}\| \leq \beta \leq \beta_{s_k}, \quad (74)$$

for all $k = 1, \dots, m$ and $t \in [0, 1]$. Since $0 < \tau_0 < 1$, this implies

$$\|\tau_0(p(t) - u(t))\|_Y \leq \beta_{s_k} \quad \text{for all } k = 1, \dots, m \text{ and } t \in [0, 1].$$

Let $t \in S$. Then, $t \in J_k$ for some k . Thus, it follows from (71) and (73) that⁹

$$\begin{aligned} H(p(t)) - H(u(t)) &= \beta\psi(t) \sum_{j=1}^m \psi_j(t) \langle H'(u(t) + \tau_0[p(t) - u(t)]), y_{s_j} \rangle \\ &\leq \beta\psi(t) \sum_{j=1}^m \psi_j(t) (-\varepsilon^{\frac{1}{2}}). \end{aligned}$$

Hence we get the following *key inequality*:

$$H(p(t)) - H(u(t)) \leq -\varepsilon^{\frac{1}{2}}\beta\psi(t) \quad \text{for all } t \in S. \quad (75)$$

If $t \in [0, 1] - S$, then $H(u(t)) < c - \varepsilon$, and hence $\psi(t) = 0$. Thus, $p(t) = u(t)$, which implies

$$H(p(t)) - H(u(t)) = 0 \quad \text{for all } t \in [0, 1] - S. \quad (76)$$

Hence

$$H(u(t)) \geq H(p(t)) \quad \text{for all } t \in [0, 1]. \quad (77)$$

Choose the number $\sigma \in [0, 1]$ in such a way that

$$H(p(\sigma)) = \max_{0 \leq t \leq 1} H(p(t)).$$

By the definition of F ,

$$H(p(\sigma)) = F(p).$$

Since $p \in M$ and $c = \inf_{q \in M} F(q)$, we get $F(q) \geq c$. By (77),

$$H(u(\sigma)) \geq H(p(\sigma)) = F(p) \geq c. \quad (78)$$

Hence $\sigma \in S$ and $\psi(\sigma) = 1$. It follows from (75) with $t := \sigma$ that

$$H(u(\sigma)) \geq \varepsilon^{\frac{1}{2}}\beta + H(p(\sigma)).$$

By the definition of F , we get $F(u) \geq H(u(\sigma))$, and hence

$$F(u) \geq \varepsilon^{\frac{1}{2}}\beta + F(p) > F(p), \quad (79)$$

which yields $p \neq u$ in X . According to (74),

$$\|p - u\|_X = \max_{0 \leq t \leq 1} \|p(t) - u(t)\|_Y \leq \beta.$$

Thus, inequality (79) tells us that

$$F(u) \geq \varepsilon^{\frac{1}{2}}\|p - u\|_X + F(p).$$

This is the desired inequality (70). \square

⁹Observe that $\psi_j(t) = 0$ for $t \notin J_j$ and $\psi_j(t) > 0$ for $t \in J_j$.

2.18 The Galerkin Method and Nonlinear Monotone Operators

We want to solve the operator equation

$$Au = b, \quad u \in X. \quad (80)$$

To this end, we assume that

- (H1) The operator $A: X \rightarrow X^*$ is *monotone* on the real, separable, reflexive Banach space X , that is,

$$\langle Au - Av, u - v \rangle \geq 0 \quad \text{for all } u, v \in X.$$

- (H2) The operator A is continuous on each finite-dimensional subspace of the Banach space X .

- (H3) The operator A is *coercive*, that is,

$$\lim_{\|u\| \rightarrow \infty} \frac{\langle Au, u \rangle}{\|u\|} = +\infty.$$

Theorem 2.K. *For each given $b \in X^*$, the original equation (80) has a solution u .*

Corollary 1. *The solution u of (80) is unique provided the operator A is strictly monotone, that is,*

$$\langle Au - Av, u - v \rangle > 0 \quad \text{for all } u, v \in X \text{ with } u \neq v. \quad (81)$$

In fact, if $Au = Av = b$, then (81) implies $u = v$. This yields Corollary 1.

Theorem 2.K is called the main theorem on monotone operators. This famous result was proved independently by Browder and Minty in 1963. Important applications to partial differential equations and integral equations can be found in Zeidler (1986), Vol. 2B.

The proof of Theorem 2.K will be based on the following three lemmas. First let us investigate the real system

$$g_j(x) = 0, \quad x \in \mathbb{R}^N, \quad j = 1, \dots, N. \quad (82)$$

Here, we set $x := (\xi_1, \dots, \xi_N)$ and $B := \{x \in \mathbb{R}^N : \|x\| \leq R\}$ for fixed $R > 0$, where $\|\cdot\|$ denotes a given norm on \mathbb{R}^N , $N \geq 1$.

Lemma 2 (Existence principle). *Suppose that*

- (i) The function $g_j: B \rightarrow \mathbb{R}$ is continuous for each $j = 1, \dots, N$.
- (ii) For all $x \in \mathbb{R}^N$ with $\|x\| = R$,

$$\sum_{j=1}^N g_j(x) \xi_j \geq 0.$$

Then, equation (82) has a solution $x \in B$.

Proof. We will use the *Brouwer fixed-point theorem* from Section 1.14 in AMS Vol. 108. Set $g(x) := (g_1(x), \dots, g_N(x))$ and suppose that $g(x) \neq 0$ for all $x \in B$. Let

$$f(x) := -\frac{Rg(x)}{\|g(x)\|} \quad \text{for all } x \in B.$$

The function $f: B \rightarrow \mathbb{R}^N$ is continuous on the compact convex set B . In addition,

$$\|f(x)\| = R \quad \text{for all } x \in B. \quad (83)$$

Thus, $f(B) \subseteq B$. By the *Brouwer fixed-point theorem*, the map f has a fixed point x :

$$f(x) = x, \quad x \in B.$$

By (83), $\|x\| = R$. Furthermore,

$$\begin{aligned} \sum_{j=1}^N g_j(x) \xi_j &= -R^{-1} \|g(x)\| \sum_{j=1}^N f_j(x) \xi_j \\ &= -R^{-1} \|g(x)\| \sum_{j=1}^N \xi_j^2 < 0. \end{aligned}$$

This contradicts assumption (ii). □

Lemma 3 (The monotonicity trick). *Assume (H1) and (H2). Let $b \in X^*$. Then it follows from*

$$u_n \rightharpoonup u \quad \text{in } X \quad \text{as } n \rightarrow \infty,$$

$$Au_n \rightharpoonup b \quad \text{in } X^* \quad \text{as } n \rightarrow \infty,$$

and $\langle Au_n, u_n \rangle \rightarrow \langle b, u \rangle$ as $n \rightarrow \infty$ that $Au = b$.

Proof. By the *monotonicity* of A ,

$$\langle Au_n, u_n \rangle - \langle Av, u_n \rangle - \langle Au_n - Av, v \rangle = \langle Au_n - Av, u_n - v \rangle \geq 0,$$

for all $v \in X$ and all n . Letting $n \rightarrow \infty$, we get¹⁰

$$\langle b, u \rangle - \langle Av, u \rangle - \langle b - Au, v \rangle \geq 0,$$

and hence

$$\langle b - Av, u - v \rangle \geq 0 \quad \text{for all } v \in X. \quad (84)$$

Next let $v := u - tw$, where $t > 0$ and $w \in X$. Then, relation (84) implies

$$\langle b - A(u - tw), w \rangle \geq 0 \quad \text{for all } w \in X.$$

Letting $t \rightarrow +0$, from (H2) we get

$$\langle b - Au, w \rangle \geq 0 \quad \text{for all } w \in X.$$

Replacing w with $-w$, this implies $\langle b - Au, w \rangle = 0$ for all $w \in X$ (i.e., $b - Au = 0$). \square

Lemma 4 (Local boundedness). *Assume (H1). Then, the operator $A: X \rightarrow X^*$ is locally bounded, that is, for each point $u \in X$, there exist numbers $r > 0$ and $\delta > 0$ such that*

$$\|Av\| \leq \delta \quad \text{for all } v \in X \text{ with } \|v - u\| \leq r.$$

Proof. Assume that A is *not* locally bounded. Then, there exist a point $u \in X$ and a sequence (u_n) with

$$u_n \rightarrow u \quad \text{and} \quad \|Au_n\| \rightarrow \infty \quad \text{as } n \rightarrow \infty.$$

Without loss of generality, we may assume that $u = 0$. We set

$$a_n := (1 + \|Au_n\| \|u_n\|)^{-1}.$$

It follows from the *monotonicity* of the operator A that

$$\langle Au_n - A(\mp v), u_n - (\mp v) \rangle \geq 0,$$

and hence

$$\begin{aligned} \pm a_n \langle Au_n, v \rangle &\leq a_n (\langle Au_n, u_n \rangle - \langle A(\pm v), u_n \mp v \rangle) \\ &\leq a_n (\|Au_n\| \|u_n\| + \|A(\pm v)\| \|u_n \mp v\|). \end{aligned}$$

¹⁰By Problem 3.5, $Au_n \rightharpoonup b$ in X^* as $n \rightarrow \infty$ implies

$$\langle Au_n, v \rangle \rightarrow \langle b, v \rangle \quad \text{as } n \rightarrow \infty \text{ for all } v \in X.$$

Since $a_n \|Au_n\| \|u_n\| \leq 1$ and $a_n \leq 1$ for all n , we get

$$\sup_n |a_n \langle Au_n, v \rangle| < \infty \quad \text{for all } v \in X.$$

According to the *Banach–Steinhaus theorem* from Section 3.3, there exists a number N such that

$$\sup_n \|a_n Au_n\| \leq N.$$

We set $b_n := \|Au_n\|$. Then,

$$b_n \leq a_n^{-1} N = (1 + b_n \|u_n\|)N \quad \text{for all } n.$$

Since $\|u_n\| \rightarrow 0$ as $n \rightarrow \infty$, the sequence (b_n) is bounded.¹¹ This is a contradiction to $\|Au_n\| \rightarrow \infty$ as $n \rightarrow \infty$.

Proof of Theorem 2.K. We will use the *Galerkin method*, which generalizes the Ritz method from Section 2.6 of AMS Vol. 108. The basic idea is to replace the original operator equation $Au = b$ by the upcoming finite-dimensional approximate equation (85) and to prove the convergence of this approximation method.

If $X = \{0\}$, then the assertion of Theorem 2.K is trivial. Therefore, let $X \neq \{0\}$.

Step 1: The Galerkin equations. Since the Banach space X is *separable*, there exists a countable set $\{x_1, x_2, \dots\}$ that is dense in X . Set

$$X_n := \text{span}\{x_1, \dots, x_n\}.$$

Thus, we get $X_1 \subseteq X_2 \subseteq \dots$ and

$$\bigcup_{n=1}^{\infty} X_n \quad \text{is dense in } X.$$

By definition, the n th *Galerkin equation* reads as follows:

$$\langle Au_n - b, v \rangle = 0 \quad \text{for fixed } u_n \in X_n \text{ and all } v \in X_n. \quad (85)$$

Step 2: Solution of the Galerkin equation. Let $\{e_1, \dots, e_N\}$ be a basis of X_n , and let

$$v := \sum_{j=1}^N \xi_j e_j, \quad u_n := \sum_{j=1}^N \xi_{nj} e_j,$$

as well as

$$x := (\xi_1, \dots, \xi_N), \quad \|x\| := \|v\|, \quad \text{and} \quad x_n := (\xi_{n1}, \dots, \xi_{nN}).$$

¹¹In fact, there is some n_0 such that $N\|u_n\| \leq 2^{-1}$ for all $n \geq n_0$. Hence $2^{-1}b_n \leq N$ for all $n \geq n_0$.

Finally, we define the real functions

$$g_j(x_n) := \langle Au_n - b, e_j \rangle = 0, \quad j = 1, \dots, N.$$

Therefore, the Galerkin equation (85) is equivalent to the following system:

$$g_j(x_n) = 0, \quad x_n \in \mathbb{R}^N, \quad j = 1, \dots, N. \quad (86)$$

By (H2), the function $g_j: \mathbb{R}^N \rightarrow \mathbb{R}$ is continuous for each $j = 1, \dots, N$. Furthermore, there exists a number $R > 0$ such that

$$\sum_{j=1}^N g_j(x) \xi_n \geq 0 \quad \text{for all } x \in \mathbb{R}^N \text{ with } \|x\| = R.$$

In fact, we have

$$\sum_{j=1}^N g_j(x) \xi_j = \langle Av - b, v \rangle.$$

By (H3), $\frac{\langle Av, v \rangle}{\|v\|} \rightarrow +\infty$ as $\|v\| \rightarrow \infty$. Hence

$$\begin{aligned} \langle Av - b, v \rangle &= \langle Av, v \rangle - \langle b, v \rangle \\ &\geq \langle Av, v \rangle - \|b\| \|v\| \geq 0 \end{aligned}$$

for all $v \in X$ with $\|v\| = R$ and fixed sufficiently large $R > 0$.

According to Lemma 2, system (86) has a solution x_n with $\|x_n\| \leq R$. Consequently, the Galerkin equation (85) has a solution u_n , where

$$\|u_n\| \leq R \quad \text{for all } n. \quad (87)$$

Step 3: Boundedness of the sequence (Au_n) . By Lemma 4, the operator A is locally bounded, that is, there exist positive numbers r and δ such that

$$\|v\| \leq r \quad \text{implies} \quad \|Av\| \leq \delta.$$

The operator A is monotone. Thus,

$$\langle Au_n - Av, u_n - v \rangle \geq 0.$$

By the Galerkin equation (85),

$$\langle Au_n, u_n \rangle = \langle b, u_n \rangle \quad \text{for all } n.$$

Hence

$$|\langle Au_n, u_n \rangle| \leq \|b\| \|u_n\| \leq \|b\| R \quad \text{for all } n.$$

By the definition of the norm in X^* and by the monotonicity of A ,

$$\begin{aligned} \|Au_n\| &= \sup_{\|v\|=r} r^{-1} \langle Au_n, v \rangle \\ &\leq \sup_{\|v\|=r} r^{-1} (\langle Av, v \rangle + \langle Au_n, u_n \rangle - \langle Av, u_n \rangle) \\ &\leq r^{-1} (\delta r + \|b\| R + \delta R). \end{aligned}$$

Step 4: Convergence of the Galerkin method. The Banach space X is *reflexive*. Thus, the bounded sequence (u_n) has a weakly convergent subsequence, again denoted by (u_n) , that is,

$$u_n \rightharpoonup u \quad \text{as } n \rightarrow \infty.$$

From the Galerkin equation (85) it follows that

$$\lim_{n \rightarrow \infty} \langle Au_n, w \rangle = \langle b, w \rangle \quad \text{for all } w \in \bigcup_{k=1}^{\infty} X_k. \quad (88)$$

Since the sequence (Au_n) is bounded and $\bigcup_{k=1}^{\infty} X_k$ is dense in X , we get

$$\lim_{n \rightarrow \infty} \langle Au_n, z \rangle = \langle b, z \rangle \quad \text{for all } z \in X. \quad (89)$$

In fact, for each $z \in X$ and given $\varepsilon > 0$, there is a $w \in \bigcup_{k=1}^{\infty} X_k$ such that $\|z - w\| < \varepsilon$. Hence

$$|\langle Au_n, z \rangle - \langle Au_n, w \rangle| \leq \sup_n \|Au_n\| \|z - w\| < \text{const} \cdot \varepsilon.$$

Thus, (88) implies (89). Since X is reflexive, relation (89) is equivalent to

$$Au_n \rightharpoonup b \quad \text{as } n \rightarrow \infty$$

(cf. Problem 3.5). Furthermore, it follows from the Galerkin equation (85) that

$$\lim_{n \rightarrow \infty} \langle Au_n, u_n \rangle = \lim_{n \rightarrow \infty} \langle b, u_n \rangle = \langle b, u \rangle.$$

Now, Lemma 3 tells us that $Au = b$. □

2.19 Symmetries and Conservation Laws (The Noether Theorem)

The following fundamental fact was discovered by Emmy Noether in 1918:

Conservation laws in nature are caused by symmetries of variational principles.

For example, as we will show here conservation of energy corresponds to invariance under time translation.

Let us start with the following fundamental *symmetry property* of the one-dimensional Lagrangian $L = L(x, u, u')$:

$$L(y, u(y, \varepsilon), u_y(y, \varepsilon))y_x(x, \varepsilon) = L(x, u(x), u'(x)) \quad (90)$$

for all $x \in]a, b[$ and $\varepsilon \in]-\varepsilon_0, \varepsilon_0[$, where the argument y has to be replaced with $y(x, \varepsilon)$. Here

$$y(x, 0) \equiv x, \quad u(x, 0) \equiv u(x).$$

We assume that

- (A1) The Lagrangian $L: \mathbb{R}^3 \rightarrow \mathbb{R}$ is C^2 . Let $-\infty \leq a < b \leq \infty$ and $\varepsilon_0 > 0$.
- (A2) The C^2 -function $u:]a, b[\rightarrow \mathbb{R}$ is a solution of the corresponding Euler–Lagrange equation

$$\frac{d}{dx} L_{u'} - L_u = 0 \quad \text{on }]a, b[. \quad (91)$$

- (A3) The symmetry relation (90) holds. Here, the function $y = y(x, \varepsilon)$ is C^2 on $]a, b[\times]-\varepsilon_0, \varepsilon_0[$. Moreover, the function $u = u(y, \varepsilon)$ is C^2 on some appropriate open set such that $(x, \varepsilon) \mapsto u(y(x, \varepsilon), \varepsilon)$ is C^2 on $]a, b[\times]-\varepsilon_0, \varepsilon_0[$. Set

$$\delta y(x) := y_\varepsilon(x, 0) \quad \text{and} \quad \delta u(x) := \frac{\partial}{\partial \varepsilon} u(y(x, \varepsilon), \varepsilon) \Big|_{\varepsilon=0}.$$

Proposition 1. *The function u satisfies the conservation law¹²*

$$\frac{d}{dx} (L\delta y + L_{u'}(\delta u - u'\delta y)) = 0 \quad \text{on }]a, b[. \quad (92)$$

Standard Example 2 (Conservation of energy). Suppose that the Lagrangian L does not depend on the real variable x . Letting

$$y = x + \varepsilon \quad \text{and} \quad u(y, \varepsilon) = u(x),$$

we get $\delta y(x) = 1$ and $\delta u(x) = 0$. Hence

$$\frac{d}{dx} (u' L_{u'} - L) = 0 \quad \text{on }]a, b[. \quad (93)$$

In mechanics this corresponds to conservation of the energy $E := u' L_{u'} - L$ provided we regard x as time. For example, if

$$L := \frac{1}{2} m u'^2 - V(u),$$

then the Euler equation (91) is identical to the equation of motion

$$m u'' = -V'(u),$$

¹²Explicitly, this means that

$$\frac{d}{dx} (L(P)\delta y(x) + L_{u'}(P)(\delta u(x) - u'(x)\delta y(x))) = 0 \quad \text{for all } x \in]a, b[,$$

where $P := (x, u(x), u'(x))$.

and (93) describes conservation of energy:

$$E := \frac{1}{2}mu'^2 + V(u) = \text{const.}$$

Proof of Proposition 1. Differentiating the symmetry relation (90) with respect to the parameter ε at $\varepsilon = 0$, we get

$$\begin{aligned} 0 &= L_x(P)y_\varepsilon(x, 0) + L_u(P)(u_x(x, 0)y_\varepsilon(x, 0) + u_\varepsilon(x, 0)) \\ &\quad + L_{u'}(P)(u_{xx}(x, 0)y_\varepsilon(x, 0) + u_{\varepsilon x}(x, 0)) + L(P)y_{\varepsilon x}(x, 0), \end{aligned}$$

where $P := (x, u(x), u'(x))$. Observing that $\delta y(x) = y_\varepsilon(x, 0)$, $u_x(x, 0) = u'(x)$ and letting $\phi(x) := u_\varepsilon(x, 0)$, we see that this is identical to

$$\begin{aligned} 0 &= L_x\delta y + L_u(u'\delta y + \phi) + L_{u'}(u''\delta y + \phi') + L(\delta y)' \\ &= L'\delta y + L_u\phi + L_{u'}\phi' + L(\delta y)'. \end{aligned}$$

Using the Euler equation (91), $L_u = (L_{u'})'$, we find that

$$0 = (L\delta y)' + (L_{u'}\phi)'. \quad (94)$$

Differentiating $u(y(x, \varepsilon), \varepsilon)$ with respect to ε at $\varepsilon = 0$, we obtain

$$\delta u(x) = u_x(x, 0)y_\varepsilon(x, 0) + u_\varepsilon(x, 0) = u'(x)\delta y(x) + \phi(x),$$

and hence $\phi = \delta u - u'\phi$. Thus, from (94) we get the assertion (92). \square

Now let us generalize this to multidimensional variational problems. Our point of departure is the following *symmetry relation* for the Lagrangian $L = L(x, u, \partial u)$:

$$L(y, u(y, \varepsilon), \partial_y u(y, \varepsilon)) \det \partial_x y(x, \varepsilon) = L(x, u(x), \partial_x u(x)) \quad (95)$$

for all $x \in G$ and $\varepsilon \in]-\varepsilon_0, \varepsilon_0[$, where the argument y has to be replaced with $y(x, \varepsilon)$. Here

$$y(x, 0) \equiv x, \quad u(x, 0) \equiv 0,$$

along with $u = (u_1, \dots, u_M)$, $x = (\xi_1, \dots, \xi_N)$, $y = (y_1, \dots, y_N)$, as well as $\partial_x u(x) = (\partial_1 u(x), \dots, \partial_N u(x))$, where $\partial_j := \partial/\partial \xi_j$. We assume that

(H1) G is a nonempty open set in \mathbb{R}^N . The Lagrangian $L: \mathbb{R}^N \times \mathbb{R}^M \times \mathbb{R}^{NM} \rightarrow \mathbb{R}$ is C^2 .

(H2) The C^2 -function $u: G \rightarrow \mathbb{R}^M$ is a solution of the corresponding Euler–Lagrange equation

$$\sum_{j=1}^N \partial_j L_{\partial_j u_m} - L_{u_m} = 0, \quad m = 1, \dots, M. \quad (96)$$

- (H3) The symmetry relation (95) holds. Here, the function $y = y(x, \varepsilon)$ is C^2 on $G \times]-\varepsilon_0, \varepsilon_0[$. Moreover, the function $u = u(y, \varepsilon)$ is C^2 on some appropriate open set such that $(x, \varepsilon) \mapsto u(y(x, \varepsilon), \varepsilon)$ is C^2 on $G \times]-\varepsilon_0, \varepsilon_0[$. Set

$$\delta y(x) := y_\varepsilon(x, 0) \quad \text{and} \quad \delta u(x) := \frac{\partial}{\partial \varepsilon} u(y(x, \varepsilon), \varepsilon) \Big|_{\varepsilon=0}.$$

Theorem 2.L (The Noether theorem). *Assume (H1) through (H3). Then, the function u satisfies the conservation law*

$$\sum_{j=1}^N \partial_j J_j = 0 \quad \text{on } G$$

with the current $J = (J_1, \dots, J_N)$, where

$$J_j := L\delta y_j + L_{\partial_j u_m}(\delta u_m - \partial_n u_m \delta y_n).$$

Here, we sum over $n = 1, \dots, N$ and $m = 1, \dots, M$.

This is one of the most important theorems in mathematical physics. The proof proceeds completely similarly to the proof of Proposition 1. Observe that

$$\frac{\partial}{\partial \varepsilon} \det \partial_x y(x, \varepsilon) \Big|_{\varepsilon=0} = 1 + \sum_{j=1}^N \partial_j (\delta y_j(x)).$$

This follows from

$$y(x, \varepsilon) = x + \varepsilon \delta y(x) + o(\varepsilon), \quad \varepsilon \rightarrow 0,$$

and hence

$$\partial_x y(x, \varepsilon) = I + \varepsilon \partial_x (\delta y(x)) + o(\varepsilon), \quad \varepsilon \rightarrow 0,$$

which implies

$$\det \partial_x y(x, \varepsilon) = 1 + \varepsilon \operatorname{tr} \partial_x (\delta y(x)) + o(\varepsilon), \quad \varepsilon \rightarrow 0.$$

Remark 3 (Interpretation of the symmetry relation). It follows from Section 2.2 that the Euler–Lagrange equation (96) represents a necessary condition for the following variational problem:

$$\int_G L(x, u(x), \partial u(x)) dx = \text{stationary !}, \quad u = \text{given} \quad \text{on } \partial G. \quad (97)$$

The decisive *symmetry relation* (95) means that the variational integral is invariant under the transformation $y = y(x, \varepsilon)$, that is,

$$\int_{G_\varepsilon} L(y, u(y, \varepsilon), \partial_y u(y, \varepsilon)) dy = \int_G L(x, u(x), \partial u(x)) dx \quad (98)$$

for all $\varepsilon \in]-\varepsilon_0, \varepsilon_0[$, where G is transformed into G_ε under $y = y(x, \varepsilon)$.

Remark 4. (The Noether theorem via local invariance of the variational integral). The following method is used frequently in applications. Let the variational problem (97) be given. Choose an arbitrary point $x \in G$. Suppose that we have the invariance condition

$$\int_{U_\varepsilon} L(y, u(y, \varepsilon), \partial_y u(y, \varepsilon)) dy = \int_U L(x, u(x), \partial u(x)) dx$$

for all $\varepsilon \in]-\varepsilon_0, \varepsilon_0[$, and for all sufficiently small neighborhoods U of the point x , where U is transformed into U_ε under $y = y(x, \varepsilon)$.

If we shrink U to the point x , then it follows from the mean value theorem for integrals that the crucial symmetry condition (95) is satisfied. Consequently, we can use the Noether theorem with respect to the transformation $y = y(x, \varepsilon)$.

Applications of the Noether theorem to gauge field theory and string theory can be found in Section 2.20 and Problem 2.15, respectively.

2.20 The Basic Ideas of Gauge Field Theory

In this section, all the functions are assumed to be sufficiently smooth.

In order to understand the basic ideas of gauge field theories in modern physics, let us first study a *simple model*. To this end, we consider the following variational problem:

$$\begin{aligned} \int_a^b (i\bar{\phi}\psi' - m\bar{\phi}\psi) dx &= \text{stationary!}, \\ \bar{\phi}(x), \psi(x) &= \text{given at the boundary points } x = a, b. \end{aligned} \quad (99)$$

Here, $\bar{\phi}(x)$ denotes the complex conjugate number to $\phi(x)$.

Proposition 1. *If $\bar{\phi}, \psi$ is a solution of (99), then*

$$i\psi' - m\psi = 0 \quad \text{on }]a, b[, \quad (100a)$$

$$i\bar{\phi}' + m\bar{\phi} = 0 \quad \text{on }]a, b[, \quad (100b)$$

and hence both ϕ and ψ satisfy equation (100a).

Proof. If we let $L := i\bar{\phi}\psi' - m\bar{\phi}\psi$, the Euler–Lagrange equations are given by

$$\frac{d}{dx}L_{\bar{\phi}'} - L_{\bar{\phi}} = 0 \quad \text{and} \quad \frac{d}{dx}L_{\psi'} - L_{\psi} = 0.$$

This yields (100). \square

Corollary 2. *If ψ is a solution of (100a), then ψ satisfies the following conservation law:*

$$(\bar{\psi}\psi)' = 0 \quad \text{on }]a, b[. \quad (100^*)$$

This means that the “density” $\bar{\psi}\psi$ is constant.

Proof. By (100), $i(\bar{\psi}\psi)' = i\bar{\psi}'\psi + i\bar{\psi}\psi' = -m\bar{\psi}\psi + m\bar{\psi}\psi = 0$. \square

Let $\alpha = \alpha(x)$ be a real function. The transformation

$$\psi_+(x) = e^{i\alpha(x)}\psi(x), \quad \phi_+(x) = e^{i\alpha(x)}\phi(x)$$

is called a (local) *gauge transformation*.¹³ If $\alpha = \text{const}$, then we speak of a *global gauge transformation*.

Corollary 3. *The Lagrangian $L = i\bar{\phi}\psi' - m\bar{\phi}\psi$ is invariant under global gauge transformations.*

Proof. Since $\psi_+ = e^{i\alpha}\psi$, $\psi'_+ = e^{i\alpha}\psi'$, and $\bar{\phi}_+ = e^{-i\alpha}\bar{\phi}$, we get

$$i\bar{\phi}_+\psi'_+ - m\bar{\phi}_+\psi_+ = i\bar{\phi}\psi' - m\bar{\phi}\psi. \quad \square$$

By Section 2.19, symmetries of the Lagrangian imply conservation laws. We want to show that

The global gauge invariance of the Lagrangian L yields the conservation law (100).*

To this end, set

$$\psi(x, \alpha) := e^{i\alpha}\psi(x), \quad \bar{\phi}(x, \alpha) := e^{-i\alpha}\bar{\phi}(x).$$

Then

$$\delta\psi(x) := \psi_\alpha(x, 0) = i\psi(x), \quad \delta\bar{\phi}(x) := \bar{\phi}_\alpha(x, 0) = -i\bar{\phi}(x).$$

Suppose that ϕ, ψ is a solution of the Euler equation (100) with $\phi = \psi$. By the Noether theorem (Theorem 2.L), it follows from the global gauge invariance

$$L(\bar{\phi}(x, \alpha), \psi(x, \alpha)) = L(\bar{\phi}(x, 0)\psi(x, 0)) \quad \text{for all } \alpha$$

¹³Gauge transformations are also called *phase transformations*.

that

$$(L_{\psi'} \delta\psi + L_{\bar{\phi}} \delta\bar{\phi})' = 0.$$

Since $L = i\bar{\phi}\psi' - m\bar{\phi}\psi$, this yields $(\bar{\phi}\phi)' = 0$. \square

In elementary particle physics, global gauge invariance implies the conservation of particle numbers or charges (e.g., the number of baryons or the electric charge).

Now to the point. The following principle is crucial.

Gauge field theories correspond to Lagrangians that are invariant under local gauge transformations (i.e., α depends on x).

In order to obtain such a Lagrangian, we have to modify L . The simplest possible ansatz consists in replacing the classical derivative d/dx by the so-called *covariant derivative*

$$\nabla := \frac{d}{dx} + i\kappa\mathcal{A},$$

where the real “field” $\mathcal{A} = \mathcal{A}(x)$ depends on x , and the positive number κ is called the coupling constant (of the interaction).

Proposition 4. *The modified Lagrangian*

$$L_{\mathcal{A}}(\bar{\phi}, \psi) := i\bar{\phi}\nabla\psi - m\bar{\phi}\psi$$

is invariant under gauge transformations provided the field \mathcal{A} is transformed into \mathcal{A}_+ , where

$$\mathcal{A}_+(x) := \mathcal{A}(x) - \kappa^{-1}\alpha'(x).$$

Proof. Letting

$$\nabla_+ := \frac{d}{dx} + i\kappa\mathcal{A}_+,$$

it follows from $\psi_+(x) = e^{i\alpha(x)}\psi(x)$ that

$$\nabla_+\psi_+ = e^{i\alpha}\nabla\psi,$$

that is, the covariant derivative ∇ transforms like the function ψ . This is the *key property* of ∇ . In fact,

$$\begin{aligned} \nabla_+\psi_+ &= (e^{i\alpha}\psi)' + i\kappa\mathcal{A}_+e^{i\alpha}\psi = e^{i\alpha}(i\alpha'\psi + \psi' + i\kappa\mathcal{A}\psi - i\alpha'\psi) \\ &= e^{i\alpha}(\psi' + i\kappa\mathcal{A}\psi) = e^{i\alpha}\nabla\psi. \end{aligned}$$

Consequently,

$$\begin{aligned} L_{\mathcal{A}_+}(\bar{\phi}_+, \psi_+) &= i\bar{\phi}_+\nabla_+\psi_+ - m\bar{\phi}_+\psi_+ \\ &= i\bar{\phi}e^{-i\alpha}e^{i\alpha}\nabla\psi - m\bar{\phi}e^{-i\alpha}e^{i\alpha}\psi = i\bar{\phi}\nabla\psi - m\bar{\phi}\psi \\ &= L_{\mathcal{A}}(\bar{\phi}, \psi). \end{aligned} \quad \square$$

Let us now replace the original variational problem (99) with the following gauge invariant problem:

$$\int_a^b (i\bar{\phi}\nabla\psi - m\bar{\phi}\psi)dx = \text{stationary}, \quad (101)$$

$\bar{\phi}(x), \psi(x)$ = given at the boundary points $x = a, b$.

That is, we replace the classical derivative d/dx with the covariant derivative ∇ .

Proposition 5. *If $\bar{\phi}, \psi$ is a solution of (101), then*

$$i\nabla\psi - m\psi = 0 \quad \text{on }]a, b[, \quad (102a)$$

$$i\nabla\phi - m\psi = 0 \quad \text{on }]a, b[. \quad (102b)$$

Proof. Letting $L := i\bar{\phi}\psi' + i\bar{\phi}i\kappa\mathcal{A}\psi - m\bar{\phi}\psi$, the Euler–Lagrange equations are given by

$$\frac{d}{dx}L_{\bar{\phi}'} - L_{\bar{\phi}} = 0 \quad \text{and} \quad \frac{d}{dx}L_{\psi'} - L_{\psi} = 0.$$

This yields (102a) and $i(\bar{\phi}' - i\kappa\mathcal{A}\bar{\phi}) + m\bar{\phi} = 0$, which is equivalent to (102b), since \mathcal{A} is real. \square

It follows from the gauge invariance of the Lagrangian L that

The gauge field equations (102) are invariant under gauge transformations.

Explicitly, this follows from

$$i\nabla_+\psi_+ - m\psi_+ = e^{i\alpha}(i\nabla\psi - m\psi).$$

Thus, if ψ is a solution of (102a), then ψ_+ is a solution of the transformed equation

$$i\nabla_+\psi_+ - m\psi_+ = 0.$$

The same is true for ϕ .

Remark 6 (Parallel transport). The function \mathcal{A} allows a geometrical interpretation. To explain this, let $C: x = x(t)$, $t_0 \leq t \leq t_1$ be a curve. Set

$$\frac{D}{dt} := x'(t)\nabla.$$

Then, by definition, the function $\psi = \psi(x)$ is *parallel along the curve C* iff

$$\frac{D\psi}{dt}(x(t)) = 0 \quad \text{on } [t_0, t_1].$$

Explicitly, this means that

$$x'(t)\psi'(x(t)) + i\kappa x'(t)\mathcal{A}(x(t))\psi(x(t)) = 0 \quad \text{on } [t_0, t_1].$$

In the special case where $\mathcal{A}(x) \equiv 0$, we obtain

$$\frac{d}{dt}\psi(x(t)) = x'(t)\psi'(x(t)) = 0 \quad \text{on } [t_0, t_1],$$

that is, ψ is constant along the curve C .

The function \mathcal{A} is called a *connection*. This interpretation of the field \mathcal{A} reflects an important relation between the gauge field theory of physicists and modern differential geometry of mathematicians based on parallel transport. This will be discussed in Remark 5 of Section 2.22.3.

Remark 7 (Gauge field theories in modern physics). In modern elementary particle physics, gauge field theories are used in order to describe mathematically the fundamental interactions in nature. In terms of our simple model given earlier, the basic ideas are roughly the following:

- (i) The functions ψ and $\bar{\phi}$ correspond to the *basic particles* \mathcal{P} and the antiparticles $\bar{\mathcal{P}}$, respectively.
- (ii) The postulate of gauge invariance forces the existence of a new field \mathcal{A} , which describes the *interaction* between the particles from (i).

For example, if ψ and $\bar{\phi}$ correspond to electrons and positrons, respectively, then the gauge field \mathcal{A} corresponds to the electromagnetic field, which is related to the photon after quantization. Thus, the existence of the photon (i.e., light) is a consequence of the postulate of *local gauge invariance*.

In the so-called *standard model* of modern elementary particle physics, there exist *six quarks* and *six leptons* (e.g., the electron and the neutrino) along with the corresponding antiparticles. According to the *principle of gauge invariance*, *local twelve particles* (gauge fields) describe the *interaction between quarks and leptons*, namely, the photon, three vector bosons (W^+ , W^- , Z^0) and eight gluons. The existence of the particles W^\pm , Z^0 was theoretically predicted by gauge field theory. These particles were detected experimentally in 1983 at CERN (Geneva, Switzerland).

The standard model corresponds to a gauge field theory with the gauge group $U(1) \times SU(2) \times SU(3)$. Both the quark model and $SU(N)$ -gauge field theories are based on the representations of Lie algebras. The basic ideas will be explained in the next three sections.

As an introduction to modern elementary particle physics from the physical point of view, we recommend the textbook by Rolnick (1994). A detailed study of the standard model can be found in the monograph by Donoghue et al. (1992). We also recommend the textbook by Sterman (1993) as an introduction both to quantum field theory and to the standard model from the physical point of view.

2.21 Representations of Lie Algebras

In the next section we will show that representations of the Lie algebra $\text{su}(3)$ play a fundamental role in elementary particle physics.

Definition 1. Let L be a linear space over \mathbb{K} . Then, L is called a *Lie algebra* over \mathbb{K} iff there exists a product $[A, B]$ on L that has the following properties:

- (i) To each given ordered pair (A, B) with $A, B \in L$, there is assigned exactly one element of L , which we call the Lie bracket $[A, B]$.
- (ii) For all $A, B, C \in L$ and $\alpha, \beta \in \mathbb{K}$,

$$\begin{aligned} [A, B] &= -[B, A] && \text{(anticommutativity),} \\ [A, [B, C]] + [B, [C, A]] + [C, [A, B]] &= 0 && \text{(Jacobi's identity),} \\ [\alpha A + \beta B, C] &= \alpha[A, C] + \beta[B, C] && \text{(distributive laws),} \\ [C, \alpha A + \beta B] &= \alpha[C, A] + \beta[C, B]. \end{aligned}$$

In addition, a linear subspace M of the Lie algebra L is called a *subalgebra* of L iff

$$A, B \in M \quad \text{implies} \quad [A, B] \in M.$$

This is equivalent to saying that M is a Lie algebra with respect to the Lie brackets of L .

A Lie algebra over \mathbb{R} (resp., \mathbb{C}) is also called a real (resp., complex) Lie algebra.

Standard Example 2 (Linear operators on Banach spaces). Let X be a Banach space over \mathbb{K} . Set

$$[A, B] := AB - BA \quad \text{for all } A, B \in L(X, X).$$

Then, the Banach space $L(X, X)$ of all linear continuous operators $A: X \rightarrow X$ is a Lie algebra over \mathbb{K} .

Proof. Note that $[A, B] = -[B, A]$ and

$$\begin{aligned} [A, [B, C]] + [B, [C, A]] + [C, [A, B]] \\ = (ABC - ACB - BCA + CBA) + (BCA - BAC - CAB + ACB) \\ + (CAB - CBA - ABC + BAC). \end{aligned} \quad \square$$

¹⁴Jacobi's identity replaces the missing associative law for the Lie product $[A, B]$.

Example 3 (The Lie algebra $\text{su}(N)$). Let X be an N -dimensional complex Hilbert space, where $N = 1, 2, \dots$. Then the following are true:

- (a) The set $\text{su}(N)$ of all traceless skew-adjoint linear operators $A: X \rightarrow X$ forms a real Lie algebra with respect to $[A, B] := AB - BA$.
- (b) $\text{su}(N)$ is a subalgebra of $L(X, X)$.
- (c) $\dim \text{su}(N) = N^2 - 1$.
- (d) $\text{su}(N)$ is a real Hilbert space with respect to the inner product

$$(A | B) := -\text{tr}(AB).$$

Proof. Ad (a). Let $A, B \in \text{su}(N)$ and $\alpha, \beta \in \mathbb{R}$. Then,

$$A^* = -A, \quad B^* = -B, \quad \text{and} \quad \text{tr } A = \text{tr } B = 0.$$

Hence

$$\begin{aligned} (\alpha A + \beta B)^* &= \alpha A^* + \beta B^* = -(\alpha A + \beta B), \\ \text{tr}(\alpha A + \beta B) &= \alpha \text{tr } A + \beta \text{tr } B = 0, \end{aligned}$$

that is, $\alpha A + \beta B \in \text{su}(N)$. Furthermore, it follows from

$$[A, B]^* = (AB - BA)^* = B^* A^* - A^* B^* = BA - AB = -[A, B]$$

and $\text{tr}[A, B] = \text{tr}(AB) - \text{tr}(BA) = 0$ that $[A, B] \in \text{su}(N)$.

Ad (b). First we want to show that there exists a one-to-one correspondence

$$A \simeq (a_{km})$$

between the operators $A \in \text{su}(N)$ and the traceless skew-adjoint $(N \times N)$ -matrices (a_{km}) , namely,

$$a_{11} + \cdots + a_{NN} = 0 \quad \text{and} \quad a_{km} = -\bar{a}_{mk} \quad \text{for all } k, m = 1, \dots, N.$$

In fact, let $\{e_1, \dots, e_N\}$ be an orthonormal basis of X . For $A \in \text{su}(N)$, set

$$a_{km} := (e_m | Ae_k), \quad k, m = 1, \dots, N.$$

Then

$$Ae_k = \sum_{m=1}^N a_{mk} e_m, \quad k = 1, \dots, N. \quad (103)$$

It follows from $\text{tr } A = \sum_{k=1}^N (e_k | Ae_k) = 0$ that $a_{11} + \cdots + a_{NN} = 0$, and $A^* = -A$ implies

$$a_{km} = (e_m | Ae_k) = -(Ae_m | e_k) = -\bar{a}_{mk}.$$

Thus, the matrix (a_{km}) is traceless and skew-adjoint. Conversely, the same argument shows that each traceless skew-adjoint matrix (a_{km}) corresponds to an operator $A \in \text{su}(N)$ given by (103).

Consider next the case $N = 3$. Each traceless skew-adjoint (3×3) -matrix can be written in the following form:

$$\begin{pmatrix} i\alpha & \alpha & b \\ -\bar{a} & i\beta & c \\ -\bar{b} & -\bar{c} & i\gamma \end{pmatrix},$$

where the real numbers α, β , and γ satisfy the trace relation $\alpha + \beta + \gamma = 0$. Since a, b , and c are complex numbers, such a matrix depends on $2 + 2(2 + 1) = 8$ real parameters. Thus,

$$\dim \text{su}(3) = 8.$$

Define the following matrices,

$$\mathcal{D}_1 := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \quad \mathcal{D}_2 := \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \quad (104)$$

$$\mathcal{A}_{12} := \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathcal{A}_{13} := \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \quad \mathcal{A}_{23} := \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}$$

$$\mathcal{B}_{12} := \begin{pmatrix} 0 & i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathcal{B}_{13} := \begin{pmatrix} 0 & 0 & i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix}, \quad \mathcal{B}_{23} := \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & i \\ 0 & i & 0 \end{pmatrix}.$$

Then, each traceless skew-adjoint (3×3) -matrix can be written as a real linear combination of the eight matrices from (104). Thus, they form a basis of the linear space $\text{su}(3)$.

In the general case, where $N = 1, 2, \dots$, the same argument yields

$$\begin{aligned} \dim \text{su}(N) &= (N - 1) + 2(1 + 2 + \cdots + N - 1) \\ &= N - 1 + (N - 1)N = N^2 - 1. \end{aligned}$$

Ad (iii). Let $A, B \in \text{su}(N)$. Then, $(iA)^* = iA$ (i.e., the operator iA is self-adjoint). Hence

$$\text{tr}(iA)^2 = \sum_{j=1}^N (e_j \mid (iA)^2 e_j) = \sum_{j=1}^N (iAe_j \mid iAe_j) = \sum_{j=1}^N \lambda_j^2,$$

where $iAe_j = \lambda_j e_j$ for all j . Thus,

$$(A \mid A) = -\text{tr } A^2 = \text{tr}(iA)^2.$$

Hence $(A \mid A) \geq 0$ for all $A \in \text{su}(N)$, and $(A \mid A) = 0$ implies $\lambda_j = 0$ for all j , that is, $A = 0$.

Furthermore, for all $A, B, C \in \text{su}(N)$ and $\alpha, \beta \in \mathbb{R}$, we get

$$(A \mid B) = -\text{tr}(AB) = -\text{tr}(BA) = (B \mid A),$$

and

$$\begin{aligned} (\alpha A + \beta B \mid C) &= -\text{tr}(\alpha AC + \beta BC) = -\alpha \text{tr}(AC) - \beta \text{tr}(BC) \\ &= \alpha(A \mid C) + \beta(B \mid C). \end{aligned} \quad \square$$

Definition 4. Let L be a Lie algebra over \mathbb{K} . By a *representation* of L on the Banach space X over \mathbb{K} , we understand a linear map

$$\phi: L \rightarrow L(X, X), \tag{105}$$

which respects the Lie brackets, that is,

$$\phi([A, B]) = [\phi(A), \phi(B)] \quad \text{for all } A, B \in L.$$

The representation ϕ is called *irreducible* iff there is no nontrivial invariant linear subspace Y of X , meaning that there is no linear subspace Y of X such that $Y \neq \{0\}$ and $Y \neq X$ as well as $\phi(A)(Y) \subseteq Y$.

Example 5 (The dual representation). Consider the representation ϕ from (105). Define

$$\phi_D(A) := -\phi(A)^T \quad \text{for all } A \in L.$$

Then, $\phi_D: L \rightarrow L(X^*, X^*)$ is a representation of L on the dual space X^* called the *dual representation* of ϕ on X^* .

Set $\mathcal{A} := \phi(A)$. Observe that the dual operator $\mathcal{A}^T: X^* \rightarrow X^*$ to $\mathcal{A}: X \rightarrow X$ is defined by

$$(\mathcal{A}^T u^*)(u) := u^*(\mathcal{A}u), \quad \text{for all } u \in X, u^* \in X^*.$$

Proof. We will show in Section 3.10 that $\mathcal{A} \in L(X, X)$ implies $\mathcal{A}^T \in L(X^*, X^*)$ and

$$(\alpha \mathcal{A} + \beta \mathcal{B})^T = \alpha \mathcal{A}^T + \beta \mathcal{B}^T \quad \text{and} \quad (\mathcal{A}\mathcal{B})^T = \mathcal{B}^T \mathcal{A}^T,$$

for all $\mathcal{A}, \mathcal{B} \in L(X, X)$ and $\alpha, \beta \in \mathbb{K}$. Hence

$$\begin{aligned} [\phi_D(A), \phi_D(B)] &= \phi(A)^T \phi(B)^T - \phi(B)^T \phi(A)^T \\ &= -(\phi(A)\phi(B) - \phi(B)\phi(A))^T = -[\phi(A), \phi(B)]^T \\ &= -\phi([A, B])^T = \phi_D([A, B]). \end{aligned} \quad \square$$

In what follows we will use tensor products, which have been discussed in detail in Problems 3.8ff of AMS Vol. 108.

Example 6 (Tensor representations). Let $\{e_1, \dots, e_N\}$ and $\{f_1, \dots, f_M\}$ be a basis of the linear space X and Y over \mathbb{K} , respectively. Suppose that $\phi: L \rightarrow L(X, X)$ and $\psi: L \rightarrow L(Y, Y)$ are representations of the Lie algebra L over \mathbb{K} . For $A \in L$ set

$$\phi_T(A)(e_j \otimes f_k) := (\phi(A)e_j \otimes f_k) + (e_j \otimes \psi(A)f_k)$$

and

$$\phi_T(A) \left(\sum_{j,k} \alpha_{jk} e_j \otimes f_k \right) := \sum_{j,k} \alpha_{jk} \phi_T(A)(e_j \otimes f_k)$$

for all $\alpha_{jk} \in \mathbb{K}$.

Then $\phi_T: L \rightarrow L(X \otimes Y, X \otimes Y)$ is a representation of L on the tensor product $X \otimes Y$, which is called the tensor product of ϕ with ψ . We also write $\phi_T := \phi \otimes \psi$.

Proof. One checks easily that ϕ_T is linear and that

$$\begin{aligned} (\phi_T(A)\phi_T(B) - \phi_T(B)\phi_T(A))(e_j \otimes f_k) &= (\phi(A)\phi(B) - \phi(B)\phi(A))e_j \otimes f_k \\ &\quad + e_j \otimes (\psi(A)\psi(B) - \psi(B)\psi(A))f_k. \end{aligned}$$

Since $\phi([A, B]) = \phi(A)\phi(B) - \phi(B)\phi(A)$ and $\psi([A, B]) = \psi(A)\psi(B) - \psi(B)\psi(A)$, we get

$$[\phi_T(A), \phi_T(B)] = \phi_T([A, B]) \quad \text{for all } A, B \in L. \quad \square$$

Example 7 (The adjoint representation). Let L be a subalgebra of the Lie algebra $L(X, X)$, where X is a Banach space over \mathbb{K} . Set

$$\phi(A)C := [A, C] \quad \text{for all } A \in L, \quad C \in L(X, X).$$

Then ϕ is a representation of L on $L(X, X)$, which is called the *adjoint representation* of L .

Proof. Recall that $[A, C] = AC - CA$. Hence $\|\phi(A)C\| \leq 2\|A\| \|C\|$. Thus, the operator $\phi(A): L(X, X) \rightarrow L(X, X)$ is linear and continuous, that is, $\phi(A) \in L(Y, Y)$, where $Y := L(X, X)$. It follows from

$$\phi(\alpha A + \beta B)C = [\alpha A + \beta B, C] = \alpha[A, C] + \beta[B, C] = (\alpha\phi(A) + \beta\phi(B))C$$

for all $A, B \in L$ and $\alpha, \beta \in \mathbb{K}$ that $\phi: L \rightarrow L(Y, Y)$ is linear.

Furthermore, we get

$$\begin{aligned} (\phi(A)\phi(B) - \phi(B)\phi(A))C &= [A, [B, C]] - [B, [A, C]] = [[A, B], C] \\ &= \phi([A, B])C, \end{aligned}$$

by the Jacobi identity. \square

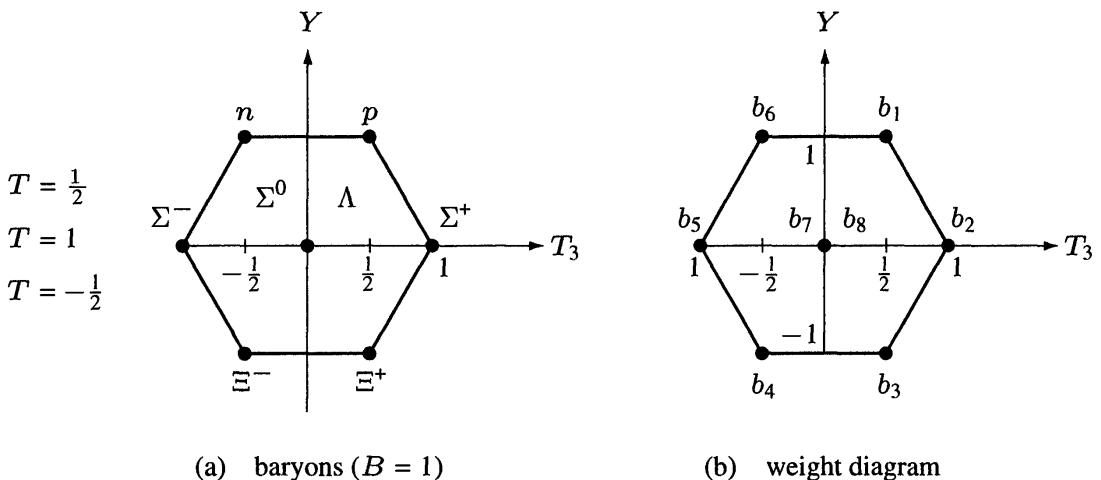


FIGURE 2.5.

2.22 Applications to Elementary Particles

Physicists introduce *quantum numbers* in order to classify elementary particles and to understand scattering processes in particle accelerators. For example, an important role is played by the isospin T , the third component T_3 of the isospin, and the hypercharge Y . The electric charge Q of an elementary particle is given by

$$Q = |e| \left(T_3 + \frac{Y}{2} \right) \quad (\text{Gell-Mann-Nishijima formula}),$$

where e = electric charge of the electron. Moreover,

$$T_3 = T, T-1, T-2, \dots, -T,$$

and $T = 0, \frac{1}{2}, 1, \frac{3}{2}, \dots$. Figure 2.5(a) displays eight baryons which behave similarly. In particular, Figure 2.5(a) tells us that

$$T_3 = \frac{1}{2}, \quad Y = 1 \quad \text{for the proton } p$$

and

$$T_3 = -\frac{1}{2}, \quad Y = 1 \quad \text{for the neutron } n.$$

Particles possess the same isospin T if they lie on the same vertical line in Figure 2.5(a). For example, we get $T = \frac{1}{2}$ for both the proton and the neutron (cf. Table 2.1). In this section we want to show that the diagram from Figure 2.5(a) corresponds to the *weight diagram* of a representation of the Lie algebra $\text{su}(3)$. As we will see later, abstract mathematical arguments lead us to the following fundamental hypothesis:

Both the proton and the neutron consist of three quarks.

This was formulated by Gell-Mann and Zweig in 1964.¹⁵

¹⁵Murray Gell-Mann was awarded the Nobel Prize in 1969.

TABLE 2.1. Quantum numbers of nucleons

	Proton	Neutron
Mixed state of three quarks	u, u, d	d, d, u
Isospin T	$\frac{1}{2}$	$\frac{1}{2}$
Third component of isospin T_3	$\frac{1}{2}$	$-\frac{1}{2}$
Hypercharge Y	1	1
$Q/ e $ (Q = electric charge)	1	0
Baryon number B	1	1
Strangeness S	0	0

2.22.1 Baryons and Quarks

Let X be a three-dimensional complex Hilbert space with the orthonormal basis $\{e_1, e_2, e_3\}$. Set

$$e_{jkm} := e_j \otimes e_k \otimes e_m.$$

Recall that the tensor product $X \otimes X \otimes X$ consists of all linear combinations

$$\sum_{j,k,m=1}^3 \alpha_{jkm} e_{jkm},$$

where $\alpha_{jkm} \in \mathbb{C}$ for all j, k, m . Moreover, $X \otimes X \otimes X$ is a complex Hilbert space with the orthonormal basis $\{e_{jkm}: j, k, m = 1, 2, 3\}$. Let $A: X \rightarrow X$ be a linear operator, namely, $A \in L(X, X)$. We write

$$A \simeq (a_{jk}), \quad j, k = 1, 2, 3, \quad \text{iff} \quad Ae_k = \sum_{j=1}^3 a_{jk} e_j \quad \text{for all } k.$$

In particular, let us introduce the four linear operators $\mathcal{Y}, \mathcal{T}_3, \mathcal{B}, \mathcal{S}: X \rightarrow X$ through

$$\mathcal{Y} \simeq \begin{pmatrix} \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & -\frac{2}{3} \end{pmatrix}, \quad \mathcal{T}_3 \simeq \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & -\frac{1}{2} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad (106)$$

and

$$\mathcal{B} \simeq \begin{pmatrix} \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} \end{pmatrix}, \quad \mathcal{S} \simeq \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

TABLE 2.2. Quantum numbers of quarks

Quantum Number	Quarks			Antiquarks		
	$u = e_1$	$d = e_2$	$s = e_3$	$\bar{u} = e_1^*$	$\bar{d} = e_2^*$	$\bar{s} = e$
Isospin T	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{2}$	0
Third component of the isospin T_3	$\frac{1}{2}$	$-\frac{1}{2}$	0	$-\frac{1}{2}$	$\frac{1}{2}$	0
Hypercharge Y	$\frac{1}{3}$	$\frac{1}{3}$	$-\frac{2}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$	$\frac{2}{3}$
Baryon number B	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$
Strangeness S	0	0	-1	0	0	1
$Q/ e $ (Q = electric charge, e = charge of the electron)	$\frac{2}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$	$-\frac{2}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Note that $i\mathcal{Y}, i\mathcal{T}_3, i\mathcal{B}, i\mathcal{S} \in su(3)$. For all $A \in L(X, X)$, define

$$\phi(A)e_{jkm} := Ae_j \otimes e_k \otimes e_m + e_j \otimes Ae_k \otimes e_m + e_j \otimes e_k \otimes Ae_m$$

and

$$\phi(A) \sum \alpha_{jkm} e_{jkm} := \sum \alpha_{jkm} \phi(A) e_{jkm}. \quad (107)$$

Then, the operator $\phi(A): X \otimes X \otimes X \rightarrow X \otimes X \otimes X$ is linear.

Furthermore, set

$$s_{jkm} := S(123)e_{jkm}, \quad a_{jkm} := A(123)e_{jkm}, \quad u_{jkm} := A(13)S(12)e_{jkm},$$

$$v_{jkm} := A(12)S(13)e_{jkm}.$$

Here, $S(123)$ (resp., $A(123)$) means symmetrization (resp., antisymmetrization) with respect to all three indices. Similarly, $S(12)$ means symmetrization with respect to the first and second indices, and so forth. Explicitly, we get

$$\begin{aligned} s_{jkm} &= e_{jkm} + e_{jmk} + e_{kmj} + e_{kjm} + e_{mjk} + e_{mkj}, \\ a_{jkm} &= e_{jkm} - e_{jmk} + e_{kmj} - e_{kjm} + e_{mjk} - e_{mkj}, \end{aligned}$$

as well as

$$\begin{aligned} u_{jkm} &= A(13)(e_{jkm} + e_{kjm}) = e_{jkm} - e_{mkj} + e_{kjm} - e_{mjk}, \\ v_{jkm} &= A(12)(e_{jkm} + e_{mkj}) = e_{jkm} - e_{kjm} + e_{mkj} - e_{kmj}. \end{aligned}$$

Finally, define the following linear subspaces of the tensor product $X \otimes X \otimes X$:

$$L_1 := \text{span}\{s_{jkm}\}, \quad L_2 := \text{span}\{a_{jkm}\},$$

$$L_3 := \text{span}\{u_{jkm}\}, \quad L_4 := \text{span}\{v_{jkm}\}.$$

By (106),

$$\begin{aligned} \mathcal{Y}e_j &= \lambda_j e_j, & \mathcal{T}_3 e_j &= \mu_j e_j, & j &= 1, 2, 3, \\ \mathcal{B}e_j &= \frac{1}{3}e_j, & \mathcal{S}e_j &= \nu_j e_j, \end{aligned} \tag{108}$$

where $\lambda_1 = \lambda_2 = \frac{1}{3}$, $\lambda_3 = -\frac{2}{3}$, $\mu_1 = \frac{1}{2}$, $\mu_2 = -\frac{1}{2}$, $\mu_3 = 0$, and $\nu_1 = \nu_2 = 0$, $\nu_3 = -1$. By (107),

$$\begin{aligned} \phi(\mathcal{Y})e_{jkm} &= (\lambda_j + \lambda_k + \lambda_m)e_{jkm}, & \phi(\mathcal{T}_3)e_{jkm} &= (\mu_j + \mu_k + \mu_m)e_{jkm}, \\ \phi(\mathcal{B})e_{jkm} &= e_{jkm}, & \phi(\mathcal{S})e_{jkm} &= (\nu_j + \nu_k + \nu_m)e_{jkm}, \end{aligned} \tag{109}$$

for all j, k, m . The same is true if we replace e_{jkm} with s_{jkm} , a_{jkm} , u_{jkm} , or v_{jkm} .

Proposition 1. *The following statements hold true.¹⁶*

- (i) $X \otimes X \otimes X = L_1 \oplus L_2 \oplus L_3 \oplus L_4$.
- (ii) *Each space L_j is invariant under the representation ϕ of the Lie algebra $su(3)$ on $X \otimes X \otimes X$. Here, ϕ is defined through (107).*
- (iii) *$\dim L_1 = 10$, and a basis of L_1 is given by*

$$s_{123}, s_{112}, s_{113}, s_{221}, s_{223}, s_{331}, s_{332}, s_{111}, s_{222}, s_{333}. \tag{110}$$

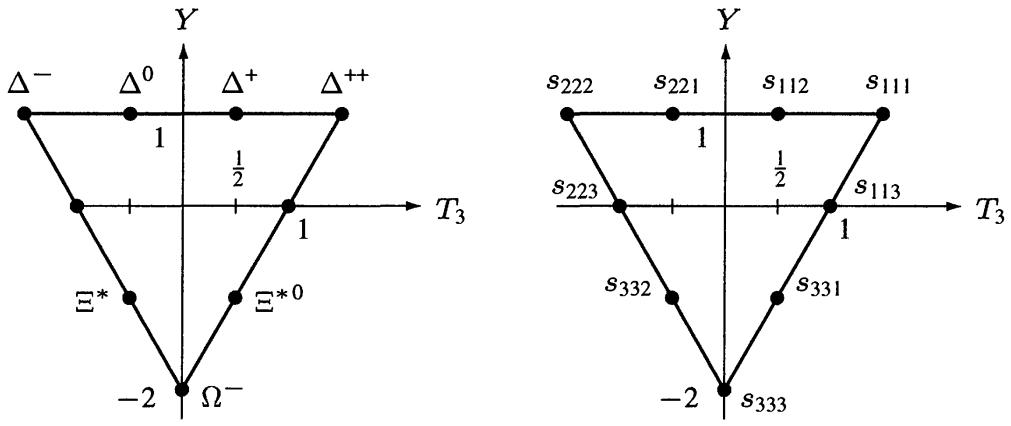
By (109), these basis vectors are common eigenvectors of the operators $\phi(\mathcal{Y})$ and $\phi(\mathcal{T}_3)$ with the eigenvalues T_3 and Y , respectively, which are pictured in Figure 2.6(b). For example,

$$\phi(\mathcal{Y})s_{112} = Ys_{112}, \quad \phi(\mathcal{T}_3)s_{112} = T_3s_{112},$$

where $Y = \lambda_1 + \lambda_1 + \lambda_2 = 1$ and $T_3 = \mu_1 + \mu_1 + \mu_2 = \frac{1}{2}$. Figure 2.6(b) is called the weight diagram of the representation of $su(3)$ on the space L_1 .

¹⁶The direct sum (i) says that, for each $u \in X \otimes X \otimes X$, there exists a unique decomposition

$$u = u_1 + u_2 + u_3 + u_4, \quad \text{where } u_j \in L_j \text{ for all } j.$$

(a) baryon resonances ($B = 1$)

(b) weight diagram

FIGURE 2.6.

(iv) $\dim L_2 = 1$, and a_{123} is a basis vector of L_2 with

$$\phi(\mathcal{Y})a_{123} = Ya_{123}, \quad \phi(T_3)a_{123} = T_3a_{123}, \quad \text{where } Y = T_3 = 0,$$

by (109).

(v) $\dim L_3 = 8$, and a basis of L_3 is given by

$$b_1 := u_{112}, \quad b_2 := u_{113}, \quad b_3 := u_{331}, \quad b_4 := u_{332}, \quad b_5 := u_{223}, \quad (111)$$

$$b_6 := u_{221}, \quad b_7 := u_{123}, \quad b_8 := u_{132}.$$

The corresponding weight diagram is pictured in Figure 2.5(b).

(vi) $\dim L_4 = 8$, and a basis of L_4 is given by

$$b_1 := v_{121}, \quad b_2 := v_{131}, \quad b_3 := v_{313}, \quad b_4 := v_{323}, \quad b_5 := v_{232},$$

$$b_6 := v_{212}, \quad b_7 := v_{132}, \quad b_8 := v_{123}.$$

The corresponding weight diagram is identical to the weight diagram of the representation of $su(3)$ on L_3 (cf. Figure 2.5(b)).

Proof. Ad (i). Use the decomposition

$$e_{jkm} = \frac{1}{6}s_{jkm} + \frac{1}{6}a_{jkm} + \frac{1}{3}u_{jkm} + \frac{1}{3}v_{jkm}.$$

Hence $X \otimes X \otimes X = L_1 + L_2 + L_3 + L_4$. As we will show, $\dim L_1 = 10$, $\dim L_2 = 1$, and $\dim L_3 = \dim L_4 = 8$. Since

$$\dim X \otimes X \otimes X = 27 \quad \text{and} \quad \dim L_1 + \dim L_2 + \dim L_3 + \dim L_4 = 27,$$

we get $X \otimes X \otimes X = L_1 \oplus L_2 \oplus L_3 \oplus L_4$.

Ad (ii). Observe that $\phi(A)$ commutes with symmetrization or antisymmetrization. For example,

$$\phi(A)S(123)e_{jkm} = S(123)\phi(A)e_{jkm}.$$

Hence $u \in L_1$ implies $\phi(A)u \in L_1$. The same argument applies to L_2 , L_3 , and L_4 .

Ad (iii). Since s_{jkm} is symmetric with respect to all indices, we get

$$s_{123} = s_{132} = s_{213} = s_{231} = s_{312} = s_{321}, \text{ and so on.}$$

Moreover, observe that $\{e_{jkm}\}$ forms an orthonormal system. Hence the vectors from (110) form an orthogonal system. That is, they are linearly independent.

Ad (v). Observe that

$$u_{jkm} = u_{kjm} \quad \text{and} \quad u_{jkm} + u_{jmk} = -u_{kmj}.$$

This implies that each u_{jkm} is a linear combination of b_1, \dots, b_8 . Moreover, an elementary computation shows that b_1, \dots, b_8 are linearly independent. For example, it follows from

$$\alpha b_1 + \beta b_2 + \gamma b_3 = 0$$

that $\alpha e_{112} + \beta e_{113} + \gamma e_{331} = 0$, and hence $\alpha = \beta = \gamma = 0$.

Ad (iv), (vi). Use similar arguments as previously. \square

One can show that the representations of $\text{su}(3)$ on L_1 , L_2 , L_3 , and L_4 are irreducible (cf. Cornwell (1989), Vol. 2, pp. 636ff).

Remark 2 (Physical interpretation). Set

- e_1 = state of the u -quark,
- e_2 = state of the d -quark,
- e_3 = state of the s -quark

and

$$\begin{aligned} e_{jkm} &\equiv e_j \otimes e_k \otimes e_m \\ &= \text{composite state of the three quark states } e_j, e_k, e_m. \end{aligned}$$

Comparing Figure 2.5(a) with Figure 2.5(b), it turns out that

The weight diagram of the representation of the Lie algebra $\text{su}(3)$ on L_3 is identical to the quantum number diagram of physicists.

Since each vector b_1, \dots, b_8 is a linear combination of e_{jkm} , we say that

The elementary particles from Figure 2.5(a) are mixed states consisting of three quarks.

For example,

$$\text{proton } p \cong b_1 = u_{112} = 2(e_1 \otimes e_1 \otimes e_2 - e_2 \otimes e_1 \otimes e_1).$$

This leads us to the following interpretation:

The proton consists of two u -quarks and one d -quark.

In quantum physics, one uses normalized states. Therefore, we have to replace u_{112} with

$$u_{112}^{(\text{norm})} = \frac{1}{\sqrt{2}}(e_1 \otimes e_1 \otimes e_2 - e_2 \otimes e_1 \otimes e_1).$$

Then, $(u_{112}^{(\text{norm})} | u_{112}^{(\text{norm})}) = 1$. Observe that $\{e_j \otimes e_k \otimes e_m\}$ forms an orthonormal system.¹⁷

The quantum numbers Y , T_3 , B , and S of quarks are given by the eigenvalues of the operators \mathcal{Y} , \mathcal{T}_3 , \mathcal{B} , and \mathcal{S} , respectively. For example, by (108),

$$\mathcal{Y}e_1 = \frac{1}{3}e_1, \quad \mathcal{T}_3e_1 = \frac{1}{2}e_1, \quad \mathcal{B}e_1 = \frac{1}{3}e_1, \quad \mathcal{S}e_1 = 0.$$

Hence the u -quark e_1 has the quantum numbers $Y = \frac{1}{3}$, $T_3 = \frac{1}{2}$, $B = 1$, and $S = 0$. This corresponds to Table 2.2.

Furthermore, the quantum numbers of composite particles correspond to the eigenvalues of the representations $\phi(\mathcal{Y})$, $\phi(\mathcal{T}_3)$, and so on. For example, if we note that $b_1 = u_{112}$, it follows from (109) that

$$\phi(\mathcal{Y})b_1 = b_1, \quad \phi(\mathcal{T}_3)b_1 = \frac{1}{2}b_1, \quad \phi(\mathcal{B})b_1 = b_1, \quad \phi(\mathcal{S})b_1 = 0.$$

Thus, the quantum numbers of the proton are given by $Y = 1$, $T_3 = \frac{1}{2}$, $B = 1$, and $S = 0$. This corresponds to Table 2.1.

Furthermore, the weight diagram of the representation of $\text{su}(3)$ on L_1 (cf. Figure 2.6(b)) corresponds to the baryon multiplet from Figure 2.6(a). Using the $\text{su}(3)$ -symmetry of this diagram, Gell-Mann and Neeman predicted the existence of the particle Ω^- in 1961. This particle was discovered shortly after its prediction.

¹⁷If we use the representation of $\text{su}(3)$ on L_4 , then we get the same weight diagram as for L_3 . In this case, the proton p corresponds to the state vector

$$b_1 = v_{121} = 2(e_1 \otimes e_2 \otimes e_1 - e_2 \otimes e_1 \otimes e_1).$$

This differs from the state vector u_{112} . Thus, the state vectors of elementary particles are not uniquely determined. We need a fixed convention. However, note that the possible proton states u_{112} and v_{121} correspond to a mixed state consisting of two u -quarks and one d -quark; the essential quark contents are the same for u_{112} and v_{121} .

2.22.2 Mesons and Pairs of Quarks and Antiquarks

We want to show that the dual representation of $\text{su}(3)$ is related to antiquarks and that the representation of $\text{su}(3)$ on $X \otimes X^*$ corresponds to mesons (pairs of quarks and antiquarks).

To begin with, define a basis $\{e_1^*, e_2^*, e_3^*\}$ of the dual space X^* by letting

$$e_j^* \left(\sum_{k=1}^3 \alpha_k e_k \right) := \alpha_j \quad \text{for all } \alpha_k \in \mathbb{C}.$$

The tensor product $X \otimes X^*$ consists of all linear combinations

$$\sum_{k,m=1}^3 \alpha_{km} e_{km}, \quad \text{where } e_{km} := e_k \otimes e_m^* \text{ and } \alpha_{km} \in \mathbb{C} \text{ for all } k, m.$$

Let $A: X \rightarrow X$ be a linear operator such that

$$A \simeq \mathcal{A} \quad \text{with the matrix } \mathcal{A} = (a_{km}).$$

Set

$$\phi_D(A) := -A^T.$$

Then, $\phi_D: L(X, X) \rightarrow L(X^*, X^*)$ is the *dual representation* of the Lie algebra $L(X, X)$ on $L(X^*, X^*)$. Explicitly, for $j = 1, 2, 3$,

$$\begin{aligned} -A^T e_k^*(e_j) &= -e_k^*(Ae_j) = -\sum_{m=1}^3 a_{mj} e_k^*(e_m) = -a_{kj} \\ &= -\sum_{m=1}^3 a_{km} e_m^*(e_j). \end{aligned}$$

Hence $-A^T e_k^* = -\sum_{m=1}^3 a_{km} e_m^*$, that is,

$$-A^T \simeq -\mathcal{A}^T,$$

where \mathcal{A}^T denotes the transposed matrix (i.e., $\mathcal{A}_{km}^T = \mathcal{A}_{mk}$). For example,

$$\mathcal{Y} \simeq \begin{pmatrix} \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & -\frac{2}{3} \end{pmatrix}, \quad \mathcal{T}_3 \simeq \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & -\frac{1}{2} & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

implies

$$\phi_D(\mathcal{Y}) \simeq \begin{pmatrix} -\frac{1}{3} & 0 & 0 \\ 0 & -\frac{1}{3} & 0 \\ 0 & 0 & \frac{2}{3} \end{pmatrix}, \quad \phi_D(\mathcal{T}_3) \simeq \begin{pmatrix} -\frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Remark 1 (Physical interpretation). We regard e_1^* , e_2^* , and e_3^* as the state vectors of the antiquarks \bar{u} , \bar{d} , and \bar{s} , respectively. Then, for $j = 1, 2, 3$,

$$\phi_D(\mathcal{Y})e_j^* = -\lambda_j e_j^*, \quad \phi_D(T_3)e_j^* = -\mu_j e_j^*,$$

where $\lambda_1 = \lambda_2 = \frac{1}{3}$, $\lambda_3 = -\frac{2}{3}$, and $\mu_1 = -\mu_2 = -\frac{1}{2}$, $\mu_3 = 0$. We say that e_j^* corresponds to a quark state with the quantum numbers $Y = -\lambda_j$ and $T_3 = -\mu_j$ (cf. Table 2.2).

For the linear operator $A: X \rightarrow X$ define

$$\phi(A)(e_k \otimes e_m^*) := Ae_k \otimes e_m^* + e_k \otimes (-A^T e_m^*).$$

Finally, set

$$\sigma := \sum_{k=1}^3 e_{kk}, \quad t_{km} := 3e_{km} - \sigma \delta_{km}$$

and

$$M_1 := \text{span}\{\sigma\}, \quad M_2 := \text{span}\{t_{km}: k, m = 1, 2, 3\}.$$

Recall that $e_{km} := e_k \otimes e_m^*$. Obviously,

$$\phi(\mathcal{Y})e_{km} = (\lambda_k - \lambda_m)e_{km}, \quad \phi(T_3)e_{km} = (\mu_k - \mu_m)e_{km}.$$

This implies

$$\phi(\mathcal{Y})\sigma = \phi(T_3)\sigma = 0, \quad \phi(\mathcal{Y})t_{km} = (\lambda_k - \lambda_m)t_{jk}, \quad \phi(T_3)t_{km} = (\mu_k - \mu_m)t_{km}. \quad (112)$$

Proposition 2. *The following are true:*

- (i) $X \otimes X^* = M_1 \oplus M_2$.
- (ii) M_1 and M_2 are invariant under the representation of $su(3)$ on $X \otimes X^*$.
- (iii) $\dim M_1 = 1$.
- (iv) $\dim M_2 = 8$, and a basis of M_2 is given by

$$b_1 := e_{12}, \quad b_2 := e_{11} - e_{22}, \quad b_3 := e_{21}, \quad b_4 := e_{13},$$

$$b_5 := e_{23}, \quad b_6 := e_{32}, \quad b_7 := e_{31}, \quad b_8 := 2e_{33} - e_{11} - e_{22}.$$

The proof will be given ahead. It can be shown that the representations of $su(3)$ on M_1 and M_2 are irreducible (cf. Cornwell (1989), Vol. 2, pp. 636ff.).

Remark 3 (Physical interpretation). We regard

$$e_{km} := e_k \otimes e_m^*$$

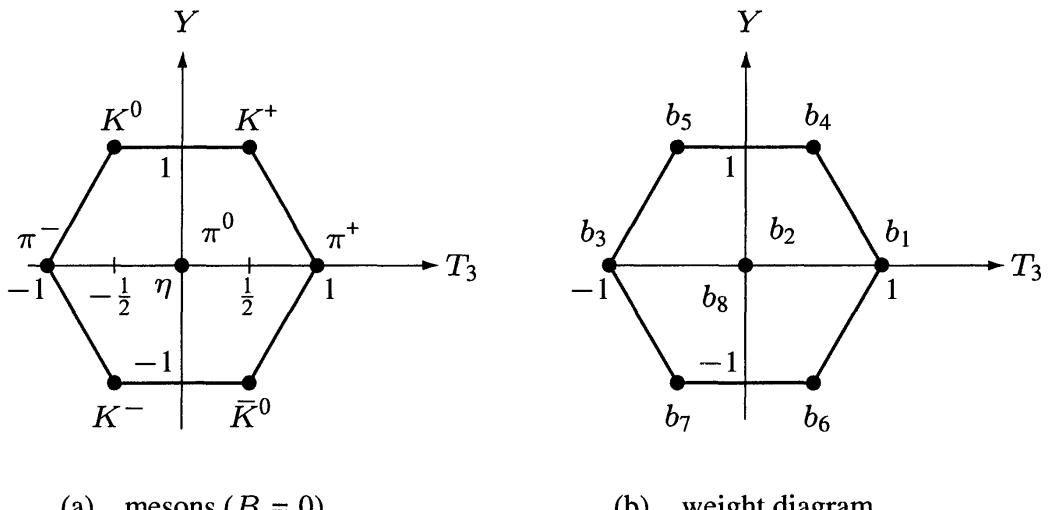


FIGURE 2.7.

as the state vector of a composite state consisting of the quark e_k and the antiquark e_m^* . Observe that b_1, \dots, b_8 are eigenvectors of $\phi(\mathcal{Y})$ and $\phi(T_3)$ with the eigenvalues Y and T_3 , respectively. The values of the quantum numbers Y and T_3 are pictured in Figure 2.7(b). This weight diagram of the representation of $\text{su}(3)$ on M_2 corresponds to the quantum number diagram of physicists for mesons (cf. Figure 2.7(a)). Therefore, physicists assume that the mesons from Figure 2.7(a) correspond to the states b_1, \dots, b_8 according to Figure 2.7(b). For example, the pion π^+ is described by the state vector

$$b_1 = e_1 \otimes e_2^*;$$

in other words, π^+ consists of one u -quark e_1 and one d -antiquark e_2^* .

The basis vector σ of M_1 corresponds to the meson η' .

Proof of Proposition 2. Ad (i). For all k, m ,

$$e_{km} = \frac{1}{3}(t_{km} + \sigma \delta_{km}).$$

Furthermore, note that $\dim(X \otimes X^*) = 9$, and $\dim M_1 + \dim M_2 = 1 + 8$, by the proof of (iv) ahead.

Ad (ii). Let $A \in \text{su}(3)$. Then

$$\begin{aligned} \phi(A) \left(\sum_{k=1}^3 e_{kk} \right) &= \sum_{k=1}^3 A e_k \otimes e_k^* - e_k \otimes A^T e_k^* \\ &= \sum_{k,m=1}^3 a_{mk} e_m \otimes e_k^* - a_{km} e_k \otimes e_m^* = 0. \end{aligned}$$

Hence $\phi(A)\sigma = 0$, i.e., M_1 is invariant under $\text{su}(3)$.

Furthermore, we get

$$\begin{aligned}\phi(A)t_{km} &= 3\phi(A)e_{km} = 3Ae_k \otimes e_m^* - 3e_k \otimes A^T e_m^* = 3 \sum_{j=1}^3 (a_{jk}e_{jm} - a_{mj}e_{kj}) \\ &= \sum_{j=1}^3 a_{jk}(3e_{jm} - \sigma\delta_{jm}) - a_{mj}(3e_{kj} - \sigma\delta_{kj}) \in M_2.\end{aligned}$$

Thus, M_2 is also invariant under $\text{su}(3)$.

Ad (iv). Observe that

$$t_{11} + t_{22} + t_{33} = 0, \quad t_{33} = b_8, \quad t_{11} - t_{22} = 3b_2. \quad \square$$

2.22.3 Applications to Gauge Field Theory

In this section, we sum over two equal lower and upper indices from 1 to 4 (the Einstein convention). We also set

$$\mathcal{L}(N) := \begin{cases} \text{su}(N) & \text{for } N \geq 2, \\ u(1) & \text{for } N = 1. \end{cases}$$

Here, $A_j \in \text{su}(N)$ iff A_j is a complex traceless skew-adjoint $(N \times N)$ -matrix, and $A_j \in u(1)$ iff iA_j is a real number.

The basic idea of gauge field theory and its importance for modern elementary physics has been discussed in Section 2.20, by considering a simple model. We are now ready to study the fundamental $\text{su}(N)$ -gauge field theory. Our point of departure is the following *variational problem*:¹⁸

$$\int_G L(\phi, \psi, A) dx = \text{stationary!}, \quad (113)$$

$$\phi, \psi, A_j = \text{given on } \partial G, \quad j = 1, \dots, 4,$$

¹⁸This variational problem refers to an arbitrary inertial system. By definition, a Cartesian coordinate system is an *inertial system* only if there exists a system time t such that each mass point, which is sufficiently distant from other masses and shielded against fields, remains at rest or moves rectilinearly with constant velocity.

Einstein's famous *principle of relativity* from 1905 postulates that all inertial systems are *physically equivalent*, meaning that physical processes are the same in all inertial systems when the initial and boundary conditions are the same. A detailed mathematical and physical discussion of this principle can be found in Zeidler (1986), Vol. 4, Chapter 75.

In order to prove that the principle (113) is indeed valid for each inertial system, one has to know the transformation rules for x , ϕ , ψ , and A under a change of the inertial system. It turns out that the four-potential A has to be a covariant vector and both ϕ and ψ have to be spinors under Lorentz transformations of x (cf. Thaller (1992), and Zeidler (1986), Vol. 5, Chapter 91).

with the Lagrangian

$$L := (\phi \mid i\gamma^4 \gamma^k \nabla_k \psi) - m(\phi \mid \gamma^4 \psi) - \frac{1}{4} \langle F_{jk}, F^{jk} \rangle, \quad (114)$$

where G is a bounded nonempty open set in \mathbb{R}^4 ,

$$\nabla_j := \partial_j + \kappa A_j, \quad A_j \in \mathcal{L}(N), \quad j = 1, \dots, 4, \quad N \geq 1,$$

and

$$F_{jk} := \partial_j A_k - \partial_k A_j + \kappa [A_j, A_k], \quad j, k = 1, \dots, 4,$$

as well as

$$\langle F_{jk}, F^{jk} \rangle := -\text{tr}(F_{jk} F^{jk}), \quad [A_j, A_k] := A_j A_k - A_k A_j.$$

Observe that

$$\kappa F_{jk} = [\nabla_j, \nabla_k] \quad \text{for all } j, k = 1, \dots, 4.$$

Furthermore, we define

$$D_j F^{jk} := \partial_j F^{jk} + \kappa [A_j, F^{jk}].$$

Here, the symbols have the following meaning:

- m = mass of the fundamental particle \mathcal{P} ,
- ψ, ϕ = field of \mathcal{P} (below we will set $\phi = \psi$),
- A = potential of the interaction between the particles \mathcal{P} ,
- F = field strength of the interaction,
- κ = coupling constant of the interaction,
- J = current generated by the particles \mathcal{P} ,
- x_1, x_2, x_3 = Cartesian coordinates,
- t = time,
- c = velocity of light, $x := (x_1, x_2, x_3, x_4)$, where $x_4 := ct$,
- $\partial_j := \partial_j / \partial x_j$ = classical derivative,
- ∇_j, D_j = covariant derivatives.

Observe that the components A_j of the potential and the components F_{jk} of the field strength are elements of the *Lie algebra* $\mathcal{L}(N)$. Furthermore, let us define the *metric matrix*

$$(g_{jk}) := \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

and the so-called *Dirac matrices*

$$\gamma^4 := \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \quad \gamma^1 := \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix},$$

$$\gamma^2 := \begin{pmatrix} 0 & 0 & 0 & -i \\ 0 & 0 & i & 0 \\ 0 & i & 0 & 0 \\ -i & 0 & 0 & 0 \end{pmatrix}, \quad \gamma^3 := \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

The crucial property of these matrices is given by the relation¹⁹

$$\gamma^j \gamma^k + \gamma^k \gamma^j = 2g^{jk} \quad \text{for all } j, k = 1, \dots, 4.$$

Here, (g^{jk}) denotes the inverse matrix to (g_{jk}) , that is, $g^{jk} = g_{jk}$ for all j, k . Moreover, $(\gamma^4)^* = \gamma^4$ and $(\gamma^k)^* = -\gamma^k$ for $k = 1, 2, 3$. We also set

$$F^{jk} := g^{jr} g^{ks} F_{rs}.$$

Finally, we want to define the Hilbert space X under consideration. First let Y be the four-dimensional complex Hilbert space of all the complex matrices

$$\psi_j = \begin{pmatrix} \psi_{j1} \\ \vdots \\ \psi_{j4} \end{pmatrix} \quad \text{with the inner product } (\phi_j | \psi_j)_Y := \sum_{k=1}^4 \bar{\phi}_{jk} \psi_{jk}.$$

Then, by definition, X is the complex Hilbert space of all the N -matrices

$$\psi = \begin{pmatrix} \psi_1 \\ \vdots \\ \psi_N \end{pmatrix} \quad \text{with } \psi_j \in Y \text{ for all } j$$

and the inner product

$$(\phi | \psi) := \sum_{j=1}^N (\phi_j | \psi_j)_Y.$$

Let $A_j \in \mathcal{L}(N)$, and denote the elements of A_j by a_{km} . Recall that the complex $(N \times N)$ -matrix $A_j = (a_{km})$ is traceless and skew-adjoint. This means that $a_{11} + \dots + a_{NN} = 0$ and $\bar{a}_{km} = -a_{mk}$ for $k, m = 1, \dots, N$. Recall that γ^k is a (4×4) -matrix.

¹⁹This relation says that the Dirac matrices γ^1 , γ^2 , γ^3 , and γ^4 generate a *Clifford algebra* (cf. Zeidler (1986), Vol. 5, Chapter 91).

Let us now extend γ^k and A_j to operators $\gamma^k, A_j: X \rightarrow X$ on the space X in a natural way. To this end, set

$$\gamma^k \psi := \begin{pmatrix} \gamma^k \psi_1 \\ \vdots \\ \gamma^k \psi_N \end{pmatrix} \quad \text{and} \quad A_j \psi := \begin{pmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & & \vdots \\ a_{N1} & \cdots & a_{NN} \end{pmatrix} \begin{pmatrix} \psi_1 \\ \vdots \\ \psi_N \end{pmatrix}.$$

Finally, set

$$\langle A, B \rangle := -\text{tr}(AB) \quad \text{for all } A, B \in \mathcal{L}(N), N \geq 1.$$

This way, we get an *inner product* $\langle \cdot, \cdot \rangle$ on the Lie algebra $\mathcal{L}(N)$, by Example 3 in Section 2.21. Let $\{B_j\}$ be an orthonormal basis of $\mathcal{L}(N)$ with respect to $\langle \cdot, \cdot \rangle$.

Theorem 2.M (Basic equations of gauge field theory). *Each sufficiently smooth solution $\phi, \psi, A_1, \dots, A_4$ of the original variational problem (113) with $\phi = \psi$ satisfies the following Euler–Lagrange equations:*

$$D_j F^{jk} = J^k, \quad k = 1, \dots, 4 \quad (\text{Yang–Mills equation}) \quad (115a)$$

$$D_j F_{km} + D_k F_{mj} + D_m F_{jk} = 0, \quad j, k, m = 1, \dots, 4 \quad (\text{Bianchi identity}) \quad (115b)$$

$$(i\gamma^k \nabla_k - mI)\psi = 0 \quad (\text{Dirac equation}). \quad (115c)$$

Here, the current J is given by

$$J^k := - \sum_{r=1}^{N^2-1} \kappa(\psi | i\gamma^4 \gamma^k B_r \psi) B_r \quad \text{for } N \geq 2,$$

and $J^k := -\kappa(\psi | i\gamma^4 \gamma^k \psi)$ for $N = 1$.

The proof of Theorem 2.M will be given ahead.

Example 1 (The Maxwell equations for the electromagnetic field). Let $N = 1$. In the special case where $\phi, \psi \equiv 0$, (115a, b) represent the Maxwell equations in vacuum. Here, the real functions $-iA_1, \dots, -iA_4$ are the components of the *four-potential*, and the real functions

$$-iF_{jk} := -i(\partial_j A_k - \partial_k A_j), \quad j, k = 1, \dots, 4,$$

are the components of the classical *electromagnetic field tensor*²⁰ (cf. Zei-

²⁰Our notation is chosen in such a way that A_j and F_{jk} are elements of the Lie algebra $\mathcal{L}(N)$. This convention simplifies the notation and clarifies the mathematics. Physicists frequently replace A_j with iA_j and F_{jk} with iF_{jk} .

dler (1986), Vol. 5, Chapter 83). Explicitly,

$$(-iF_{jk}) = \begin{pmatrix} 0 & -B_3 & B_2 & -E_1/c \\ B_3 & 0 & -B_1 & -E_2/c \\ -B_2 & B_1 & 0 & -E_3/c \\ E_1/c & E_2/c & E_3/c & 0 \end{pmatrix},$$

where E_1, E_2, E_3 and B_1, B_2, B_3 are the Cartesian components of the electric field vector \mathbf{E} and magnetic field vector \mathbf{B} , respectively.

Note that, in this special classical case, we get $[A_j, A_k] = 0$, and hence $\nabla_j = D_j = \partial_j$. In the language of vector calculus, the Maxwell equations in vacuum (115a) and (115b) with $\phi, \psi \equiv 0$ are identical to

$$\operatorname{curl} \mathbf{E} = 0, \quad \operatorname{curl} \mathbf{B} = c^{-2} \mathbf{E}_t, \quad (115\text{a}^*)$$

$$\operatorname{div} \mathbf{E} = 0, \quad \operatorname{div} \mathbf{B} = 0. \quad (115\text{b}^*)$$

This shows that

The basic equations (115) of gauge field theory generalize the classical Maxwell equations of electromagnetism.

Example 2 (The Dirac equation). Let $N = 1$ and $A_j \equiv 0$ for all j .

In this case, system (115) passes over to the Dirac equation (115c) with $\nabla_j = \partial_j$. This equation describes a free relativistic electron.

Example 3 (The Maxwell–Dirac equations). Let $N = 1$ and $\kappa = e$ = electric charge of the electron. Then, system (115) represents the Maxwell–Dirac equations of quantum electrodynamics.

Let us now discuss the fundamental gauge invariance. By a *gauge transformation*

$$\psi \mapsto \psi', \quad \phi \mapsto \phi', \quad A_j \mapsto A'_j, \quad \gamma^k \mapsto \gamma'^k, \quad (116)$$

we understand²¹

$$\psi'(x) = U(x)\psi(x), \quad \phi'(x) = U(x)\phi(x), \quad \gamma'^k = U\gamma^k U^{-1}, \quad (116^*)$$

$$A'_j(x) = U(x)A_j(x)U(x)^{-1} + \kappa^{-1}U(x)\partial_j U(x)^{-1},$$

where

$$U(x) := e^{A(x)} \quad \text{with } A(x) \in \mathcal{L}(N) \quad \text{for all } x \in \mathbb{R}^4.$$

Corollary 4 (Gauge invariance). *If we set $\nabla'_j := \partial_j + \kappa A'_j$, then*

$$\nabla'_j \psi'(x) = U(x)\nabla_j \psi(x), \quad (117)$$

²¹Note that the prime does not denote any derivative.

$$F'_{jk}(x) = U(x)F_{jk}(x)U(x)^{-1}, \quad (118)$$

$$L(\phi'(x), \psi'(x), A'(x), \gamma'^k) = L(\phi(x), \psi(x), A(x), \gamma^k), \quad (119)$$

for all $j, k = 1, \dots, 4$ and $x \in \mathbb{R}^4$.

From (119) we get the following fundamental result:

The basic equations (115) from Theorem 2.M are invariant under the gauge transformation (116).

Explicitly, this means the following. If ϕ, ψ, A is a solution of (115), then ϕ', ψ', A' is a solution of (115) provided a prime is assigned to all symbols.

Proof of Corollary 4. This follows from (116) by a straightforward computation.

Ad (117). Since

$$\partial_j U(x) = (\partial_j A(x))U(x) \quad \text{and} \quad \partial_j U(x)^{-1} = -(\partial_j A(x))U(x)^{-1},$$

we get

$$\begin{aligned} \nabla'_j \psi' &= (\partial_j + \kappa A'_j)U\psi = (\partial_j U)\psi + U\partial_j \psi + \kappa(UA_j U^{-1})U\psi + (U\partial_j U^{-1})U\psi \\ &= U(\partial_j + \kappa A_j)\psi = U\nabla_j \psi. \end{aligned}$$

Observe that $AU = UA$ and $(\partial_j A)U = U\partial_j A$, since $e^A = I + A + \frac{1}{2}A^2 + \dots$.

Ad (118). Note that $\kappa F_{jk} = [\nabla_j, \nabla_k] \equiv \nabla_j \nabla_k - \nabla_k \nabla_j$. By (117), $\nabla'_j \phi = U\nabla_j(U^{-1}\phi)$. Hence

$$\kappa F'_{jk} = [\nabla'_j, \nabla'_k] = [U\nabla_j U^{-1}, U\nabla_k U^{-1}] = U[\nabla_j, \nabla_k]U^{-1} = \kappa UF_{jk}U^{-1}.$$

Ad (119). Observe that the operator U is unitary. Hence

$$\begin{aligned} (\phi' | \gamma'^4 \psi') &= (U\phi | U\gamma^4 \psi) = (\phi | \gamma^4 \psi), \\ (\phi' | \gamma'^4 \gamma'^k \nabla'_k \psi') &= (U\phi | U\gamma^4 \gamma^k \nabla_k \psi) = (\phi | \gamma^4 \gamma^k \nabla_k \psi), \end{aligned}$$

and

$$\begin{aligned} \langle F'_{jk}, F'^{jk} \rangle &= \langle UF_{jk}U^{-1}, UF^{jk}U^{-1} \rangle = -\text{tr}(UF_{jk}U^{-1}UF^{jk}U^{-1}) \\ &= -\text{tr}(UF_{jk}F^{jk}U^{-1}) = -\text{tr}(U^{-1}UF_{jk}F^{jk}) = \langle F_{jk}, F^{jk} \rangle. \quad \square \end{aligned}$$

Proof of Theorem 2.M. Let $\tau \in \mathbb{R}$. Set

$$S(\tau) := \int_G L(\phi + \tau\delta\phi, \psi + \tau\delta\psi, A + \tau\delta A)dx,$$

where $\delta\phi(x), \delta\psi(x) \in Y$, $\delta A_j(x) \in \mathcal{L}(N)$, and the components of $\delta\phi, \delta\psi$, and δA_j are $C_0^\infty(G)$ -functions.

Suppose that ϕ, ψ, A is a solution of the original variational problem (113). Then $S'(0) = 0$, that is,

$$\begin{aligned} 0 = S'(0) &= \int_G (\delta\phi \mid \gamma^4(i\gamma^k \nabla_k - mI)\psi) + (\phi \mid \gamma^4(i\gamma^k \nabla_k - mI)\delta\psi) dx \\ &\quad + \int_G (\phi \mid i\gamma^4 \gamma^k \kappa \delta A_k \psi) - \frac{1}{2} \langle \delta F_{jk}, F^{jk} \rangle dx, \end{aligned}$$

where

$$\delta F_{jk} := \partial_j \delta A_k - \partial_k \delta A_j + \kappa[\delta A_j, A_k] + \kappa[A_j, \delta A_k].$$

Step 1: Let $\delta\psi \equiv 0$ and $\delta A_j \equiv 0$ for all j . Then

$$\int_G (\delta\phi \mid \gamma^4(i\gamma^k \nabla_k - mI)\psi) dx = 0,$$

for all $\delta\phi$ with $C_0^\infty(G)$ -components. Hence

$$\gamma^4(i\gamma^k \nabla_k - mI)\psi = 0. \quad (120)$$

Step 2: Let $\delta\phi \equiv 0$ and $\delta A_j \equiv 0$ for all j . Observe that

$$(i\gamma^4 \gamma^k)^* = -i\gamma^4 \gamma^k \quad \text{for } k = 1, 2, 3, 4. \quad (121)$$

In fact, $\gamma^{4*} = \gamma^4$ and

$$(i\gamma^4 \gamma^k)^* = -i\gamma^{k*} \gamma^4 = i\gamma^k \gamma^4 = -i\gamma^4 \gamma^k \quad \text{for } k = 1, 2, 3.$$

Furthermore, it follows from $(A_k)^* = -A_k$ and $A_j \gamma^k = \gamma^k A_j$ for $k, j = 1, 2, 3, 4$ that

$$(i\gamma^4 \gamma^k A_k)^* = A_k^* (i\gamma^4 \gamma^k)^* = i\gamma^4 \gamma^k A_k \quad \text{for } k = 1, 2, 3, 4. \quad (122)$$

Thus, integration by parts yields

$$0 = S'(0) = \int_G (\phi \mid B\delta\psi) dx = \int_G (B\phi \mid \delta\psi) dx$$

for all $\delta\psi$ with $C_0^\infty(G)$ -components, where $B := \gamma^4(i\gamma^k(\partial_k + \kappa A_k) - mI)$. Hence

$$B\phi \equiv \gamma^4(i\gamma^k \nabla_k - mI)\phi = 0.$$

By (120), the function ψ satisfies the same equation. This shows that our assumption $\phi = \psi$ from Theorem 2.M makes sense. In what follows we will use this assumption.

Step 3: Let $\delta\phi \equiv 0$ and $\delta\psi \equiv 0$. Set $\delta A_k(x) := a_k(x)B_s$, where $a_k \in C_0^\infty(G)$. Then

$$\begin{aligned} (\psi \mid i\gamma^4 \gamma^k \kappa \delta A_k \psi) &= (\psi \mid i\gamma^4 \gamma^k \kappa B_s \psi) a_k \\ &= \sum_r \langle (\psi \mid i\gamma^4 \gamma^k \kappa B_r \psi) B_r, B_s \rangle a_k \\ &= -\langle J^k, B_s \rangle a_k = -\langle J^k, \delta A_k \rangle, \end{aligned}$$

because $\langle B_r, B_s \rangle = \delta_{rs}$. Since $(B_r)^* = -B_r^*$, it follows as in (122) that

$$(i\gamma^4 \gamma^k B_r)^* = i\gamma^4 \gamma^k B_r.$$

Consequently, $(\psi | i\gamma^4 \gamma^k \kappa B_r \psi)$ is real, and $B_r \in \mathcal{L}(N)$ hence implies that

$$J^k \in \mathcal{L}(N).$$

Furthermore, note that $\text{tr}(DE) = \text{tr}(ED)$, and hence

$$\begin{aligned} \langle [A, B], C \rangle &= -\text{tr}(ABC) + \text{tr}(BAC) \\ &= -\text{tr}(CAB) + \text{tr}(ACB) = -\langle [A, C], B \rangle. \end{aligned}$$

Since $F^{jk} = -F^{kj}$, this implies

$$\begin{aligned} \frac{1}{2} \langle \delta F_{jk}, F^{jk} \rangle &= \langle \partial_j \delta A_k, F^{jk} \rangle + \langle \kappa [A_j, \delta A_k], F^{jk} \rangle \\ &= \langle \partial_j \delta A_k, F^{jk} \rangle - \langle \kappa [A_j, F^{jk}], \delta A_k \rangle. \end{aligned}$$

Integration by parts yields

$$\begin{aligned} 0 = S'(0) &= \int_G (\psi | i\gamma^4 \gamma^k \kappa \delta A_k \psi) - \frac{1}{2} \langle \delta F_{jk}, F^{jk} \rangle dx \\ &= \int_G \langle -J^k + \partial_j F^{jk} + \kappa [A_j, F^{jk}], \delta A_k \rangle dx \\ &\equiv \int_G \langle -J^k + D_j F^{jk}, \delta A_k \rangle dx = \int_G \langle -J^k + D_j F^{jk}, B_s \rangle a_k dx \end{aligned}$$

for all $a_k \in C_0^\infty(G)$ and all s . Hence

$$\langle -J^k + D_j F^{jk}, B_s \rangle = 0 \quad \text{for all } s.$$

Since $\{B_s\}$ is an orthonormal basis of $\mathcal{L}(N)$ with respect to the inner product $\langle \cdot, \cdot \rangle$, this implies

$$-J^k + D_j F^{jk} = 0.$$

In this connection, observe that J^k , F^{jk} , and $D_j F^{jk}$ are elements of the Lie algebra $\mathcal{L}(N)$.

Step 4: A straightforward computation shows that the Bianchi identity (115b) is a consequence of

$$F_{jk} = \partial_j A_k - \partial_k A_j + \kappa [A_j, A_k]. \quad \square$$

Remark 5 (Gauge theory and modern differential geometry). A more detailed study of the mathematical and physical aspects of gauge field theory can be found in Zeidler (1986), Vol. 5, and in the handbook article Zeidler

(1995), Chapter 19. There it is shown that, in terms of modern differential geometry, the potential A and the field F of interaction are related to the connection and the curvature of principal fiber bundles. Roughly speaking,

potential A = connection of the principal fiber bundle \mathcal{F} ,

field F of interaction = curvature of \mathcal{F} ,

particle field ψ = section of the associated vector bundle \mathcal{V} .

Let us briefly discuss this. To simplify notation, let $\kappa = 1$. In what follows we sum over j from 1 to 4.

(i) *The Lie group \mathcal{G} .* Set

$$\mathcal{G}(N) := \begin{cases} \mathrm{SU}(N) & \text{for } N \geq 2, \\ \mathrm{U}(1) & \text{for } N = 1. \end{cases}$$

Recall that $\mathrm{SU}(N)$ denotes the set of all complex unitary $(N \times N)$ -matrices \mathbf{U} with $\det \mathbf{U} = 1$. Moreover, $\mathrm{U}(1)$ denotes the set of all complex numbers u with $|u| = 1$.

The Lie algebra $\mathcal{L}(N)$ introduced at the beginning of Section 2.22.3 is precisely the Lie algebra corresponding to $\mathcal{G}(N)$.

(ii) *Principal fiber bundle.* The product space

$$\mathcal{F} := \mathbb{R}^4 \times \mathcal{G}(N)$$

is a special case of a fiber bundle. The set $\{x\} \times \mathcal{G}(N)$ is called the *fiber* over the point x . Since the typical fiber $\mathcal{G}(N)$ is a Lie group, the fiber bundle \mathcal{F} is called a principal fiber bundle.

(iii) *The associated vector bundle \mathcal{V} .* The values $\psi(x)$ of the field ψ live in the linear space X . Set

$$\mathcal{V} := \mathbb{R}^4 \times X.$$

This is a fiber bundle. Since the typical fiber X of \mathcal{V} is a linear space, \mathcal{V} is called a vector bundle associated to the principal fiber bundle \mathcal{F} .

The map $x \mapsto (x, \psi(x))$ from the base space \mathbb{R}^4 into the vector bundle \mathcal{V} is called a *section* of \mathcal{V} .

(iv) *Parallel transport in the principal fiber bundle \mathcal{F} .* Let $x = x(\sigma)$, $a \leq \sigma \leq b$, be a C^1 -curve in \mathbb{R}^4 . Recall that $A_j \in \mathcal{L}(N)$. The differential equation

$$g'(\sigma) + x'_j(\sigma)A'_j(x(\sigma))g(\sigma) = 0, \quad a \leq \sigma \leq b, \tag{P}$$

is called the equation of parallel transport in \mathcal{F} . We are given the matrix $g(0) \in \mathcal{G}(N)$. We are looking for a function

$$g = g(\sigma), \quad \text{where } g(\sigma) \in \mathcal{G}(N) \text{ for all } \sigma \in [a, b].$$

If $g = g(\sigma)$ is a solution of (P), then we say that the curve in \mathcal{F} ,

$$\sigma \mapsto (x(\sigma), g(\sigma)),$$

corresponds to a parallel transport of the initial point $(x(0), g(0)) \in \mathcal{F}$ along the base curve $x = x(\sigma)$.

The coefficients A_j of equation (P) are called a *connection* on \mathcal{F} .

In order to show that equation (P) makes sense, observe the following: If $g(\sigma)$ lives in the Lie group $\mathcal{G}(N)$ for all σ , then the derivative $g'(\sigma)$ lives in the corresponding Lie algebra $\mathcal{L}(N)$.

(v) *Parallel transport in the associated vector bundle \mathcal{V} .* Let the particle field ψ be given, where

$$\psi(x) \in X \quad \text{for all } x \in \mathbb{R}^4.$$

Set $\Psi(\sigma) := \psi(x(\sigma))$. By definition, the field ψ is *parallel* along the base curve $x = x(\sigma)$ iff

$$\Psi'(\sigma) + x'_j(\sigma)A_j(x(\sigma))\Psi(\sigma) = 0, \quad a \leq \sigma \leq b. \quad (\text{P}^*)$$

If we introduce the covariant *directional derivative* along the curve $x = x(\sigma)$ through

$$\frac{D}{d\sigma} := x'_j(\sigma)\nabla_j,$$

then equation (P^{*}) can be written elegantly as

$$\frac{D\psi}{d\sigma}(x(\sigma)) = 0, \quad a \leq \sigma \leq b.$$

Observe that $\Psi'(\sigma) = x'_j(\sigma)\partial_j\Psi(x(\sigma))$ and $\nabla_j = \partial_j + A_j$.

(vi) *Curvature of the principal fiber bundle \mathcal{F} .* In modern differential geometry, curvature is measured by the commutator of covariant derivatives. Therefore, let us describe the curvature of \mathcal{F} through the commutators

$$F_{jk} = [\nabla_j, \nabla_k],$$

where $[\nabla_j, \nabla_k] = \nabla_j\nabla_k - \nabla_k\nabla_j$. This yields

$$F_{jk} = \partial_j A_k - \partial_k A_j + [A_j, A_k],$$

which is precisely the field F of interaction.

General fiber bundles are manifolds that possess a local product structure. The intrinsic formulation of parallel transport and curvature is then based on the language of differential forms with values in a Lie algebra (cf. Zeidler (1986), Vol. 5, (1995)).

As an introduction to this subject, we recommend Isham (1989).

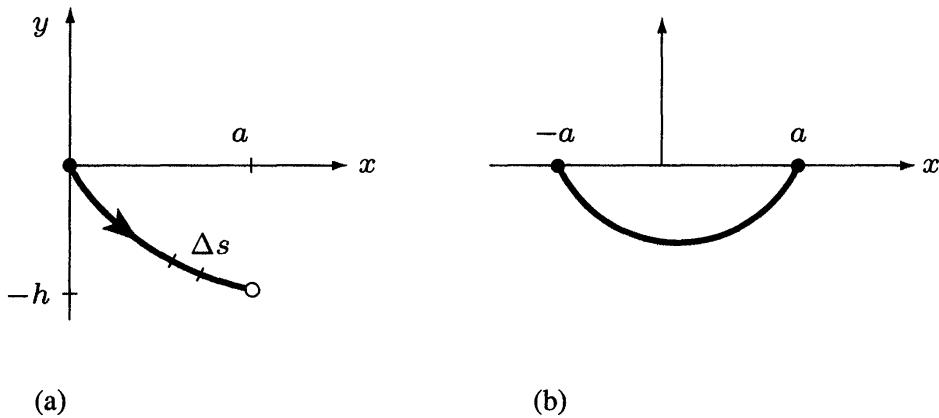


FIGURE 2.8.

Problems

2.1. Bernoulli's brachistochrone problem. Two points, at different distances from the ground and not in a vertical line, should be connected by such a curve

$$y = y(x)$$

so that a body under the influence of gravitational forces passes in the shortest possible time from the upper to the lower point. Compute this curve. Show that this corresponds to the following variational problem (see Figure 2.8(a)):

$$\int_0^a \frac{\sqrt{1+y'^2}}{\sqrt{-y}} dx = \min !, \quad y(0) = 0, \quad y(a) = -h. \quad (123)$$

Solution: The energy E of the body is given by

$$E = \frac{1}{2}mv^2 + mgy = \text{const},$$

where m = mass, v = velocity, and g = gravitational acceleration. At the beginning of the motion, we have $y = v = 0$, and hence $E = 0$. This implies

$$v = \sqrt{-2gy}.$$

Let s denote the arclength of the curve. Then, $v = \frac{\Delta s}{\Delta t}$. Thus,

$$\text{time} = \sum \Delta t = \sum \frac{\Delta s}{v} = \min !.$$

Using $\Delta s = \sqrt{(\Delta x)^2 + (\Delta y)^2}$ and letting $\Delta x \rightarrow 0$, we get (123).

The Euler equation to (123) is given by

$$\frac{d}{dx} L_{y'} - L_y = 0,$$

where $L = \frac{\sqrt{1+y'^2}}{\sqrt{-y}}$. Since L is independent of x , we get

$$\frac{d}{dx} (L - y' L_{y'}) = L_y y' + L_{y'} y'' - y'' L_{y'} - y' \frac{d}{dx} L_{y'} = 0.$$

Hence $L - y'L_{y'} = \text{const}$, that is,

$$\frac{\sqrt{1+y'^2}}{\sqrt{-y}} - \frac{y'^2}{\sqrt{-y}\sqrt{1+y'^2}} = \frac{1}{\sqrt{2C}}. \quad (124)$$

One checks that the cycloid

$$x = C(u - \sin u), \quad y = C(\cos u - 1)$$

is a solution of (124).

2.2. The hanging rope. Compute the shape $y = y(x)$ of a hanging rope (Figure 2.8(b)). Motivate that this corresponds to the following variational problem:

$$\int_{-a}^a \rho g \sqrt{1+y'^2} dx = \min !, \quad y(a) = y(-a) = 0, \quad (125)$$

where ρ = density and g = gravitational acceleration.

Solution: A piece of the hanging rope has the potential energy $\rho g \Delta s$, where s denotes the arclength. Problem (125) corresponds to the *principle of minimal potential energy*:

$$E_{\text{pot}} = \sum \rho g \Delta s = \min !.$$

The Euler equation to (125) reads as follows:

$$\frac{d}{dx} \frac{y'}{\sqrt{1+y'^2}} = 0.$$

Hence $\frac{y'}{\sqrt{1+y'^2}} = \text{const}$. This yields $y = \cosh x - \cosh a$.

2.3. Minimal surfaces. Compute the Euler equation to the variational problem

$$\int_G \sqrt{1+u_x^2+u_y^2} dxdy = \min !,$$

u = given on ∂G .

This is the problem of the least area for a prescribed boundary curve (see Figure 2.9).

Solution: We get

$$\frac{\partial}{\partial x} \frac{u_x}{\sqrt{1+u_x^2+u_y^2}} + \frac{\partial}{\partial y} \frac{u_y}{\sqrt{1+u_x^2+u_y^2}} = 0.$$

Existence theorems can be found in Problem 2.13.

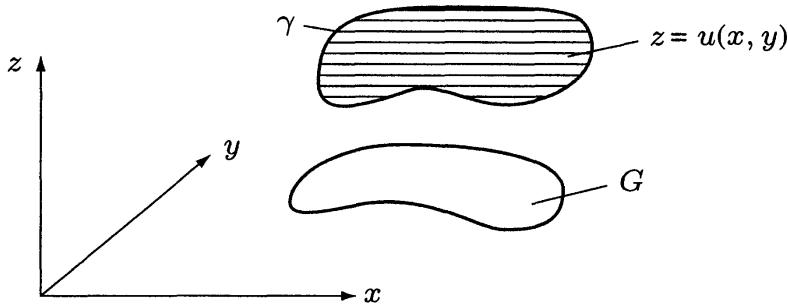


FIGURE 2.9.

2.4. The generalized Dirichlet problem. Compute the Euler equation to the variational problem

$$\int_G (F(u_x^2 + u_y^2) - 2fu) dx dy = \min !, \quad u \text{ given on } \partial G.$$

Solution: We obtain

$$\frac{\partial}{\partial x} (F'(u_x^2 + u_y^2)u_x) + \frac{\partial}{\partial y} (F'(u_x^2 + u_y^2)u_y) = f.$$

This is identical to the *conservation law*

$$\operatorname{div} \mathbf{j} = f,$$

where

$$\mathbf{j} = F'(|\operatorname{grad} u|^2) \operatorname{grad} u.$$

Equations of this type appear frequently in mathematical physics (cf. Zeidler (1986), Vol. 2B, Sections 25.9ff).

2.5. Motion of relativistic particles. Compute the Euler equation to the variational problem

$$\int_{t_0}^{t_1} \left(-m_0 c^2 \sqrt{1 - \frac{x'^2}{c^2}} - U(x) \right) dt = \min !,$$

$$x(t_0) = a, \quad x(t_1) = b,$$

where $x = (\xi_1, \xi_2, \xi_3)$ and $x'^2 = \sum_{j=1}^3 \xi_j'^2$.

Solution: We get

$$\frac{d}{dt} (m \xi'_j) = -\frac{\partial U}{\partial \xi_j}, \quad j = 1, 2, 3,$$

with the so-called relativistic mass

$$m = \frac{m_0}{\sqrt{1 - \frac{x'^2}{c^2}}}.$$

This can be written briefly as

$$(mx')' = -\text{grad } U.$$

This is the equation of motion $x = x(t)$ for a relativistic particle in an inertial system under the influence of the force $K = -\text{grad } U$. Observe that the relativistic mass m of the particle depends on its velocity. This mass goes to infinity if the particle approaches the velocity c of light.

A detailed study of the physical and mathematical meanings of this problem in special relativity can be found in Zeidler (1986), Vol. 4, Section 75.11.

2.6. The n th variation.

Let

$$F(u) := \int_a^b L(x, u(x), u'(x), \dots, u^{(n)}(x)) dx,$$

where $-\infty < a < b < \infty$ and $n \geq 1$. Suppose that the Lagrangian $L: [a, b] \times \mathbb{R}^{n+1} \rightarrow \mathbb{R}$ is C^2 . Set $X := C^n[a, b]$ (cf. Problem 1.6c in AMS Vol. 108). Show that

- (i) The functional $F: X \rightarrow \mathbb{R}$ is continuous.
- (ii) If L is convex with respect to the arguments $u, u', \dots, u^{(n)}$, then F is convex on X .
- (iii) Compute the first variation $\delta F(u; h)$ and the second variation $\delta^2 F(u; h)$ for all $u, h \in X$.

Hint: Letting $\phi(t) := F(u + th)$ for $t \in \mathbb{R}$, we get $\delta^r F(u; h) = \phi^{(r)}(0)$, and hence

$$\begin{aligned} \delta F(u; h) &= \int_a^b \sum_{k=0}^n [D_k L(x, u(x), \dots, u^{(n)}(x))] h^{(k)}(x) dx, \\ \delta^2 F(u; h) &= \int_a^b \sum_{m=0, k=0}^n [D_k D_m(x, u(x), \dots, u^{(n)}(x))] h^{(m)}(x) h^{(k)}(x) dx, \end{aligned}$$

where $D_k := \partial/\partial u^{(k)}$.

2.7a. Lower semicontinuous functionals. Let $F, G: M \subseteq X \rightarrow \mathbb{R}$ be two functionals on the subset M of the real normed space X . Show that

- (i) If F and G are convex on the convex set M , then so is $F + G$.
- (ii) If F and G are lower semicontinuous on the closed set M , then so is $F + G$.

Solution: Ad (i). Use the definition of convex functionals.

Ad (ii). For $r \in \mathbb{R}$, let $\mathcal{M}_r(F) := \{u \in M : F(u) \leq r\}$. The set

$$\mathcal{N}_r(F) := \{u \in M : F(u) > r\} = M - \mathcal{M}_r$$

is relatively open (on M) iff \mathcal{M}_r is closed (cf. Problem 1.12d). Thus, F is lower semicontinuous on M iff $\mathcal{N}_r(F)$ is relatively *open* for all $r \in \mathbb{R}$. Using all the possible decompositions $r = \alpha + \beta$, we get

$$\mathcal{N}_r(F + G) = \bigcup_{\alpha + \beta = r} (\mathcal{N}_\alpha(F) \cap \mathcal{N}_\beta(G)).$$

By assumption, $\mathcal{N}_r(F)$ and $\mathcal{N}_r(G)$ are relatively *open* for all $r \in \mathbb{R}$. Since the union of an arbitrary family of relatively open sets is again relatively open, the set $\mathcal{N}_r(F + G)$ is relatively open for all $r \in \mathbb{R}$, and hence $F + G$ is lower semicontinuous on M .

2.7b. *The classical symbols $\underline{\lim}$ and $\overline{\lim}$.* Recall the following facts from calculus. Let (a_n) be a sequence of real numbers. Then, the point $\alpha \in [-\infty, \infty]$ is called a *cluster point* of (a_n) iff there exists a subsequence $(a_{n'})$ such that

$$\alpha = \lim_{n' \rightarrow \infty} a_{n'}.$$

On $[-\infty, \infty]$ there are always a smallest and a largest cluster point of (a_n) , which we denote by

$$\underline{\lim}_{n \rightarrow \infty} a_n \quad \text{and} \quad \overline{\lim}_{n \rightarrow \infty} a_n$$

respectively. Instead of $\underline{\lim}$ and $\overline{\lim}$, one also uses the notations \liminf and \limsup , respectively. For example, if $a_n = (-1)^n$, then

$$\underline{\lim}_{n \rightarrow \infty} a_n = -1 \quad \text{and} \quad \overline{\lim}_{n \rightarrow \infty} a_n = 1.$$

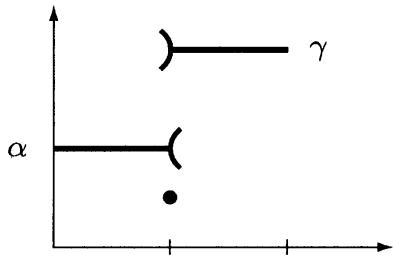
Let (a_n) and (b_n) be two sequences of real numbers. Show that

$$\underline{\lim}_{n \rightarrow \infty} a_n + \underline{\lim}_{n \rightarrow \infty} b_n \leq \underline{\lim}_{n \rightarrow \infty} (a_n + b_n),$$

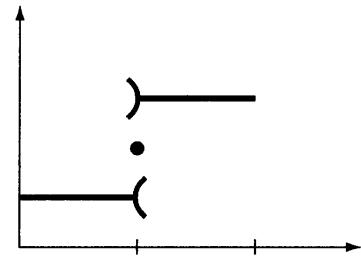
provided the left-hand side does not correspond to the meaningless expressions “ $\infty - \infty$ ” or “ $-\infty + \infty$.”

2.8. Sequentially lower semicontinuous functionals. Let X be a real normed space. The functional $F: M \subseteq X \rightarrow \mathbb{R}$ is called *sequentially lower semicontinuous* at the point $u \in M$ iff

$$F(u) \leq \underline{\lim}_{n \rightarrow \infty} F(u_n)$$



(a) lower semicontinuous



(b) not lower semicontinuous

FIGURE 2.10.

for each sequence (u_n) in M with $u_n \rightarrow u$ as $n \rightarrow \infty$.

Furthermore, $F: M \subseteq X \rightarrow \mathbb{R}$ is called sequentially lower semicontinuous iff it is sequentially lower continuous at each point of M .

Show that

- (i) If $F, G: M \subseteq X \rightarrow \mathbb{R}$ are sequentially lower semicontinuous at the point $u \in M$, then so is $F + G$.
- (ii) Let M be a closed set. Then, the functional $F: M \subseteq X \rightarrow \mathbb{R}$ is lower semicontinuous on M iff it is sequentially lower semicontinuous.
- (iii) Let M be a closed set and let $\dim X < \infty$ (e.g., $X = \mathbb{R}$). Then, the following three statements are mutually equivalent for the functional $F: M \subseteq X \rightarrow \mathbb{R}$:
 - (a) F is lower semicontinuous.
 - (b) F is sequentially lower semicontinuous.
 - (c) F is weakly sequentially lower semicontinuous.
- (iv) Let $X := \mathbb{R}$ and $M := [0, 2]$. For $\alpha \leq \gamma$, define

$$F(u) := \begin{cases} \alpha & \text{if } 0 \leq u < 1 \\ \beta & \text{if } u = 1 \\ \gamma & \text{if } 1 < u \leq 2. \end{cases}$$

Then, $F: [0, 2] \rightarrow \mathbb{R}$ is *sequentially lower semicontinuous* at the point $u = 1$ iff $\beta \leq \alpha$ (cf. Figure 2.10).

Moreover, F is *lower semicontinuous* on $[0, 2]$ iff $\beta \leq \alpha$.

By (iii), F is weakly sequentially lower semicontinuous on $[0, 2]$ iff it is lower semicontinuous on $[0, 2]$.

Finally, F is *continuous* on $[0, 2]$ iff $\alpha = \beta = \gamma$.

Hints: Ad (i). Use Problem 2.8.

Ad (ii). Cf. Zeidler (1986), Vol. 3, p. 165.

Ad (iii). Weak and strong convergence coincide if $\dim X < \infty$.

2.9. *Applications to the famous N-body problem.* Let us describe the motion of the sun and of $N - 1$ planets through

$$\mathbf{x} = \mathbf{x}_j(t), \quad j = 1, \dots, N,$$

where $\mathbf{x}_j(t)$ denotes the position vector of the j th body at time t . The motion of these bodies is governed by the Newtonian equations:

$$m_j \mathbf{x}'' = \mathbf{K}_j, \quad j = 1, \dots, N, \quad (126)$$

where the force is given through

$$\mathbf{K}_j = \sum_{k=1, k \neq j}^N \frac{\gamma m_j m_k (\mathbf{x}_k - \mathbf{x}_j)}{|\mathbf{x}_k - \mathbf{x}_j|^3}.$$

Here, m_j denotes the mass of the j th body, and γ denotes the gravitational constant.

Existence theorem: Let $T > 0$. Then there exist infinitely many T -periodic noncollision solutions.

2.9a. Show that equation (126) is the Euler equation to the following classical variational problem:

$$\int_a^b L(\mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{x}'_1, \dots, \mathbf{x}'_N) dt = \text{stationary!},$$

$$\mathbf{x}_j(a) = \text{given}, \quad \mathbf{x}_j(b) = \text{given}, \quad j = 1, \dots, N,$$

where the Lagrangian is given by

$$L = \text{kinetic energy} - \text{potential energy}.$$

That is,

$$L = \sum_{j=1}^N \frac{1}{2} m_j \mathbf{x}'_j^2 - V_j,$$

along with $\mathbf{K}_j = -\text{grad}_{\mathbf{x}_j} V_j$ and

$$V_j = \sum_{k=1, k \neq j}^N -\frac{\gamma m_j m_k}{|\mathbf{x}_j - \mathbf{x}_k|}.$$

2.9b.* Study the proof of the existence theorem in the monograph by Ambrosetti and Coti-Zelati (1993). The proof is based on modern variational methods (e.g., the Ljusternik–Schnirelman theory, the Morse theory,

and the mountain pass theorem). Observe that the force \mathbf{K}_j becomes singular for $\mathbf{x}_j = \mathbf{x}_k$, which corresponds to a collision. This complicates the proof considerably.

As an introduction to the N -body problem we recommend Meyer and Hall (1992).

2.10. Nonlinear elasticity. Let G be a nonempty, bounded, open, connected set in \mathbb{R}^3 that has a sufficiently smooth boundary. We want to describe the *deformation* of an elastic body by the vector equation

$$\mathbf{y} = \mathbf{x} + \mathbf{u}(\mathbf{x}). \quad (127)$$

Here, \mathbf{x} denotes the position vector of a point in the undeformed region G . Under a deformation, the position vector \mathbf{x} is transformed into the new position vector \mathbf{y} (cf. Figure 2.11). The basic variational problem in nonlinear elasticity reads as follows:

$$\int_G L(\mathcal{E}(\mathbf{u}'), \mathbf{x}) d\mathbf{x} - \int_G \mathbf{K}\mathbf{u} d\mathbf{x} = \min !, \quad (128)$$

$$\mathbf{u} = \mathbf{u}_0 \quad \text{on } \partial G.$$

This is the principle of minimal stored energy. We use the following notation:

- \mathbf{u} = deformation vector,
- L = stored energy function,
- \mathbf{K} = density of the outer forces,
- \mathcal{E} = deformation tensor,
- σ = stress tensor (first Piola–Kirchhoff stress tensor).

The deformation \mathbf{u}_0 of the boundary is given. We are looking for the deformation \mathbf{u} of the body. The terms appearing in (128) possess the following physical meaning:

$$\begin{aligned} \int_G L(\mathcal{E}(\mathbf{u}'), \mathbf{x}) d\mathbf{x} &= \text{elastic energy of the body stored} \\ &\quad \text{by the deformation,} \\ - \int_G \mathbf{K}\mathbf{u} d\mathbf{x} &= -(\text{work done by the outer forces}) \\ &= \text{stored potential energy.} \end{aligned}$$

Let us use a Cartesian coordinate system with the orthonormal basic vectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$. In what follows we sum over two equal indices from 1 to 3. Set

$$\mathbf{x} = x_j \mathbf{e}_j, \quad \mathbf{u} = u_j \mathbf{e}_j, \quad \mathbf{K} = K_j \mathbf{e}_j,$$

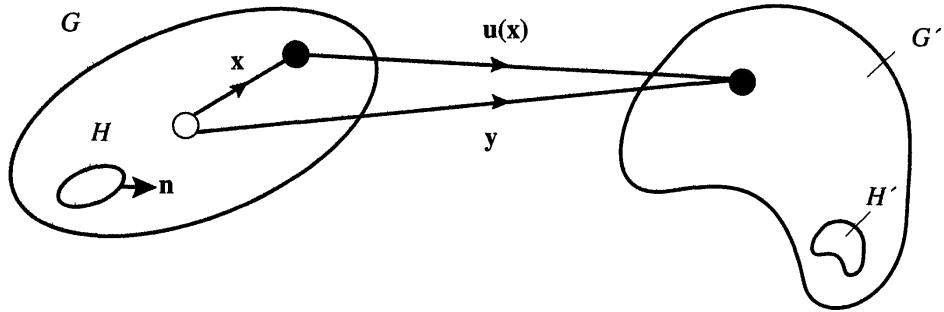


FIGURE 2.11.

and $\partial_j = \partial/\partial x_j$. Then

$$\mathcal{E} = \frac{1}{2}(\mathbf{u}'(\mathbf{x}) + \mathbf{u}'(\mathbf{x})^* + \mathbf{u}'(\mathbf{x})^* \mathbf{u}'(\mathbf{x})).$$

Explicitly,

$$\mathcal{E} = \mathcal{E}_{ij} \mathbf{e}_i \otimes \mathbf{e}_j,$$

where

$$\mathcal{E}_{ij} = \frac{1}{2}(\partial_i u_j + \partial_j u_i + \partial_i u_k \partial_j u_k).$$

Using components, the variational problem (128) reads as follows:

$$\begin{aligned} \int_G L(\mathcal{E}, x) dx - \int_G K_j u_j dx &= \text{stationary!}, \\ u_j &= u_{j0} \quad \text{on } \partial G, \end{aligned} \tag{128*}$$

where $\mathcal{E} = (\mathcal{E}_{ij})$.

2.10a. The equilibrium equation. Show that each sufficiently smooth solution to the principle (128) of stationary energy satisfies the following Euler–Lagrange equation:

$$\begin{aligned} \operatorname{div} \sigma &= \mathbf{K} && \text{on } G, \\ \mathbf{u} &= \mathbf{u}_0, && \text{on } \partial G. \end{aligned} \tag{129}$$

Explicitly,

$$\begin{aligned} \partial_i \sigma_{ij} &= K_j && \text{on } G, \\ u_j &= u_{j0} && \text{on } \partial G, \quad j = 1, 2, 3, \end{aligned} \tag{129*}$$

where

$$\sigma_{ij} := \frac{\partial L}{\partial(\partial_i u_j)}.$$

Solution: Use Remark 4 in Section 2.2.

Remark: Equation (129) is called the *equilibrium condition*. It describes the balance between the outer forces and the stress forces. Let H be a sufficiently regular subregion of G , where \mathbf{n} denotes the outer unit normal vector to a boundary point of H . Under the deformation (127), the subregion H is transformed onto the deformed subregion H' . Then

$$\int_H \mathbf{K} dx = \text{outer force acting on the deformed subregion } H',$$

$$-\int_H (\operatorname{div} \boldsymbol{\sigma}) dx = \text{stress force acting on } H'.$$

Integration by parts yields

$$\int_H (\operatorname{div} \boldsymbol{\sigma}) dx = \int_{\partial H} (\boldsymbol{\sigma} \mathbf{n}) dS,$$

where $\boldsymbol{\sigma} \mathbf{n} = (\sigma_{ij} n_j) \mathbf{e}_i$.

A detailed discussion of both the physical and the mathematical background can be found in Zeidler (1986), Vol. 4, Chapter 61.

2.10b. Convex approximation models. Since it is difficult to solve the highly nonlinear equilibrium equations (129), physicists and engineers consider approximation models. To this end, they replace the nonlinear deformation tensor \mathcal{E} with the linear approximation

$$\gamma_{ij} = \frac{1}{2}(\partial_i u_j + \partial_j u_i).$$

If the stored energy function L is now convex with respect to the first partial derivatives $\partial_i u_j$, then we can apply the existence theorem from Section 2.6.

- (i) Generalize Proposition 1 from Section 2.6 to problem (128).
- (ii) In addition, study such approximation models along with a duality theory in Zeidler (1986), Vol. 4, Chapter 62.

2.10c.* Polyconvex material. It turns out that convex models, as considered in Problem 2.10b, are never rigorous models. In 1977 John Ball introduced a class of rigorous models in nonlinear elasticity based on the notion of polyconvexity.

Theorem: *Problem (128) has a solution*

$$u_j \in W_p^1(G), \quad j = 1, 2, 3, \quad 2 \leq p < \infty,$$

where

$$\det(I + u'(x)) > 0 \quad \text{for almost all } x \in G,$$

provided the following assumptions are satisfied:

- (H1) *Polyconvexity.* The stored energy function L is polyconvex, that is, we have

$$L = P(A, \text{adj } A, \det A),$$

where

$$A := I + u'(x),$$

and P is a convex continuous function of the three arguments A , $\text{adj } A$, and $\det A > 0$. Explicitly, $A = (a_{ij})$ is a real (3×3) -matrix with

$$a_{ij} := \delta_{ij} + \partial_i u_j.$$

Furthermore, $\text{adj } A := (\det A)A^{-1}$. Denote the space of all real (3×3) -matrices by $M(3, 3)$. Polyconvexity means that

$$\begin{aligned} P(tA + (1-t)B, t \text{adj } A + (1-t)\text{adj } B, t \det A + (1-t) \det B) \\ \leq tP(A, \text{adj } A, \det A) + (1-t)P(B, \text{adj } B, \det B) \end{aligned}$$

for all $t \in [0, 1]$ and all $A, B \in M(3, 3)$ with $\det A, \det B > 0$.

- (H2) *Coerciveness.* There is a constant $c > 0$ such that

$$P(A, \text{adj } A, \det A) \geq c(|A|^p + |\text{adj } A|^r + |\det A|^s) + \text{const}$$

for all $A \in M(3, 3)$ with $\det A > 0$. Here, $|A| := \max_{ij} |a_{ij}|$, and

$$2 \leq p < \infty, \quad \frac{1}{p} + \frac{1}{q} = 1, \quad q \leq r < \infty, \quad 1 < s < \infty.$$

- (H3) *The limit $\det A \rightarrow 0$.* We have

$$\lim_{n \rightarrow \infty} P(A_n, B_n, d_n) = +\infty$$

if $|A_n - A| + |B_n - B| \rightarrow 0$ as $n \rightarrow \infty$ in $M(3, 3)$ and $d_n \rightarrow d$ as $n \rightarrow \infty$ in \mathbb{R} .

- (H4) *Boundary displacement.* Let \mathbf{u}_0 be a given C^1 -vector field on \overline{G} such that

$$\det(I + u'_0(x)) > 0 \quad \text{on } \overline{G}.$$

- (H5) *Outer forces.* Let $K_j \in L_q(G)$ for all j .

For $q = 2$, the spaces $L_q(G)$ and $W_q^1(G)$ were introduced in Chapter 2 of AMS Vol. 108. The general definition for $q \geq 1$ can be found in Problems 5.9 and 5.12.

Hint: Study the proof in Zeidler (1986), Vol. 4, Section 62.13. The point is that, surprisingly enough, $\det A$ and $\text{adj } A$ possess nice properties with respect to weak convergence in Sobolev spaces.

For example, rubberlike material is polyconvex (cf. Zeidler (1986), Vol. 4, Section 61.8).

2.11. A special Lagrange multiplier rule. Consider the minimum problem

$$\begin{aligned} f(u) &= \min!, & u \in X, \\ \Gamma(u) &= \alpha & \text{(side condition)} \end{aligned} \quad (130)$$

for fixed real $\alpha \neq 0$. Assume the following:

- (H1) $f: X \rightarrow \mathbb{R}$ is a functional on the real Banach space X .
- (H2) The first variation $\delta f(u; h)$ exists for all $u, h \in X$, and the map $h \mapsto \delta f(u; h)$ is linear on X for each $u \in X$.
- (H3) The functional $\Gamma: X \rightarrow \mathbb{R}$ is linear and continuous.

Show that if u_0 is a solution to problem (130), then there exists a real number λ such that

$$\begin{aligned} \delta f(u_0; h) + \lambda \delta \Gamma(u_0; h) &= 0 & \text{for all } h \in X, \\ \Gamma(u_0) &= \alpha. \end{aligned} \quad (131)$$

The number λ is called a Lagrange multiplier.

Note that the multiplier λ is uniquely determined by (131). In fact, since $\delta \Gamma(u_0; h) = \Gamma(h) = \alpha$ for all $h \in X$, we obtain $\delta f(u_0; u_0) + \lambda \Gamma(u_0) = 0$. Hence

$$\lambda = -\alpha^{-1} \delta f(u_0; u_0).$$

In particular, if the Gâteaux-derivative $f'(u_0)$ exists, then (131) is equivalent to the following condition:

$$\begin{aligned} f'(u_0) + \lambda \Gamma'(u_0) &= 0, \\ \Gamma(u_0) &= \alpha. \end{aligned} \quad (131^*)$$

This is a special case of the general Lagrange multiplier rule to be considered in Section 4.14.

Solution: Define

$$Ph := h - \alpha^{-1} \Gamma(h) u_0 \quad \text{for all } h \in X. \quad (132)$$

Then, $\Gamma(Ph) = 0$ for all $h \in X$. Fix $h \in X$. Set

$$\phi(t) := f(u_0 + tPh) \quad \text{for all } t \in \mathbb{R}.$$

Since $\Gamma(u_0 + tPh) = \alpha$ for all $h \in X$, the function ϕ has a minimum at the point $t = 0$. This implies $\phi'(0) = 0$. Hence

$$\phi'(0) = \delta f(u_0; Ph) = 0 \quad \text{for all } h \in X.$$

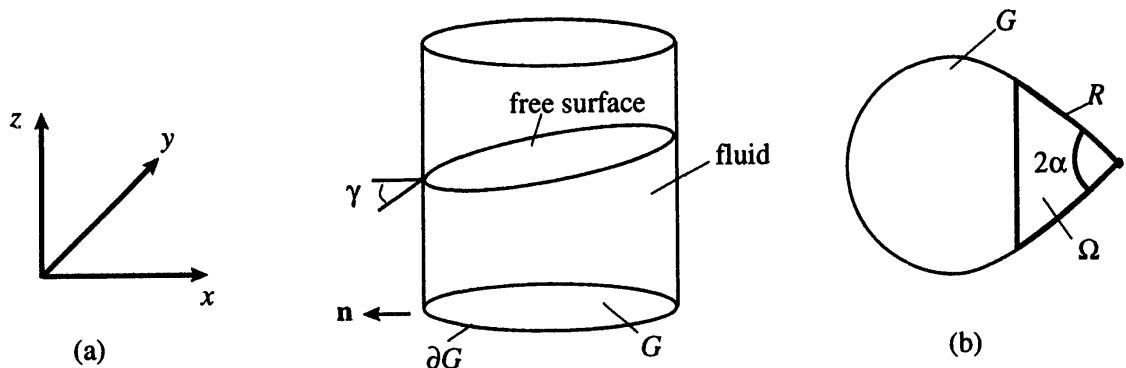


FIGURE 2.12.

By (132),

$$\delta f(u_0; h) - \alpha^{-1} \delta f(u_0; u_0) \Gamma(h) = 0 \quad \text{for all } h \in X.$$

It follows from (H3) that $\delta \Gamma(u_0; h) = \Gamma(h)$ for all $h \in X$. Thus, we obtain (131).

2.12. Capillary surfaces, natural boundary conditions, and experiments performed in space shuttles. We want to show how to use the preceding Lagrange multiplier rule in order to obtain important information about an interesting problem in fluid dynamics.

2.12a. The principle of minimal stored energy for capillary surfaces due to Gauss. Let us consider a fluid of constant density in a container under the influence of the gravitational force. We are looking for the shape of the free surface of the fluid (cf. Figure 2.12).

From the physical point of view, two additional forces appear, namely,

(i) the capillary force at the free surface,

(ii) the adhesion force at the wall.

The capillary force is due to the fact that the fluid molecules at the free surface are not completely surrounded by fluid molecules.

Let us use a Cartesian (x, y, z) -coordinate system with the corresponding orthonormal basic vectors \mathbf{i} , \mathbf{j} , and \mathbf{k} . Then the equation of the free surface is given by

$$z = u(x, y) \quad \text{on } G.$$

The fundamental variational problem for determining the free surface reads as follows:

$$\begin{aligned} \sigma \int_G \sqrt{1 + u_x^2 + u_y^2} dx dy - \sigma \beta \int_{\partial G} u ds - \int_G \rho g u dx dy &= \min!, \quad (133) \\ \int_G u dx dy = V &\quad (\text{side condition}). \end{aligned}$$

We are given the volume V of the fluid. The set G is assumed to be nonempty, bounded, open, and connected, where the boundary ∂G is sufficiently regular.

This variational principle corresponds to the principle of minimal stored energy. It resembles the principle (128) of minimal energy in elasticity. The terms appearing in (133) possess the following physical meaning:

$$\begin{aligned} \sigma \int_G \sqrt{1 + u_x^2 + u_y^2} dx dy &= \text{energy stored by the} \\ &\quad \text{deformation of the free surface,} \\ -\sigma \beta \int_{\partial G} u ds &= -(\text{work done by the adhesion forces at the wall}) \\ &= \text{stored potential energy,} \\ \int_G \rho g u dx dy &= -(\text{work done by the gravitational force}) \\ &= \text{stored potential energy.} \end{aligned}$$

Here we start with the simplest possible assumption, which says that the surface energy is proportional to the surface area. Moreover, we assume that the work done by the adhesion forces at the wall is also proportional to the area of the wall wetted by the fluid ($s = \text{arclength of the boundary curve } \partial G$). In addition, we are given the following positive constants:

$$\begin{aligned} \sigma &= \text{surface tension,} \\ \rho &= \text{density of the fluid,} \\ \beta &= \text{relative adhesion coefficient,} \\ g &= \text{gravitational acceleration,} \\ \kappa &= \text{capillary constant } \left(\kappa = \frac{g\rho}{\sigma} \right). \end{aligned}$$

Let

$$\mathbf{T} := \frac{\nabla u}{|\nabla u|}.$$

Show that a sufficiently smooth solution u of the variational problem (133) satisfies the following conditions:

$$\begin{aligned} \operatorname{div} \mathbf{T} &= \kappa u + \lambda && \text{on } G, \\ \mathbf{T} \mathbf{n} &= \beta && \text{on } \partial G, \end{aligned} \quad (134)$$

where \mathbf{n} denotes the outer unit normal vector to the boundary ∂G . Furthermore,

$$\lambda = \beta \frac{\text{length}(\partial G)}{\text{meas}(G)} - \kappa \frac{V}{\text{meas}(G)}, \quad (135)$$

and

$$\cos \gamma = \beta. \quad (136)$$

Here, γ denotes the contact angle between the free surface and the wall (cf. Figure 2.12(a)).

Explicitly, equation (134) reads as follows:

$$\frac{\partial}{\partial x} \left(\frac{u_x}{\sqrt{1 + u_x^2 + u_y^2}} \right) + \frac{\partial}{\partial y} \left(\frac{u_y}{\sqrt{1 + u_x^2 + u_y^2}} \right) = \kappa u + \lambda \quad \text{on } G.$$

Geometrical discussion. The outer unit normal vector \mathbf{N} to the free surface is given by

$$\mathbf{N} = \frac{\mathbf{k} - u_x \mathbf{i} - u_y \mathbf{j}}{\sqrt{1 + u_x^2 + u_y^2}}.$$

Hence

$$\cos \gamma = -\mathbf{n} \cdot \mathbf{N} = \mathbf{T} \cdot \mathbf{n}.$$

By the boundary condition in (134), $\cos \gamma = \beta$. This is (136). The quantity

$$H := \frac{1}{2} \operatorname{div} \mathbf{T} \quad (137)$$

represents the *mean curvature* of the free surface.

Physical discussion. From (136) we obtain the important physical fact that the contact angle γ is constant, and γ depends only on the material constant β , not on the shape of ∂G .

In a space ship, the gravitational force is weak, that is, the gravitational acceleration g is small. If we set $g = 0$, then $\kappa = 0$. In this case, the basic equation (134) is specialized to the following equation for capillary surfaces without gravity:

$$\begin{aligned} \operatorname{div} \mathbf{T} &= 2H, && \text{on } G, \\ \mathbf{T} \cdot \mathbf{n} &= \beta, && \text{on } \partial G, \end{aligned} \quad (138)$$

and

$$\cos \gamma = \beta, \quad 2H = \beta \frac{\text{length}(\partial G)}{\text{meas}(G)}.$$

Consequently, capillary surfaces without gravity are surfaces of constant mean curvature that have a constant contact angle.

Since the gravitational force in a space ship is weak, the capillary force plays a fundamental role for handling fluids like fuel. Recently, NASA performed experiments on a space shuttle based on the mathematical theory of capillary surfaces. Further experiments are planned.

Historical remark. The differential equation (138) along with the corresponding boundary condition date back to papers written by Young in 1805 and Laplace in 1806. They used ingenious heuristic arguments that have become standard in the engineering literature. The rigorous approach given here is based on a method Gauss proposed in 1830.

Solution: We will use the Lagrange multiplier rule from Problem 2.11. Set

$$f(u) := \sigma \int_G \sqrt{1 + u_x^2 + u_y^2} dx dy - \sigma \beta \int_{\partial G} u ds + \int_G \rho g u dx dy$$

and

$$\Gamma(u) := \int_G u dx dy.$$

Let $X := C^1(\bar{G})$. Choose $h \in X$ and $t \in \mathbb{R}$. Recall that

$$\delta f(u; h) := \left. \frac{df(u + th)}{dt} \right|_{t=0}.$$

Then

$$\delta f(u; h) = \sigma \int_G \frac{u_x h_x + u_y h_y}{\sqrt{1 + u_x^2 + u_y^2}} dx dy - \sigma \beta \int_{\partial G} h ds + \int_G \rho g h dx dy,$$

and

$$\delta \Gamma(u, h) = \int_G h dx dy.$$

By Problem 2.12a, there is a real number λ such that

$$\delta f(u; h) + \sigma \lambda \delta \Gamma(u; h) = 0 \quad \text{for all } h \in C^1(\bar{G}).$$

Integration by parts yields

$$\int_G (\sigma \lambda + \rho g u - \sigma \operatorname{div} \mathbf{T}) h dx dy + \sigma \int_{\partial G} (\mathbf{T} \mathbf{n} - \beta) h ds = 0 \quad (139)$$

for all $h \in C^1(\bar{G})$.

(a) From (139) we obtain

$$\int_G (\sigma \lambda + \rho g u - \sigma \operatorname{div} \mathbf{T}) h dx dy = 0 \quad \text{for all } h \in C_0^\infty(G).$$

Hence

$$\sigma \lambda + \rho g u - \sigma \operatorname{div} \mathbf{T} = 0 \quad \text{on } G. \quad (140)$$

(b) By (139),

$$\int_G (\mathbf{T} \mathbf{n} - \beta) h ds = 0 \quad \text{for all } h \in C^1(\bar{G}). \quad (141)$$

If the boundary ∂G is sufficiently regular, then $C^1(\overline{G})$ is dense in $L_2(\partial G)$ (cf. Zeidler (1986), Vol. 2A, Section 21.3). Therefore, it follows from (141) that

$$\mathbf{T}\mathbf{n} - \beta = 0 \quad \text{on } \partial G.$$

Since this boundary condition follows automatically from the variational principle, it is called a *natural boundary condition*.

(c) Integrating equation (140) over G and using integration by parts, we obtain

$$\sigma\lambda \int_G dx dy + \int_G \rho g u dx dy - \sigma \int_{\partial G} \mathbf{T}\mathbf{n} ds = 0.$$

Hence

$$\sigma\lambda \text{meas}(G) + \rho g V - \sigma\beta \text{length}(\partial G) = 0.$$

2.12b. Regions that have a corner. Suppose that the region G contains a corner with interior angle 2α . Show the following:

- (i) If a (sufficiently regular) capillary surface without gravity exists, then the angle α must be sufficiently large, that is,

$$\alpha \geq \frac{\pi}{2} - \gamma, \quad (142)$$

where γ denotes the contact angle.

- (ii) For each angle α that satisfies condition (142), there exists a capillary surface without gravity provided the boundary of G is a polygon.

This theorem was proved by Concus and Finn in 1974.

Solution: Ad (i). Consider a situation as pictured in Figure 2.12(b). We are given a capillary surface $z = u(x, y)$ that satisfies equation (138). Integration over the subregion Ω yields

$$\int_{\Omega} \operatorname{div} \mathbf{T} dx dy = 2H \int_{\Omega} dx dy.$$

If we use integration by parts, then we obtain

$$\int_{\partial\Omega} \mathbf{T}\mathbf{n} ds = 2H \text{meas}(\Omega). \quad (143)$$

Set $\partial_1\Omega := \partial G \cap \partial\Omega$. Then, $\partial\Omega = \partial_1\Omega \cup \partial_2\Omega$. By (138),

$$\mathbf{T}\mathbf{n} = \beta \quad \text{on } \partial_1\Omega,$$

where $\beta = \cos\gamma$. Moreover, $|\mathbf{T}| < 1$. Hence

$$\mathbf{T}\mathbf{n} < 0 \quad \text{and} \quad |\mathbf{T}\mathbf{n}| < 1 \quad \text{on } \partial_2\Omega.$$

Thus, from (143) we obtain the key inequality

$$2H \operatorname{meas}(\Omega) > (\cos \gamma) \operatorname{length}(\partial_1 \Omega) - \operatorname{length}(\partial \Omega_2). \quad (144)$$

By Figure 2.12(b),

$$\operatorname{length}(\partial_1 \Omega) = 2R, \quad \operatorname{length}(\partial_2 \Omega) = 2R \sin \alpha, \quad \operatorname{meas}(\Omega) = \text{const.}$$

Dividing relation (144) by R and letting $R \rightarrow 0$, we get

$$\cos \gamma \leq \sin \alpha.$$

This implies $\sin(\frac{\pi}{2} - \gamma) \leq \sin \alpha$. Hence, $\frac{\pi}{2} - \gamma \leq \alpha$.

Ad (ii). Use spherical caps as free surfaces (cf. Finn (1984), p. 136).

The monograph by Finn (1984) is a modern standard text on capillary surfaces.

2.13. *The Dirichlet principle and existence theorems for minimal surfaces.* Let us discuss some important results concerning minimal surfaces. This is still a very active area of research. The minimal surface problem (or the Plateau problem) reads as follows: For a given boundary curve γ , we are looking for a minimal surface spanned through γ .

By definition, a surface is called a minimal surface iff the *mean curvature vanishes* at each interior point. For example, if the smooth surface is described by the equation $z = u(x, y)$, then it is a minimal surface iff it is a solution of the partial differential equation given in Problem 2.3.

Each minimal surface locally represents a surface of least area.

2.13a. *Conformal coordinates.* The introduction of such coordinates elegantly reduces the minimal surface problem to the Laplacian. Let us consider a Cartesian coordinate system in \mathbb{R}^3 with the orthonormal vectors \mathbf{i} , \mathbf{j} , and \mathbf{k} . Set $\mathbf{x} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$. We are looking for a minimal surface

$$\mathbf{x} = \mathbf{x}(v, w), \quad (v, w) \in \overline{D},$$

where $D := \{(v, w) : v^2 + w^2 < 1\}$ denotes the unit disk. The parameters v, w are supposed to be conformal, that is, the partial derivatives \mathbf{x}_v and \mathbf{x}_w satisfy the two conditions

$$\mathbf{x}_v^2 = \mathbf{x}_w^2 \quad \text{and} \quad \mathbf{x}_v \cdot \mathbf{x}_w = 0 \quad \text{on } D. \quad (145a)$$

If we use conformal coordinates, then the vanishing of mean curvature means that

$$\Delta \mathbf{x} = 0 \quad \text{on } D. \quad (145b)$$

Finally, we have to add the boundary condition:

$$\mathbf{x} = \text{boundary curve } \gamma \quad \text{on } \partial D. \quad (145c)$$

2.13b. *The classical existence theorem for the Plateau problem.* We are given the closed C^1 -Jordan curve γ .²² Then there exists a minimal surface bounded by the curve γ .

More precisely, there exists a continuous map $\mathbf{x}: \overline{D} \rightarrow \mathbb{R}^3$, which is analytic on D , such that the conditions in (145) are satisfied. Moreover, the map $\mathbf{x}: \partial D \rightarrow \gamma$ is a homeomorphism.

Hint: A fairly elementary proof of this theorem (essentially due to Courant) can be found in the lecture notes by Jost (1994). The idea of the proof is to use the following variational problem:

$$\int_D (\mathbf{x}_v^2 + \mathbf{x}_w^2) dv dw = \min!, \quad (146)$$

\mathbf{x} = boundary curve γ on ∂D .

Observe that the Euler–Lagrange equation to (146) is precisely (145b). First one constructs a minimal sequence (\mathbf{x}_n) to (146). The point is then to show that this sequence converges to a solution to (145). To this end, one uses the famous Courant–Lebesgue lemma and the Arzelà–Ascoli theorem.

Study this proof in Jost (1994). More general results can be found in Dierkes, Hildebrandt, Küster, and Wohlrab (1992), Vol. 1.

2.13c.* Regularity up to the boundary.²³ If the boundary curve γ is $C^{m,\alpha}$ for fixed $m = 1, 2, \dots$, and $0 < \alpha < 1$, then the map $\mathbf{x}: \overline{D} \rightarrow \mathbb{R}^3$ from Problem 2.13b is also $C^{m,\alpha}$.

Hint: Cf. Dierkes, Hildebrandt, Küster, and Wohlrab (1992), Vol. 2, Chapter 7. The proof is based on sophisticated properties of harmonic functions and on differential inequalities.

2.13d.* Generic finiteness. For most boundary curves γ , the number of minimal surfaces spanned through γ is finite.

Hint: This is a typical modern result based on the Sard–Smale theorem (cf. Section 5.15). A detailed study can be found in Tromba (1977).

Historical remarks. Lagrange formulated the minimal surface equation in 1762 (cf. Problem 2.3). In the nineteenth century, the physicist Plateau performed numerous soap film experiments. Douglas and Radó in 1930 independently proved the general existence theorem from Problem 2.13b. Douglas also proved existence theorems in the case where a finite number of boundary curves is given. For his research on minimal surfaces, Jesse Douglas was awarded the first Fields medal at the International Mathematical Congress held in Oslo in 1936.

²²Recall that a closed Jordan curve is the image $\psi(\partial D)$ of a homeomorphism $\psi: \partial D \rightarrow \mathbb{R}^3$. In addition, we assume that the map ψ is C^1 . Roughly speaking, reasonable smooth closed curves are of this type.

²³The class $C^{m,\alpha}$ of Hölder spaces was introduced in Problem 1.8 of AMS Vol. 108.

The boundary theorem from Problem 2.13c was first proved by Hildebrandt in 1966 (for $m \geq 4$). Böhme and Tromba obtained the generic finiteness theorem in 1977.

The Plateau problem in higher dimensions. In order to solve the minimal surface problem in higher dimensions, according to Giorgi, one uses the concept of *generalized surface area* based on functions of bounded variations. This leads to extremely weak solutions. The main task is then to prove the regularity or the partial regularity of those very weak solutions. This can be found in the monograph by Giusti (1984). The main ideas are explained in Zeidler (1986), Vol. 2B, pp. 1114ff.

Standard texts on minimal surfaces include the two monographs by Dierkes, Hildebrandt, Küster, and Wohlrab (1992), Vols. 1, 2, and that by Nitsche (1992). We also recommend the monograph by Struwe (1988).

The minimal surface problem demonstrates how a simple question arising from our real world leads to the invention of sophisticated techniques in mathematics.

Harmonic maps. The solution $\mathbf{x} = \mathbf{x}(v, w)$ of the variational problem (146) represents a special case of a harmonic map. As an introduction to the modern theory of harmonic map, we recommend the monograph by Jost (1984). Many problems in physics that are governed by the principle of minimal stored energy are described by harmonic maps (e.g., liquid crystals²⁴). In Problem 2.14 we will consider the Landau–Ginzburg model in superconductivity and superfluidity.

2.14. Singular variational problems, phase transitions, and the Landau–Ginzburg model in superconductivity, and superfluidity. Let G be a nonempty bounded open set in \mathbb{R}^2 . Set $S^1 := \{z \in \mathbb{C}: |z| = 1\}$.

2.14a. Harmonic maps from G to \mathbb{C} . Let us consider a map

$$\psi: \overline{G} \rightarrow \mathbb{C},$$

where $\psi(x) = u(x) + v(x)i$, and $x = (\xi, \eta)$. Set $|\nabla\psi|^2 := |\psi_\xi|^2 + |\psi_\eta|^2$. Hence

$$|\psi|^2 = u^2 + v^2, \quad \text{and} \quad |\nabla\psi|^2 = u_\xi^2 + u_\eta^2 + v_\xi^2 + v_\eta^2.$$

Show that each sufficiently smooth solution to the variational problem

$$\begin{aligned} \int_G |\nabla\psi|^2 dx &= \min!, \\ \psi &= g \quad \text{on } G \end{aligned} \tag{147}$$

satisfies the Euler–Lagrange equation

$$\begin{aligned} \Delta\psi &= 0 && \text{on } G, \\ \psi &= g && \text{on } \partial G. \end{aligned} \tag{148}$$

²⁴As an introduction to the modern mathematical theory of liquid crystals, we recommend the proceedings edited by Ericksen and Kinderlehrer (1987).

Each solution $\psi: G \rightarrow \mathbb{C}$ of the first equation in (148) is called a harmonic map from G to \mathbb{C} .

2.14b. *The Landau–Ginzburg model.* Let G be the interior of a closed C^∞ -Jordan curve \mathcal{C} in \mathbb{R}^2 . We are given the C^∞ -map $g: \mathcal{C} \rightarrow \mathbb{C}$, where $|g(x)| = 1$ on \mathcal{C} . Let $d > 0$ denote the winding number of g . That is, if \mathcal{C} is surrounded once counterclockwise, then the origin is surrounded d times counterclockwise by the image curve $g(\mathcal{C})$. Let $\varepsilon > 0$. Consider the following problem of minimal stored energy:

$$\int_G |\nabla \psi_\varepsilon|^2 dx + \frac{1}{2\varepsilon^2} \int_G (1 - |\psi_\varepsilon|^2)^2 dx = \min !, \quad (149)$$

$$\psi_\varepsilon = g \quad \text{on } \partial G,$$

$$\operatorname{Re} \psi_\varepsilon, \quad \operatorname{Im} \psi_\varepsilon \in W_2^1(G).$$

Show that each sufficiently smooth solution of (131) satisfies the Landau–Ginzburg equation:

$$\begin{aligned} -\Delta \psi_\varepsilon &= \frac{1}{\varepsilon^2} \psi_\varepsilon (1 - |\psi_\varepsilon|^2) && \text{on } G, \\ \psi_\varepsilon &= g && \text{on } \partial G. \end{aligned} \quad (150)$$

2.14c.** *The delicate limiting process $\varepsilon \rightarrow 0$.* The following conditions are met:

- (i) For each $\varepsilon > 0$, the variational problem (149) has a smooth solution ψ_ε , which satisfies equation (150). In addition, $|\psi_\varepsilon(x)| \leq 1$ on \overline{G} . If ε is sufficiently smooth (i.e., $0 < \varepsilon < \varepsilon_0$), then ψ_ε has precisely d zeros on G . These zeros possess the index one.²⁵
- (ii) There exist precisely d points P_1, \dots, P_d in G and a sequence, $\varepsilon_n \rightarrow 0$ as $n \rightarrow \infty$, such that the limit

$$\lim_{n \rightarrow \infty} \psi_{\varepsilon_n}(x) = \psi(x)$$

exists for all points $x \in \overline{G} - \{P_1, \dots, P_d\}$, uniformly on each compact subset of $\overline{G} - \{P_1, \dots, P_d\}$.

- (iii) The function ψ satisfies the equation

$$\begin{aligned} -\Delta \psi(x) &= 0 \quad \text{and} \quad |\psi(x)| = 1 && \text{on } G - \{P_1, \dots, P_d\}, \\ \psi &= g && \text{on } \partial G. \end{aligned}$$

²⁵By definition, the index of an isolated zero P_j of ψ is the winding number of ψ with respect to a sufficiently small circle centered at the point P_j . This definition does not depend on the choice of the circle (cf. the mapping degree in Zeidler (1986), Vol. 1, Chapter 12).

- (iv) Let us write the complex number $z = \xi + \eta i$ instead of x , and let z_j correspond to the point P_j . Then, ψ behaves like

$$a_j \frac{z - z_j}{|z - z_j|}$$

near the singular point z_j , where a_j is a complex constant with $|a_j| = 1$. More precisely,

$$\left| \psi(z) - a_j \frac{z - z_j}{|z - z_j|} \right| \leq \text{const} |z - z_j|^2 \quad \text{as } z \rightarrow z_j.$$

Thus, the limit function ψ is a harmonic map from $G - \{P_1, \dots, P_d\}$ onto the unit circle S^1 that has singularities at the points P_1, \dots, P_d .

Hint: The proofs along with further information about the computation of the singular points can be found in the monograph by Bethuel, Brézis, and Hélein (1994).

Physical interpretation. For a superconductor, there is a critical (low) absolute temperature T_c such that, for absolute temperatures T ,

$$0 < T < T_c,$$

superconducting regions appear. These regions correspond to supercurrents of electrons. The passage from normal conductivity to superconductivity is called a *phase transition*.

The complex function $\psi = \psi(x)$ is called an *order parameter* (or a Higgs field²⁶). By definition,

$$|\psi(x)| := \text{density of superconducting electrons at the point } x.$$

The physical units are normalized such that $0 \leq |\psi(x)| \leq 1$. Let $\delta > 0$ be a sufficiently small number. By definition,

if $1 - \delta < |\psi(x)| \leq 1$, then there is a superconducting state at the point x ;
 if $|\psi(x)| < \delta$, then there is a normal state at the point x .

If we use the representation

$$\psi(x) = \rho(x) e^{iS(x)},$$

²⁶In the standard model of elementary particle physics, the *Higgs field* corresponds to a hypothetical Higgs particle that is responsible for the mass of the gauge particles W^\pm and Z detected in 1983. These gauge particles are responsible for the weak interaction in nature (e.g., the radioactive decay). If we do not introduce the Higgs field, then the gauge particles are massless, contradicting physical experiments.

where $S(x)$ is a real phase factor, then the vector $\text{grad } S(x)$ is proportional to the velocity vector of the supercurrent at the point x .

The Landau–Ginzburg model. The variational problem (149) due to the physicists Landau and Ginzburg represents a highly simplified mathematical model for superconducting or superfluid materials.²⁷ The Landau–Ginzburg term

$$\frac{1}{2\varepsilon^2} \int_G (1 - |\psi_\varepsilon|^2)^2 dx$$

can be regarded as a penalty term. The penalty is maximal for normal states. The minimum problem (149) forces the value $|\psi_\varepsilon(x)|$ to be close to 1 provided the positive parameter ε is sufficiently small. Thus, the penalty term forces the appearance of superconducting states.

Renormalization of energy. The limiting function ψ as $\varepsilon \rightarrow 0$ corresponds to an (idealized) superconducting state on G , where the singular points P_1, \dots, P_d are called *defects*. An infinite amount of energy is concentrated at the defects. That is, we have

$$\int_{U(P_j)} |\nabla \psi|^2 dx = \infty$$

on each small neighborhood $U(P_j)$ of the defect P_j . The appearance of infinite energies is typical for modern physics (e.g., for quantum field theory and elementary particle physics). This phenomenon indicates that the mathematical modeling is wrong. To overcome this serious mathematical difficulty, physicists invented the technique of renormalization near 1950. The idea is to pass from the original infinite energy to a renormalized finite energy by subtracting terms that correspond to the singularities.

From a mathematical viewpoint, it is quite remarkable that one can define a renormalized energy E_{ren} such that the location of the defects is obtained by minimizing E_{ren} with respect to all possible defects (cf. Bethuel, Brézis, and Hélein (1994)).

Quantization on a classical level. It is quite interesting that the number d of defects only depends on a topological invariant of the boundary values (the winding number of g). Thus, d can be regarded as a topological quantum number (on a classical level). Such quantization effects are typical for the behavior of superconductors and superfluids in nature.

Cooper pairs. The Landau–Ginzburg approach to superconductivity represents a purely phenomenological theory. A deeper physical understanding of superconductivity can be gained by using the methods of quantum statistics based on second quantization in quantum field theory. The theory of Bardeen, Cooper, and Schrieffer from 1957 explains superconductivity by means of Cooper pairs consisting of two electrons that have

²⁷In superfluidity (e.g., supercooled Helium), $|\psi(x)|$ represents the density of the superfluid component, and $\text{grad } S(x)$ is proportional to the velocity vector of the superfluid component.

the same energy, but antiparallel spin. Cooper pairs of particles are also responsible for superfluidity (cf. Landau and Lifšic (1988), Vols. 9 and 10; Schrieffer (1964); and Bogoliubov (1967)).

Singular variational problems and phase transitions. The variational problem (149) represents a singular perturbation of the Dirichlet problem (147), where the perturbation is given by the Landau–Ginzburg term. In modern mathematical physics, singular variational problems are used in order to model all kinds of phase transitions in materials (cf. Zeidler (1986), Vol. 5).

The variational approach to free boundary problems. Free boundary problems appear in many fields of physics (e.g., in hydrodynamics, elasticity, or plasticity). For example, the free boundary may correspond to any of the following: the surface of a rotating star, the surface of a water wave, the boundary of a groundwater zone, the boundary of a plasticity zone, or, more generally, the boundary of a phase transition zone (e.g., melting ice). From a mathematical viewpoint, such problems can frequently be formulated as constrained variational problems. Using appropriate Lagrange multipliers, we arrive at unconstrained variational problems with an additional (possibly singular) term (cf. Section 4.14). A detailed study can be found in the monograph by Friedman (1982) (cf. also Zeidler (1986), Vol. 5).

2.15. String theory and the Noether theorem. In what follows we will use physical units such that

$$c = 1 \text{ (velocity of light)}, \quad h = 2\pi \text{ (Planck's quantum of action)}.$$

Let \mathbb{Z} denote the set of integers.

The role of string theory in modern physics will be discussed at the end of this problem.

The basic idea of string theory. Consider a Cartesian (x^1, x^2, x^3) -system Σ that corresponds to an inertial system with time t . Set $x^4 := t$. Then the motion of a point particle is described by the *one-parameter* equation

$$x^j = x^j(\tau), \quad 0 \leq \tau \leq \tau_0, \quad j = 1, \dots, 4.$$

This corresponds to a curve in the four-dimensional space–time manifold. This curve is called a *world line* (cf. Figure 2.13(a)).

In contrast to this, the motion of a closed string is described by the *two-parameter* equation

$$x^j = x^j(\sigma, \tau), \quad 0 \leq \sigma \leq 2\pi, \quad 0 \leq \tau \leq \tau_0, \quad j = 1, \dots, d, \quad (151)$$

along with the periodicity condition

$$x^j(0, \tau) = x^j(2\pi, \tau), \quad 0 \leq \tau \leq \tau_0, \quad j = 1, \dots, d.$$

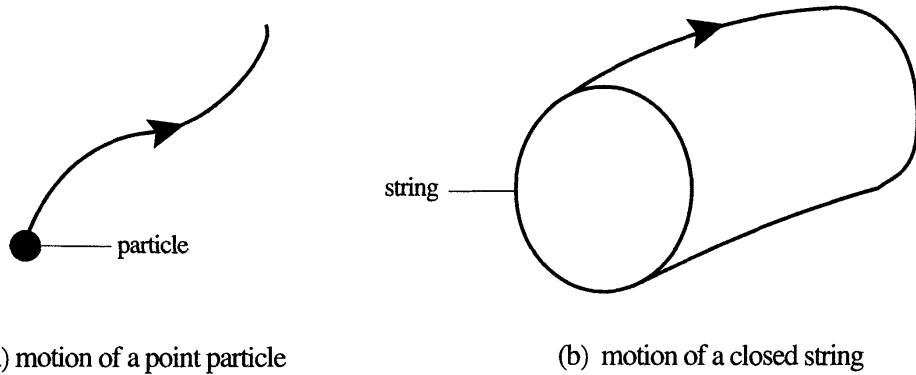


FIGURE 2.13.

Let $d = 4$. Then this corresponds to the motion of a closed curve in the Cartesian coordinate system Σ . Moreover, equation (151) represents a two-dimensional surface in the space-time manifold. This surface is called a *world sheet* (cf. Figure 2.13(b)).

Observe that the appearance of a general d -dimensional space-time manifold in string theory is motivated by the quantum physics of strings.

Notation on the world sheet. Consider the parameter space

$$W := \{(\sigma, \tau) : 0 \leq \sigma \leq 2\pi, 0 \leq \tau \leq \tau_0\}.$$

Set $\sigma^1 := \sigma$, $\sigma^2 := \tau$, and $\partial_\alpha := \partial/\partial\sigma^\alpha$. In what follows we will sum over equal lower and upper Greek indices from 1 to 2. Consider a curve $\sigma^\alpha = \sigma^\alpha(p)$ on W . By definition, the derivative of arclength s with respect to the curve parameter p satisfies the differential equation

$$\left(\frac{ds(p)}{dp} \right)^2 = g_{\alpha\beta}(\sigma(p), \tau(p)) \frac{dx^\alpha(p)}{dp} \frac{dx^\beta(p)}{dp}.$$

Here, we assume that $g_{\alpha\beta} = g_{\beta\alpha}$ on W for all α, β ,

$$g_{22} > 0 \quad \text{and} \quad g := \det(g_{\alpha\beta}) < 0 \quad \text{on } W.$$

Let $(g^{\alpha\beta})$ denote the inverse matrix to $(g_{\alpha\beta})$. Then

$$g_{\alpha\gamma} g^{\gamma\beta} = \delta_\alpha^\beta \quad \text{on } W \quad \text{for all } \alpha, \beta, \tag{152}$$

where

$$\delta_\alpha^\beta := \begin{cases} 1 & \text{if } \alpha = \beta, \\ 0 & \text{if } \alpha \neq \beta. \end{cases}$$

Notation on the space–time manifold \mathbb{R}^d . Let

$$x = (x^1, \dots, x^d), \quad \text{where } x^j \in \mathbb{R} \text{ for all } j.$$

In what follows we sum over equal lower and upper Roman indices from 1 to d . Define

$$xy := x^j x^k \eta_{jk},$$

where

$$\eta^{jk} = \eta_{jk} := \begin{cases} 1 & \text{if } j = k = d, \\ -1 & \text{if } j = k = 1, \dots, d-1, \\ 0 & \text{if } j \neq k. \end{cases}$$

The pseudo-inner product xy on \mathbb{R}^d corresponds to the Minkowski metric on \mathbb{R}^d .

2.15a. *The equation of motion for the bosonic string.* The basic variational principle due to Polyakov reads as follows:

$$-\frac{T}{2} \int_W g^{\gamma\delta} \partial_\gamma x \partial_\delta x \sqrt{-g} d\sigma d\tau = \text{stationary!}, \quad (153)$$

$$x(0, \tau) = x(2\pi, \tau) \quad \text{for all } \tau \in [0, \tau_0].$$

Here, T is a given positive constant called the string tension. We are looking for $x = x(\sigma, \tau)$ and $g^{\alpha\beta} = g^{\alpha\beta}(\sigma, \tau)$. This variational principle tells us that $x: W \rightarrow \mathbb{R}^d$ is a harmonic map with respect to the metrics on W and \mathbb{R}^d .

Show that each sufficiently smooth solution to the variational problem (153) satisfies the following equations:

$$\partial_\alpha (\sqrt{-g} g^{\alpha\beta} \partial_\beta x) = 0 \quad (\text{equation of motion}), \quad (154)$$

$$W_{\alpha\beta} = 0, \quad \alpha, \beta = 1, 2 \quad (\text{constraints}). \quad (155)$$

By definition,

$$W_{\alpha\beta} := \frac{1}{2} \partial_\alpha x \partial_\beta x - \frac{1}{4} g_{\alpha\beta} g^{\gamma\delta} \partial_\gamma x \partial_\delta x.$$

Solution: Introduce the Lagrangian

$$L := -\frac{T}{2} \sqrt{-g} g^{\gamma\delta} \partial_\gamma x \partial_\delta x.$$

Then the Euler–Lagrange equations to (153) read as follows:

$$\partial_\alpha \left(\frac{\partial L}{\partial (\partial_\alpha x^m)} \right) = 0, \quad m = 1, \dots, d \quad (\text{equation of motion}), \quad (156)$$

$$\frac{\partial L}{\partial g^{\alpha\beta}} = 0, \quad \alpha, \beta = 1, 2 \quad (\text{constraints}). \quad (157)$$

We want to show that (156) and (157) correspond to (154) and (155), respectively.

Ad (156). Obviously,

$$\frac{\partial L}{\partial(\partial_\alpha x^m)} = -T \sqrt{-g} (g^{\alpha\delta} \partial_\delta x^k \eta_{km}).$$

Ad (157). Recall that $g = \det(g_{\alpha\beta})$. If $g_{\alpha\beta}$ depends on a parameter p , then a classical formula for the derivative \dot{g} with respect to p tells us that

$$\dot{g} = gg^{\alpha\beta}\dot{g}_{\alpha\beta}.$$

By (152), $g_{\alpha\beta}g^{\alpha\beta} = \delta_\beta^\beta = 2$. Hence

$$\dot{g}_{\alpha\beta}g^{\alpha\beta} + g_{\alpha\beta}\dot{g}^{\alpha\beta} = 0.$$

This implies

$$\dot{g} = -gg_{\alpha\beta}\dot{g}^{\alpha\beta}.$$

Consequently,

$$\frac{\partial g}{\partial g^{\gamma\delta}} = -gg_{\alpha\beta} \frac{\partial g^{\alpha\beta}}{\partial g^{\gamma\delta}} = -gg_{\gamma\delta}.$$

Therefore, equation (157) implies (155).

2.15b. Conservation laws.

Set $P^\alpha := -T \sqrt{-g} g^{\alpha\beta} \partial_\beta x$, $J^\alpha := x \wedge P^\alpha$.

Explicitly,

$$(P^\alpha)^j := -T \sqrt{-g} g^{\alpha\beta} \partial_\beta x^j,$$

$$(J^\alpha)^{jk} := x^j (P^\alpha)^k - x^k (P^\alpha)^j.$$

Use a simple computation in order to show that the equation of motion in (154) implies the following two conservation laws:

$$\partial_\alpha P^\alpha = 0 \quad (\text{energy-momentum conservation}), \tag{158}$$

$$\partial_\alpha J^\alpha = 0 \quad (\text{angular-momentum conservation}). \tag{159}$$

2.15c. Poincaré transformations, the Noether theorem, and conservation laws. Show that the conservation laws (158) and (159) are consequences of the Noether theorem from Section 2.19.

Hint: Ad (158). Use the translation

$$y^j = x^j + a^j. \tag{160}$$

Note that the Lagrangian L is invariant under this transformation.

Ad (159): Let $j, k = 1, \dots, d - 1$, where $j \neq k$. Use both the Lorentz transformation

$$\begin{pmatrix} y^j \\ y^d \end{pmatrix} = \begin{pmatrix} \cosh \psi & \sinh \psi \\ \sinh \psi & \cosh \psi \end{pmatrix} \begin{pmatrix} x^j \\ x^d \end{pmatrix} \quad (161)$$

and the rotation

$$\begin{pmatrix} y^j \\ y^k \end{pmatrix} = \begin{pmatrix} \cos \phi & \sin \phi \\ -\sin \phi & \cos \phi \end{pmatrix} \begin{pmatrix} x^j \\ x^k \end{pmatrix}. \quad (162)$$

Show first that $yy = xx$ for these transformations. Thus, the Lagrangian L remains invariant. Furthermore, observe that these transformations can be written as

$$y^m = x^m + \varepsilon_r^m x^r + o(|\varepsilon|), \quad \text{as } |\varepsilon| \rightarrow 0, \quad (163)$$

where

$$\varepsilon_{mr} = -\varepsilon_{rm} \quad \text{for all } m, r,$$

and $\varepsilon_r^m = \eta^{ms} \varepsilon_{sr}$. In addition, $|\varepsilon| := \max_{m,r} |\varepsilon_r^m|$. To see this, note that, for example,

$$\cosh \psi = 1 + o(\psi), \quad \sinh \psi = \psi + o(\psi), \quad \psi \rightarrow 0.$$

Apply now the Noether theorem to this situation.

Remark: The transformations (161) and (162) generate the Lorentz group. If we add the translations from (160) to the Lorentz group, we obtain the Poincaré group, which plays a fundamental role in relativistic physics.

The same considerations apply to each physical theory that is invariant under Poincaré transformations. Such theories are called relativistically invariant. In this respect, the Noether theorem combined with Poincaré translations in (160) leads to conservation of the energy-momentum tensor, whereas the Lorentz group is responsible for conservation of the angular-momentum tensor.

2.15d. The conformal gauge. Use the following two-dimensional Minkowski metric on the world sheet:

$$(g_{\alpha\beta}) = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Physicists call this choice of metric the *conformal gauge*. Set

$$\partial_{\pm} := \frac{1}{2}(\partial_{\tau} \pm \partial_{\sigma}).$$

Show that the basic equations (154) and (155) now read as follows for all $\sigma, \tau \in \mathbb{R}$:

$$(\partial_{\tau}^2 - \partial_{\sigma}^2)x = 0 \quad (\text{equation of motion}), \quad (164)$$

$$\partial_+ x \partial_+ x = 0, \quad \partial_- x \partial_- x = 0 \quad (\text{Virasoro constraints}), \quad (165)$$

$$x(\sigma + 2\pi, \tau) = x(\sigma, \tau) \quad (\text{periodicity}). \quad (166)$$

Note that equation (164) coincides with the classic equation for a vibrating string (cf. Section 5.12 in AMS Vol. 108).

Solution: Ad (164). This follows immediately from (154).

Ad (165). Equation (155) yields

$$\partial_\tau x \partial_\sigma x = 0 \quad \text{and} \quad \partial_\tau x \partial_\tau x + \partial_\sigma x \partial_\sigma x = 0.$$

This is equivalent to

$$(\partial_\tau \pm \partial_\sigma)x(\partial_\tau \pm \partial_\sigma)x = 0.$$

2.15e. Explicit solutions via Fourier series. Define the Virasoro charges

$$L_m^\pm := T \int_0^{2\pi} d\sigma e^{\pm im\sigma} \partial_\pm x \partial_\pm x, \quad m \in \mathbb{Z}. \quad (167)$$

Show that a solution to the basic equations (164) through (166) is given by

$$\begin{aligned} x = x_0 &+ \frac{\tau - \sigma}{\sqrt{4\pi T}} \alpha_0^- + \frac{i}{\sqrt{4\pi T}} \sum_{n \neq 0} \frac{1}{n} \alpha_n^- e^{-in(\tau - \sigma)} \\ &+ \frac{\tau + \sigma}{\sqrt{4\pi T}} \alpha_0^+ + \frac{i}{\sqrt{4\pi T}} \sum_{n \neq 0} \frac{1}{n} \alpha_n^+ e^{-in(\tau + \sigma)}, \end{aligned} \quad (168)$$

where the Fourier coefficients α satisfy the equations

$$L_m^\pm \equiv \frac{1}{2} \sum_{n=-\infty}^{\infty} \alpha_{m-n}^\pm \alpha_n^\pm = 0, \quad m \in \mathbb{Z}, \quad (169)$$

along with

$$(\alpha_n^\pm)^* = \alpha_n^\pm, \quad n \in \mathbb{Z}. \quad (170)$$

Moroever, $x_0 \in \mathbb{R}^d$ is given.

Here, $\alpha = (\alpha^1, \dots, \alpha^d)$, where the components α^j are complex numbers. The asterisk stands for a passage to the conjugate complex components. In order to obtain classic solutions, assume that

$$|\alpha_n^\pm| \leq \frac{\text{const}}{n^4}, \quad n = \pm 1, \pm 2, \dots$$

This condition allow us to differentiate the Fourier series in (168) twice.

Solution: An explicit computation shows that the function $x = x(\sigma, \tau)$ from (168) solves the equation of motion in (164). Relation (170) guarantees that the components of $x(\sigma, \tau)$ are real numbers.

Substituting the function $x = x(\sigma, \tau)$ from (168) into (167), we obtain the expression for L_m^\pm given in (169). Now, to the point: It follows from $L_m^\pm = 0$ for all m that

$$\partial_\pm x \partial_\pm x = 0 \quad \text{for all } \sigma, \tau \in \mathbb{R},$$

since the quantities L_m^\pm are proportional to the Fourier coefficients of the functions $\partial_\pm x \partial_\pm x$. Thus, the Virasoro constraints from (165) are fulfilled.

2.15f. The Virasoro algebra and infinite-dimensional Lie algebras. Consider the unit circle $S^1 := \{z \in \mathbb{C}: |z| = 1\}$. Let X be the space of all C^∞ -functions $f: S^1 \rightarrow \mathbb{C}$ that can be extended to a holomorphic function on an open neighborhood of S^1 . Define the operator

$$\mathcal{L}_m: X \rightarrow X$$

through

$$\mathcal{L}_m f := -z^{m+1} \frac{df}{dz}, \quad m \in \mathbb{Z}.$$

Set $[\mathcal{A}, \mathcal{B}] := \mathcal{A}\mathcal{B} - \mathcal{B}\mathcal{A}$. Moreover, let

$$\text{Vir}_0 := \text{span}\{\mathcal{L}_m: m \in \mathbb{Z}\}.$$

Finally, choose a symbol C , and set

$$\text{Vir} := \text{span}\{C, \mathcal{L}_m: m \in \mathbb{Z}\}.$$

(i) Show that, for all $m, n \in \mathbb{Z}$, we have the commutation relations:

$$[\mathcal{L}_n, \mathcal{L}_m] = (n - m)\mathcal{L}_{n+m}.$$

This way, the complex linear space Vir_0 becomes an infinite-dimensional Lie algebra called the special Virasoso algebra.

(ii) Determine the numbers $a(n, m)$ in such a way that, for all $n, m \in \mathbb{Z}$,

$$[\mathcal{L}_n, \mathcal{L}_m] = (n - m)\mathcal{L}_{n+m} + a(n, m)C,$$

$$[\mathcal{L}_n, C] = 0.$$

This way, the complex linear space Vir becomes an infinite-dimensional Lie algebra called the Virasoro algebra, which is also called the *central extension* of Vir_0 .

Hint: Use the Jacobi identity to show that

$$a(n, m) = \text{const}(n^3 - n)\delta_{n, -m}.$$

2.15g. Supernumbers. Consider an infinite number of symbols

$$\theta_1, \theta_2, \dots .$$

Introduce a product $\theta_k \theta_m$ that has the following decisive property:

$$\theta_k \theta_m = -\theta_m \theta_k \quad \text{for all } k, m.$$

Let α_j be a complex number. Each (finite) complex linear combination of finite products

$$\alpha_0 + \sum_{k,m} \alpha_{km} \theta_k \theta_m + \sum_{k,m,l} a_{kml} \theta_k \theta_m \theta_l + \dots$$

is called a *supernumber*.

Show that the Grassmann algebra from Problem 4.14d is a model for supernumbers.

The symbols θ are called Grassmann variables. From an historical point of view, it is interesting that such quantities were already introduced by Hermann Grassmann in 1844. From a physical point of view, supermathematics allows us to formulate theories that describe bosons (integer-spin particles) and fermions (half-numberly spin particles) in a unique way.

Solution: Let X be a complex linear space with $\dim X = \infty$. Suppose that $X = \text{span}\{\theta_1, \theta_2, \dots\}$, where $\theta_1, \dots, \theta_n$ are linearly independent for each n . Define

$$\theta_k \theta_m := \theta_k \wedge \theta_m, \quad \theta_k \theta_m \theta_l := \theta_m \wedge \theta_k \wedge \theta_l,$$

and so on. For example, $\theta_k \theta_m$ is a bilinear form on X^T , where

$$(\theta_k \theta_m)(u, v) = \theta_k(u)\theta_m(v) - \theta_k(v)\theta_m(u) \quad \text{for all } u, v \in X^T.$$

2.15h. Supermathematics. It is possible to construct a reach mathematics based on supernumbers. This is called *supermathematics*. Let us consider some examples.

(i) Differentiation:

$$\partial_j \theta_j := 1, \quad \partial_j (\theta_j \theta_k \theta_m \cdots \theta_r) := \theta_k \theta_m \cdots \theta_r.$$

(ii) Integration:²⁸

$$\int (\alpha + \beta \theta) d\theta := \beta.$$

²⁸This definition is motivated by the classic formula

$$\int_{-\infty}^{\infty} f(x + \text{const}) dx = \int_{-\infty}^{\infty} f(x) dx$$

(translation invariance of the integral).

Compute the following expressions:

$$e^\theta, \quad \int e^\theta d\theta, \quad \frac{de^\theta}{d\theta}, \quad \partial_1(\theta_2\theta_1), \quad \partial_1(e^{\theta_2} \sin \theta_1).$$

Solution:

(a) Note that $\theta^2 = \theta\theta = -\theta\theta$, and hence $\theta^2 = 0$. Therefore,

$$e^\theta = 1 + \theta + \frac{\theta^2}{2} + \dots = 1 + \theta.$$

$$(b) \int e^\theta d\theta = \int (1 + \theta) d\theta = 1.$$

$$(c) (e^\theta)' = (1 + \theta)' = 1.$$

$$(d) \partial_1(\theta_2\theta_1) = \partial_1(-\theta_1\theta_2) = -\theta_2.$$

$$(e) e^{\theta_2} \sin \theta_1 = (1 + \theta_2)\theta_1.$$

$$(f) \partial_1(\theta_1 + \theta_2\theta_1) = 1 - \theta_2.$$

As an introduction to modern supermathematics, we recommend the monographs by Berezin (1987), Bagger and Wess (1991) (supersymmetry and supergravity), and Constantinescu and de Groote (1994) (sheaf-theoretical approach).

Remark (The importance of string theory in modern physics). We will use only formal arguments.

(i) *The quantization of the bosonic string.* Consider first the bosonic string from (168). In order to quantize this string, we have to replace the Fourier coefficients $(\alpha_m^\pm)^j$ by operators in a “Hilbert space”²⁹ H satisfying the following commutation rules:

$$[(\alpha_m^\pm)^j, (\alpha_n^\pm)^k] = m\delta_{n,-m}\eta^{jk},$$

$$[(\alpha_m^\pm)^j, (\alpha_n^\mp)^k] = 0, \quad n, m \in \mathbb{Z}.$$

Then the Virasoro charges become operators, that is,

$$L_m^\pm = \frac{1}{2} \sum_{n=-\infty}^{\infty} \alpha_{m-n}^\pm \alpha_n^\pm, \quad m \in \mathbb{Z},$$

by (169). It turns out that

$$[L_m, L_n] = (m - n)L_{m+n} + \frac{c}{12}(n^3 - n)\delta_{m,-n}I, \quad m, n \in \mathbb{Z},$$

²⁹Note that the inner product on H is not positive definite.

where c is a constant, and where I denotes the unit operator. This means that the Virasoro charges form a Virasoro algebra. This fact plays a fundamental role in string theory.

The classic constraints $L_m^\pm = 0$ have to be replaced by the conditions

$$L_m^\pm \psi = 0, \quad (L_0 - 1)\psi = 0, \quad \text{for all } m \in \mathbb{Z}. \quad (171)$$

The set of all $\psi \in H$ that satisfy condition (171) form a linear subspace H_{phys} of H . By definition, the elements ψ of H_{phys} are called *physical states*. Such a physical state ψ is called a *ghost state* iff

$$(\psi | \psi) < 0.$$

An important result (based on formal arguments) says that the physical states are ghost-free iff $d = 26$ (the dimension of the space–time manifold). The theory makes sense in this case only.

This quantization procedure corresponds to the old canonical quantization. In fact, modern string theory is based on quantization via the Feynman path integral. Such a quantization procedure has the decisive advantage that the relativistic invariance of the theory can be seen immediately.

(ii) *Superstring theory*. If one formulates string theories on the basis of supernumbers, then one obtains so-called superstring theories. In particular, the critical dimension for the heterotic string is $d = 10$. The use of supernumbers allows us to describe both bosonic and fermionic strings in a unified manner. It is fascinating that the simplest vibration of a closed superstring corresponds to a massless spin-two particle. Moreover, the gauge symmetries of string theory show that this spin-two particle has all the properties of the hypothetical graviton that is responsible for the gravitational force.

For many physicists, superstring theory is therefore the leading candidate for the unification of all fundamental forces in the universe (gravitation, weak interaction, electromagnetic interaction, and strong interaction). However, to date there has been no experimental evidence for the existence of strings. String theory predicts that the physical effects generated by superstrings are striking if the particle mass is near 10^{19} proton masses. However, even the largest particle accelerators or observations from cosmic ray detectors and satellites will, at best, be able to probe only indirect signals emerging from such extremely high energies.

As an introduction to string theory, we recommend the lecture notes by Lüst and Theissen (1989), and the monograph by Green, Schwarz, and Witten (1987). From a mathematical point of view, it is fascinating that string theory is related to many topics in mathematics (e.g., Riemannian surfaces, Kähler manifolds, Calabi–Yau manifolds, topology, characteristic classes and vector bundles, knot theory, Morse theory, Floer cohomology, Korteweg–de Vries hierarchy and solitons, Kac–Moody algebras, quantum groups, algebraic geometry, and number theory). In this connection, there

is now a fascinating flow of ideas from physics to pure mathematics, and vice versa. This can be found in the monographs by Kaku (1987), (1991). See also Waldschmidt et al. (1992).

The fundamental principle of stationary action in physics. The variational principles in Problems 2.1 through 2.15 correspond to the principle of stationary action. This is the most important principle in physics needed to obtain the basic equations in all fields of physics. Further applications of this principle to important physical problems can be found in the monographs by Soper (1975) and by Landau and Lifšic (1988), Vols. 1–10.

3

Principles of Linear Functional Analysis

I love mathematics not only because it is applicable to technology but also because it is beautiful.

Roszá Péter (1905–1977)

It is true that a mathematician, who is not somewhat of a poet, will never be a perfect mathematician.

Karl Weierstrass (1815–1897)

A mathematician, like a painter or poet, is a maker of patterns. If his patterns are more permanent than theirs, it is because they are made with ideas.

Godfrey Harold Hardy (1877–1947)

Linear functional analysis is based on the following two important principles:

- (i) the Hahn–Banach theorem, and
 - (ii) the Baire theorem.

The Hahn–Banach theorem on the extension of linear functionals has been studied in Chapter 1. In this chapter we will investigate some applications of the Baire category theorem to linear operation equations. The most important consequences of the Baire theorem are the following (cf. Figure 3.1):

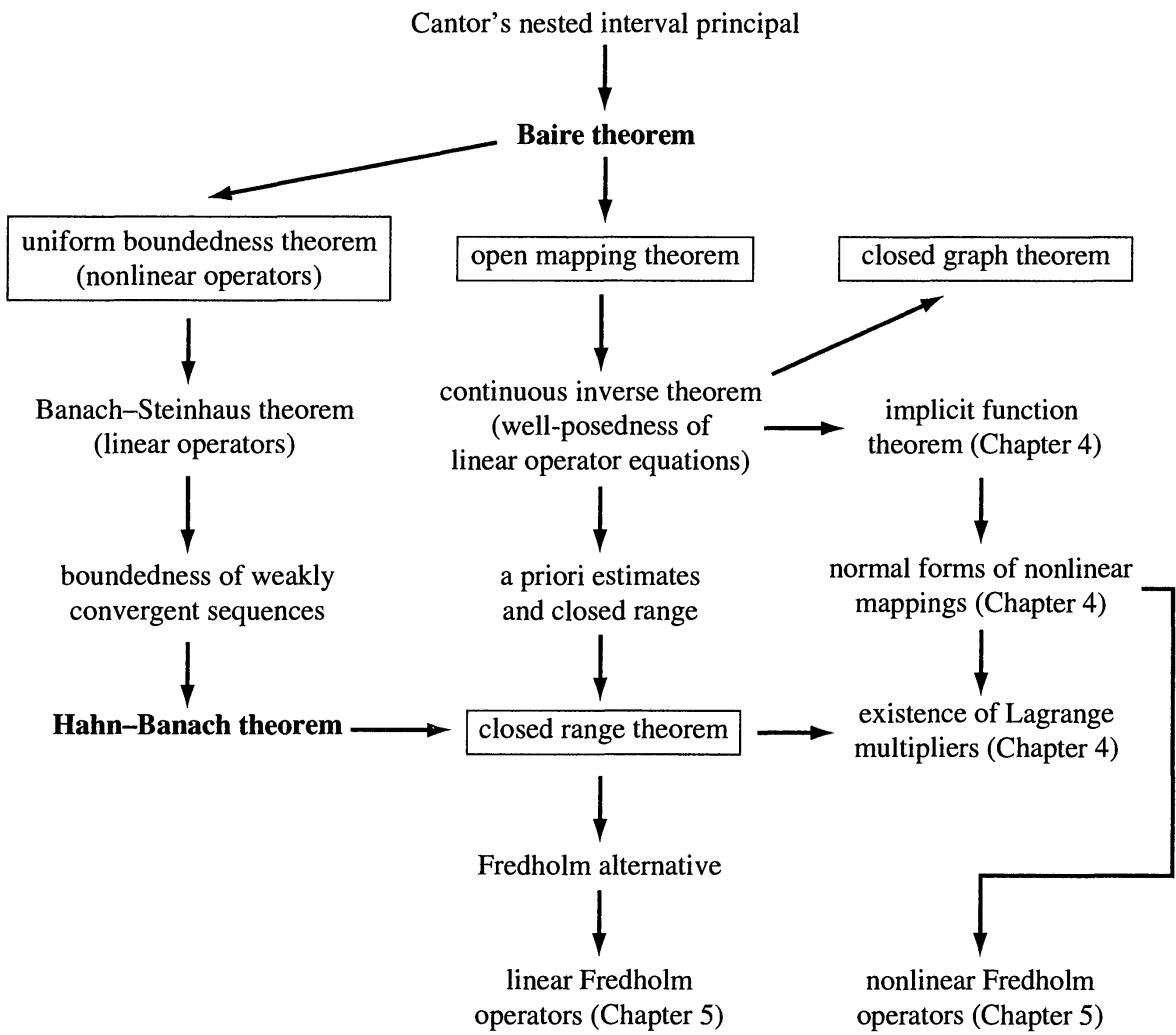


FIGURE 3.1.

- (a) the uniform boundedness theorem;
- (b) the open mapping theorem;
- (c) the closed graph theorem; and
- (d) the closed range theorem.

These fundamental results were proved by Banach in the late 1920s. The prototype of the Baire theorem was proved by Baire in 1899 before the creation of functional analysis.

A detailed presentation of the fascinating history of linear functional analysis can be found in the monograph by Dieudonné (1981).

3.1 The Baire Theorem

Definition 1. Let M be a subset of a normed space X over \mathbb{K} . Then

- (i) M is called *nowhere dense* in X iff

$$\text{int } \overline{M} = \emptyset,$$

that is, the closure \overline{M} of M does not contain any interior points.

- (ii) M is said to be of the *first category* in X iff M is the countable union of nowhere dense subsets M_n of X , that is,

$$M = \bigcup_{n=1}^{\infty} M_n.$$

Sets of the first category are also called *meager*.

- (iii) M is said to be of the *secondary category* in X iff M is *not* of the first category. Such sets are also called *fat*.

Standard Example 2 (Sets in \mathbb{R}).

- (i) Each finite set $\{x_1, \dots, x_n\}$ in \mathbb{R} is nowhere dense in \mathbb{R} .
- (ii) Each at most countable subset of \mathbb{R} is of the first category in \mathbb{R} .
- (iii) The set of rational numbers \mathbb{Q} is of the first category in \mathbb{R} .
- (iv) Each nonempty open subset of \mathbb{R} (e.g., \mathbb{R} itself) is of the second category in \mathbb{R} .

Proof. Ad (i), (ii). This is obvious.

Ad (iii). Note that \mathbb{Q} is countable.

Ad (iv). This is a special case of the Baire theorem (Theorem 3.A). \square

Proposition 3 (Cantor's nested interval principle). *Let $M_1 \supseteq M_2 \supseteq \dots$ be a sequence of nonempty closed subsets M_n of a Banach space X such that*

$$\lim_{n \rightarrow \infty} \text{diam } M_n = 0. \quad (1)$$

Then there exists a unique point u with $u \in M_n$ for all n .

Proof. Existence. Choose a point $u_n \in M_n$ for each n . By (1), the sequence (u_n) is Cauchy, and hence there is a point u such that $u_n \rightarrow u$ as $n \rightarrow \infty$. Since $u_n \in M_k$ for all $n \geq k$ and the set M_k is closed, $u \in M_k$ for each k .

Uniqueness. Let $u, v \in M_n$ for all n . By (1), $\|u - v\|$ is arbitrarily small. Hence $u = v$. \square

Theorem 3.A (The Baire theorem). *Each nonempty open subset U of a Banach space X over \mathbb{K} (e.g., $U = X$) is of the second category in X .*

Proof. If U were not of the second category, then U would be of the first category. Then there would exist a family $\{M_n\}$ of sets in X such that

$$U = \bigcup_{n=1}^{\infty} M_n \quad \text{and} \quad \text{int } \overline{M}_n = \emptyset \quad \text{for all } n.$$

Let us introduce the closed ball

$$B_r(a) := \{u \in X : \|u - a\| \leq r\}$$

of radius $r > 0$. First choose a point $a \in U$. Since the set U is open,

$$B_r(a) \subseteq U \quad \text{for some } r > 0.$$

Since $\text{int } \overline{M}_1 = \emptyset$, there exists a point $a_1 \in \text{int } B_r(a)$ such that¹ $\text{dist}(a_1, \overline{M}_1) > 0$. Thus, there is a number r_1 with $0 < r_1 < \frac{r_0}{2}$ such that

$$\overline{M}_1 \cap B_{r_1}(a_1) = \emptyset.$$

Continuing this argument, we obtain a sequence of balls

$$B_r(a) \supseteq B_{r_1}(a_1) \supseteq B_{r_2}(a_2) \cdots \quad \text{with } \lim_{n \rightarrow \infty} r_n = 0 \tag{2}$$

such that

$$\overline{M}_n \cap B_{r_n}(a_n) = \emptyset \quad \text{for all } n = 1, 2, \dots. \tag{3}$$

It follows from (2) and the *nested interval principle* (Proposition 1) that there exists a point u with $u \in B_{r_n}(a_n)$ for all n . By (3), $u \notin M_n$ for all n . This is a contradiction to

$$u \in B_r(a) \subseteq U = \bigcup_{n=1}^{\infty} M_n. \quad \square$$

¹Otherwise, $\text{dist}(b, \overline{M}_1) = 0$ for all $b \in \text{int } B_r(a)$. Since \overline{M}_1 is closed, this implies $b \in \overline{M}_1$ for all $b \in \text{int } B_r(a)$, and hence $\text{int } \overline{M}_1 \neq \emptyset$. This is a contradiction.

3.2 Application to the Existence of Nondifferentiable Continuous Functions

Proposition 1 (Weierstrass). *There exists a nondifferentiable continuous function $f: [0, 1] \rightarrow \mathbb{R}$.*

This will be proved by using the following general principle.

Existence Principle 2. Let M be a subset of a Banach space X , and let M be of the first category in X .

Then, there exists a point $u \in X$ such that

$$u \notin M.$$

Moreover, the set $X - M$ is of the second category in X .

Proof. By the Baire theorem (Theorem 3.A), X is of the second category. Since $X = M \cup (X - M)$ and M is of the first category, the set $X - M$ must be of the second category. Note that the union of two sets of the first category yields a set of the first category. \square

Define

$$M := \{f \in C[0, 1] : \text{there exists a point } x_* \in [0, 1[\text{ such that the right-hand derivative } f'_+(x_*) \text{ exists}\}.$$

Proposition 3. *The set M is of the first category in $C[0, 1]$.*

This implies that the set $C[0, 1] - M$ is of the second category in the Banach space $C[0, 1]$. Hence Proposition 3 implies Proposition 1. Roughly speaking, Proposition 3 tells us that “most” continuous functions $f: [0, 1] \rightarrow \mathbb{R}$ are nondifferentiable. In 1806 Ampère tried to prove that “each continuous function is differentiable.” More than fifty years later, Weierstrass showed that such a statement is wrong.

Proof of Proposition 3. Let M_n denote the set of all functions $f \in C[0, 1]$ such that there exists a point $x_* \in [0, 1[$ with

$$|f(x_* + h) - f(x_*)| \leq nh \quad \text{for all } h \in [0, 1] \text{ with } x_* + h \leq 1. \quad (4)$$

If $f \in M$, then $f'_+(x_*)$ exists and f is continuous on $[0, 1]$. Thus, $f \in M_n$ for some n , and hence

$$M \subseteq \bigcup_{n=1}^{\infty} M_n.$$

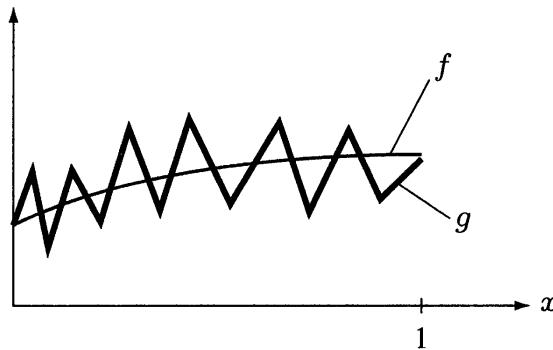


FIGURE 3.2.

We have to show that each set M_n is *nowhere dense* in $C[0, 1]$. Then M is of the *first category* in $C[0, 1]$.

We first prove that M_n is *closed*. To this end, let (f_k) be a sequence in M_n such that $f_k \in M_n$ for all $k = 1, 2, \dots$. Then there exist points x_k such that

$$|f_k(x_k + h) - f_k(x_k)| \leq nh$$

for all $h \in [0, 1]$ with $x_k + h \leq 1$, and $k = 1, 2, \dots$. (5)

Since $x_k \in [0, 1]$ for all k , there is a subsequence, again denoted by (x_k) , such that $x_k \rightarrow x_*$ as $k \rightarrow \infty$. Letting $k \rightarrow \infty$ in (5), we have²

$$|f(x_* + h) - f(x_*)| \leq nh \quad \text{for all } h \in [0, 1] \text{ with } x_* + h \leq 1.$$

Hence $f \in M_n$, that is, M_n is closed.

We now show that $\text{int } M_n = \emptyset$. Let $f \in M_n$. For each $\varepsilon > 0$, there exists a piecewise linear, continuous function $g: [0, 1] \rightarrow \mathbb{R}$ such that

$$\|f - g\| \equiv \max_{0 \leq x \leq 1} |f(x) - g(x)| < \varepsilon$$

and $|g'_+(x)| > n$ for all $x \in [0, 1[$ (see Figure 3.2). This implies $g \notin M_n$. Hence f is not an interior point of M_n . □

3.3 The Uniform Boundedness Theorem

Theorem 3.B (The uniform boundedness theorem). *Let \mathcal{F} be a nonempty set of continuous maps*

$$F: X \rightarrow Y,$$

where X is a Banach space over \mathbb{K} and Y is a normed space over \mathbb{K} . Suppose that

$$\sup_{F \in \mathcal{F}} \|Fu\| < \infty \quad \text{for all } u \in X.$$

²This limit exists, since $f_k(x) \rightarrow f(x)$ as $k \rightarrow \infty$ uniformly on $[0, 1]$ and the function f is uniformly continuous on $[0, 1]$.

Then there exists a closed ball B in X of positive radius such that

$$\sup_{u \in B} \left(\sup_{F \in \mathcal{F}} \|Fu\| \right) < \infty.$$

Proof. Set

$$M_k := \bigcap_{F \in \mathcal{F}} \{u \in X : \|Fu\| \leq k\}.$$

Obviously,

$$X = \bigcup_{n=1}^{\infty} M_n.$$

Since F is continuous, the set M_k is closed.³ By the Baire theorem (Theorem 3.A), $\text{int } M_k \neq \emptyset$ for some k . Hence the set M_k contains a closed ball B of positive radius. Then, by the definition of M_k ,

$$\sup_{u \in B} \left(\sup_{F \in \mathcal{F}} \|Fu\| \right) \leq k. \quad \square$$

Corollary 1 (The Banach–Steinhaus theorem). *Let \mathcal{L} be a nonempty set of linear continuous operators*

$$L: X \rightarrow Y,$$

where X is a Banach space over \mathbb{K} and Y is a normed space over \mathbb{K} . Suppose that

$$\sup_{L \in \mathcal{L}} \|Lu\| < \infty \quad \text{for all } u \in X.$$

Then $\sup_{L \in \mathcal{L}} \|L\| < \infty$.

Proof. By Theorem 3.B, there exists a closed ball B of positive radius in X such that

$$\sup_{x \in B} \left(\sup_{L \in \mathcal{L}} \|Lu\| \right) < \infty. \quad (6)$$

Since L is linear, we get $\|Lr(u - u_0)\| \leq r\|Lu\| + r\|Lu_0\|$ for all $r > 0$ and $u_0 \in X$. Thus, relation (6) remains true if B denotes the closed unit ball. Hence

$$\sup_{L \in \mathcal{L}} \|L\| = \sup_{L \in \mathcal{L}} \left(\sup_{\|u\| \leq 1} \|Lu\| \right) < \infty. \quad \square$$

³Obviously, the set $\{u \in X : \|Fu\| \leq k\}$ is closed. Furthermore, observe that the intersection of an arbitrary number of closed sets is again closed (cf. Problem 1.12).

Proposition 2. Let (L_n) be a sequence of linear continuous operators

$$L_n: X \rightarrow Y,$$

where X is a Banach space over \mathbb{K} and Y is a normed space over \mathbb{K} .

Then the following two conditions are equivalent:

(i) There exists a linear continuous operator $L: X \rightarrow Y$ such that

$$Lu = \lim_{n \rightarrow \infty} L_n u \quad \text{for all } u \in X.$$

(ii) There is a dense subset D of X such that $\lim_{n \rightarrow \infty} L_n u$ exists for all $u \in D$, and $\sup_n \|L_n\| < \infty$.

Proof. (i) \Rightarrow (ii). This follows from the Banach–Steinhaus theorem (Corollary 1).

(ii) \Rightarrow (i). Let $u \in X$. Then, for each $\varepsilon > 0$, there exists a point $v \in D$ such that

$$\|u - v\| < \varepsilon.$$

Since $(L_n v)$ is Cauchy,

$$\|L_n v - L_m v\| < \varepsilon \quad \text{for all } n, m \geq n_0(\varepsilon).$$

Hence

$$\begin{aligned} \|L_n u - L_m u\| &\leq \|L_n u - L_n v\| + \|L_n v - L_m v\| + \|L_m v - L_m u\| \\ &\leq 2 \left(\sup_n \|L_n\| \right) \|u - v\| + \varepsilon \quad \text{for all } n, m \geq n_0(\varepsilon). \end{aligned}$$

Thus, the sequence $(L_n u)$ is Cauchy and is hence convergent. Define

$$Lu := \lim_{n \rightarrow \infty} L_n u.$$

Obviously, the operator $L: X \rightarrow Y$ is linear. Moreover,

$$\|Lu\| \leq \left(\sup_n \|L_n\| \right) \|u\|$$

(i.e., L is also continuous). □

Standard Example 3 (Weak convergence). Let (u_n) be a sequence in the normed space X over \mathbb{K} . Then the following two conditions are equivalent:

(i) $u_n \rightharpoonup u$ as $n \rightarrow \infty$.

- (ii) The sequence $(\|u_n\|)$ is bounded, and there is a dense subset D of X^* such that

$$\langle f, u_n \rangle \rightarrow \langle f, u \rangle \quad \text{as } n \rightarrow \infty \text{ for all } f \in D.$$

Proof. Set $L_n f := \langle f, u_n \rangle$ for all $f \in X^*$ and fixed n . Since

$$|L_n f| \leq \|f\| \|u_n\| \quad \text{for all } f \in X^*,$$

the operator $L_n: X^* \rightarrow \mathbb{K}$ is linear and continuous. By Corollary 2 in Section 1.1,

$$\|L_n\| = \|u_n\| \quad \text{for all } n.$$

The assertion follows now from Proposition 2. Note that X^* is a Banach space, by Section 1.21 in AMS Vol. 108. \square

3.4 Applications to Cubature Formulas

Let $-\infty < a = x_0^{(n)} < x_1^{(n)} < \dots < x_n^{(n)} = b < \infty$. By a cubature formula, we understand a formula of the following form:

$$\int_a^b u(x) dx = L_n u + r_n(u), \tag{7}$$

where

$$L_n u := \sum_{k=0}^n c_k^{(n)} u(x_k^{(n)}), \quad n = 1, 2, \dots,$$

and $r_n(u)$ denotes the remainder. Our problem is to choose the real numbers $c_k^{(n)}$ in such a way that we obtain a *convergent* cubature formula for the given function u :

$$\lim_{n \rightarrow \infty} r_n(u) = 0.$$

Proposition 1. *The following two conditions are equivalent:*

- (i) *The cubature formula (7) is convergent for all continuous functions $u: [a, b] \rightarrow \mathbb{R}$.*
- (ii) *The cubature formula (7) is convergent for all polynomials u and*

$$\sup_{n \geq 1} \sum_{k=0}^n |c_k^{(n)}| < \infty. \tag{8}$$

Corollary 2. Suppose that all the numbers $c_k^{(n)}$ are nonnegative and that the cubature formula is exact for the function $u \equiv 1$; then condition (8) is satisfied.

In fact, letting $u(x) \equiv 1$ in (7), we get

$$\sum_{k=0}^n c_k^{(n)} = \int_a^b dx.$$

Proof of Proposition 1. Let $X := C[a, b]$, and set

$$Lu := \int_a^b u(x)dx.$$

Then, the operators $L, L_n: C[a, b] \rightarrow \mathbb{R}$ are linear and continuous. Moreover,

$$\|L_n\| = \sum_{k=0}^n |c_k^{(n)}| \quad \text{for } n = 1, 2, \dots. \quad (9)$$

To prove this, let $u \in C[a, b]$. Fix the number n . Then

$$|L_n u| \leq \sum_{k=0}^n |c_k^{(n)}| \max_{a \leq x \leq b} |u(x)| = \sum_{k=0}^n |c_k^{(n)}| \|u\|.$$

Furthermore, let us construct a piecewise linear, continuous function $w: [a, b] \rightarrow \mathbb{R}$ by prescribing the values

$$w(x_k^{(n)}) := \operatorname{sgn} c_k^{(n)}, \quad k = 0, \dots, n,$$

at all the node points $x_k^{(n)}$. Then

$$|L_n w| = \left| \sum_{k=0}^n c_k^{(n)} \operatorname{sgn} c_k^{(n)} \right| = \sum_{k=0}^n |c_k^{(n)}| \|w\|,$$

since $\|w\| = 1$. This yields (9).

By the Weierstrass approximation theorem, the set of polynomials is dense in the Banach space $C[a, b]$. Therefore, the assertion follows from the Banach–Steinhaus theorem (Proposition 2 in Section 3.3). \square

Example 3 (The trapezoid formula). Let $x_k^{(n)} := \frac{k(b-a)}{n}$, $k = 0, 1, \dots$. Then

$$L_n u := \frac{b-a}{n} \left(\frac{u(b) + u(a)}{2} + u(x_1^{(n)}) + \dots + u(x_{n-1}^{(n)}) \right) \quad (10)$$

is called the *trapezoid formula*, where $n = 1, 2, \dots$.

(i) For each $u \in C^2[a, b]$, we get the following error estimates:

$$|r_n(u)| \leq \frac{(b-a)^3}{12n^2} \max_{a \leq x \leq b} |u''(x)|, \quad n = 1, 2, \dots. \quad (11)$$

(ii) For each $u \in C[a, b]$, the trapezoid formula converges as $n \rightarrow \infty$.

Proof. Ad (i). We set $\alpha := x_k^{(n)}$ and $\beta := x_{k+1}^{(n)}$. Let

$$r := \int_{\alpha}^{\beta} u(x) dx - \frac{\beta - \alpha}{2}(u(\alpha) + u(\beta)).$$

For given y with $\alpha < y < \beta$, define the linear function

$$p(x) := u(\alpha) + (x - \alpha) \frac{u(\beta) - u(\alpha)}{\beta - \alpha}$$

and set

$$\rho(x) := u(x) - p(x) - \frac{u(y) - p(y)}{(y - \alpha)(y - \beta)}(x - \alpha)(x - \beta). \quad (12)$$

Then, $\rho(\alpha) = \rho(y) = \rho(\beta) = 0$. By the *mean value theorem*, this implies the existence of numbers ξ and η with $\alpha < \xi < y < \eta < \beta$ such that

$$\rho'(\xi) = \rho'(\eta) = 0.$$

Again by the mean value theorem, there is a number ζ with $\xi < \zeta < \eta$ such that

$$\rho''(\zeta) = 0.$$

According to (12), this implies

$$u''(\zeta) - 2 \frac{u(y) - p(y)}{(y - \alpha)(y - \beta)} = 0,$$

and hence

$$u(x) - p(x) = \frac{u''(\zeta(x))}{2} (x - \alpha)(x - \beta) \quad \text{for all } x \in [\alpha, \beta].$$

Integration yields

$$\left| \int_{\alpha}^{\beta} u(x) dx - \int_{\alpha}^{\beta} p(x) dx \right| \leq \frac{1}{2} \max_{\alpha \leq x \leq \beta} |u''(x)| \int_{\alpha}^{\beta} (x - \alpha)(x - \beta) dx.$$

Hence

$$\left| \int_{\alpha}^{\beta} u(x) dx - \frac{1}{2}(\beta - \alpha)(u(\alpha) + u(\beta)) \right| \leq \max_{\alpha \leq x \leq b} |u''(x)| \frac{(\beta - \alpha)^3}{12}.$$

Recall that $\alpha = \frac{k(b-a)}{n}$ and $\beta = \frac{(k+1)(b-a)}{n}$; then summation over k yields (11).

Ad (ii). If u is a polynomial, then $r_n(u) \rightarrow 0$ as $n \rightarrow \infty$, by (11). Moreover, the trapezoid formula (10) is exact for $u \equiv 1$, again by (11).

Thus, assertion (ii) is a special case of Proposition 1 and Corollary 2. \square

3.5 The Open Mapping Theorem

Theorem 3.C (Banach's open mapping theorem). *Let $A: X \rightarrow Y$ be a linear continuous operator, where X and Y are Banach spaces over \mathbb{K} . Then, the following two conditions are equivalent:*

- (i) *A is surjective.*
- (ii) *A is open, that is, A maps open sets onto open sets.*

Proof. (i) \Rightarrow (ii). Let us introduce the open ball

$$\mathcal{B}_R := \{u \in X : \|u\| < R\}.$$

Step 1: Since A is surjective,

$$Y = \bigcup_{n=1}^{\infty} \overline{A(\mathcal{B}_n)}. \quad (13)$$

By the *Baire theorem* (Theorem 3.A), there is some index m such that the closure $\overline{A(\mathcal{B}_m)}$ is *not* nowhere dense. Thus, there is a point $w \in Y$ such that

$$w \in \text{int } \overline{A(\mathcal{B}_m)}.$$

Since A is surjective, there exists some point $u \in X$ such that $w = Au$. Hence

$$0 \in \text{int } \overline{A(\mathcal{B}_m - u)}.$$

Finally, choose the number $R > 0$ so large that $\mathcal{B}_m - u \subseteq \mathcal{B}_R$. Then

$$0 \in \text{int } \overline{A(\mathcal{B}_R)}. \quad (14)$$

Step 2: Let us prove the stronger result that

$$0 \in \text{int } A(\mathcal{B}_R). \quad (15)$$

Condition (14) means that there is some number $r > 0$ such that

$$\|v\| < r \quad \text{with } v \in Y \quad \text{implies } v \in \overline{A(\mathcal{B}_R)}.$$

In particular, this implies the following:

$$\begin{aligned} \text{For each } v \in Y \text{ with } \|v\| < r, \text{ there is a point} \\ u \in \mathcal{B}_r \text{ such that } \|v - Au\| < \frac{r}{2}. \end{aligned} \quad (16)$$

To prove (15) it is sufficient to show that, for each $v \in Y$ with $\|v\| < r$, there is some point $u \in \mathcal{B}_{3R}$ such that

$$v = Au. \quad (17)$$

In fact, this means that $0 \in \text{int } A(\mathcal{B}_{3R})$, and hence we get (15), by the linearity of A .

Let $v \in Y$ be given with $\|v\| < r$. Using (16), we construct a sequence (u_n) in the ball \mathcal{B}_R such that $v_0 := v$ and

$$\|2(v_n - Au_n)\| < r, \quad v_{n+1} = 2(v_n - Au_n), \quad n = 0, 1, \dots.$$

Hence

$$2^{-n-1}v_{n+1} = 2^{-n}v_n - A(2^{-n}u_n), \quad n = 0, 1, \dots.$$

This implies

$$A\left(\sum_{n=0}^m 2^{-n}u_n\right) = v_0 - 2^{-m-1}v_{m+1}. \quad (18)$$

Since $\sum_{n=0}^m \|2^{-n}u_n\| \leq \sum_{n=0}^m 2^{-n}R \leq 2R$, the series

$$u := \sum_{n=0}^{\infty} 2^{-n}u_n$$

is convergent, by Section 1.22 in AMS Vol. 108. Hence $\|u\| < 3R$. Letting $m \rightarrow \infty$ in (18), we get (17).

Step 3: Let U be an open subset of X , and let $u \in U$. Then there is some $r > 0$ such that

$$u + r\mathcal{B}_R \subseteq U.$$

Using the linearity of the operator A and (15), we obtain

$$Au \in \text{int } A(u + r\mathcal{B}_R),$$

and hence $Au \in \text{int } A(U)$. Thus, the set $A(U)$ is open.

(ii) \Rightarrow (i). Since A is open, the set $A(X)$ contains an interior point. This implies $A(X) = Y$, by the linearity of A . \square

Proposition 1 (Banach's continuous inverse theorem). *Let $A: X \rightarrow Y$ be a linear continuous operator, where X and Y are Banach spaces over \mathbb{K} . If the inverse operator*

$$A^{-1}: Y \rightarrow X$$

exists, then it is continuous.

Proof. By the open mapping theorem (Theorem 3.C), the operator A is open. Thus, if the set W is open in X , then $A(W)$ is open in Y . A general result about continuous maps on topological spaces tells us that this implies the continuity of A^{-1} (cf. Problem 1.13a).

A direct proof goes like this. Set $\mathcal{B}_\varepsilon := \{u \in X : \|u\| < \varepsilon\}$. Since A^{-1} is linear, it is sufficient to prove that A^{-1} is continuous at the point $v = 0$. In fact, for each given $\varepsilon > 0$, the set $A(\mathcal{B}_\varepsilon)$ is open, since A is open. Hence $0 \in \text{int } A(\mathcal{B}_\varepsilon)$ because $A(0) = 0$. Thus, there is some number $\delta(\varepsilon) > 0$ such that $\|Au\| < \delta(\varepsilon)$ implies $u \in \mathcal{B}_\varepsilon$, that is,

$$\|Au\| < \delta(\varepsilon) \quad \text{implies} \quad \|u\| < \varepsilon.$$

Hence $\|v\| < \delta(\varepsilon)$ implies $\|A^{-1}v\| < \varepsilon$. This means that the operator A^{-1} is continuous at $v = 0$. \square

The following corollary represents an important reformulation of Proposition 1 in terms of the operator equation

$$Au = v, \quad u \in X. \tag{19}$$

Corollary 2 (The well-posedness principle). *Let $A: X \rightarrow Y$ be a linear continuous operator, where X and Y are Banach spaces over \mathbb{K} . Then the following two conditions are equivalent:*

- (i) *Equation (19) is well posed, that is, by definition, for each given $v \in Y$, equation (19) has a unique solution u , which depends continuously on v .*
- (ii) *For each $v \in Y$, equation (19) has a solution u , and $Aw = 0$ implies $w = 0$.*

3.6 Product Spaces

Definition 1. Let X_1, \dots, X_n be normed spaces over \mathbb{K} . The *product space*

$$X_1 \times \cdots \times X_n$$

consists of all the n -tuples

$$(u_1, \dots, u_n), \quad \text{where } u_k \in X_k \text{ for } k = 1, \dots, n.$$

For $\alpha \in \mathbb{K}$, we set

$$\begin{aligned} \alpha(u_1, \dots, u_n) &:= (\alpha_1 u_1, \dots, \alpha_n u_n), \\ (u_1, \dots, u_n) + (v_1, \dots, v_n) &:= (u_1 + v_1, \dots, u_n + v_n), \end{aligned}$$

and

$$\|(u_1, \dots, u_n)\| := \sum_{k=1}^n \|u_k\|. \quad (20)$$

Then, $X_1 \times \dots \times X_n$ becomes a *normed space over \mathbb{K}* .

Proposition 2. *If X_1, \dots, X_n are Banach spaces, then so is the product space $X_1 \times \dots \times X_n$.*

Proof. Let us consider the case where $n = 2$. The general case proceeds analogously.

Suppose that the sequence of the points (u_n, v_n) is Cauchy in $X_1 \times X_2$. Then

$$\|(u_n, v_n) - (u_m, v_m)\| = \|u_n - u_m\| + \|v_n - v_m\| < \varepsilon \quad \text{for all } n, m \geq n_0(\varepsilon).$$

Thus, (u_n) and (v_n) is Cauchy in X_1 and X_2 , respectively. Hence

$$u_n \rightarrow u \text{ in } X_1 \quad \text{and} \quad v_n \rightarrow v \text{ in } X_2 \quad \text{as } n \rightarrow \infty.$$

This implies

$$\|(u_n, v_n) - (u, v)\| = \|u_n - u\| + \|v_n - v\| \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

that is, $(u_n, v_n) \rightarrow (u, v)$ in $X_1 \times X_2$ as $n \rightarrow \infty$. □

By the same argument, it follows from (20) that

$$(u_1^{(k)}, \dots, u_n^{(k)}) \rightarrow (u_1, \dots, u_n) \text{ in } X_1 \times \dots \times X_n \text{ as } k \rightarrow \infty$$

iff all the components converge, that is,

$$u_m^{(k)} \rightarrow u_m \text{ in } X_m \text{ as } k \rightarrow \infty$$

for all $m = 1, \dots, n$.

3.7 The Closed Graph Theorem

Definition 1. Let X and Y be normed spaces over \mathbb{K} . By the *graph $G(A)$* of the operator

$$A: D(A) \subseteq X \rightarrow Y,$$

we mean the subset

$$G(A) := \{(u, Au) : u \in D(A)\}$$

of the product space $X \times Y$.

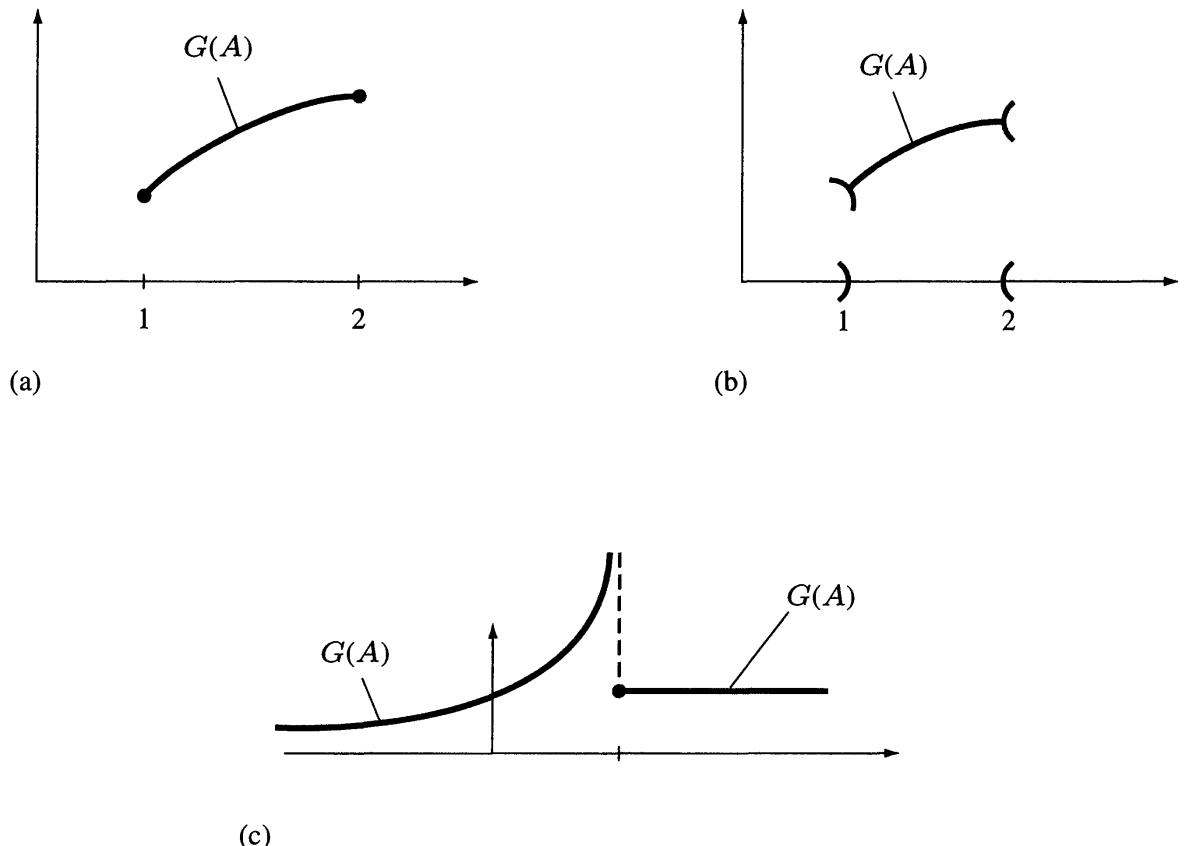


FIGURE 3.3.

The operator A is called *graph-closed* iff $G(A)$ is closed in $X \times Y$. This means that for each sequence (u_n) in the set $D(A)$ it follows from

$$u_n \rightarrow u \text{ in } X \quad \text{as } n \rightarrow \infty \quad (21)$$

and

$$Au_n \rightarrow v \text{ in } Y \quad \text{as } n \rightarrow \infty$$

that $u \in D(A)$ and $v = Au$.

Example 2. Let $X = Y = \mathbb{R}$. Then the following are true:

- (i) The function $A: [1, 2] \rightarrow \mathbb{R}$ pictured in Figure 3.3(a) is continuous and graph-closed in $\mathbb{R} \times \mathbb{R}$.
- (ii) The function $A:]0, 1[\rightarrow \mathbb{R}$ pictured in Figure 3.3(b) is continuous but is *not* graph-closed in $\mathbb{R} \times \mathbb{R}$.
- (iii) The function $A: \mathbb{R} \rightarrow \mathbb{R}$ pictured in Figure 3.3(c) is *not* continuous but is graph-closed.

It follows from (21) that each continuous operator $A: X \rightarrow Y$ is also graph-closed. The converse is *not* always true, by Example 2(iii). However, the following theorem tells us that the situation is nice in the *linear* case.

Theorem 3.D (Banach's closed graph theorem). *Let X and Y be Banach spaces over \mathbb{K} .*

Then, each graph-closed linear operator $A: X \rightarrow Y$ is continuous.

Proof. Let us define the following two linear continuous operators,

$$P: G(A) \rightarrow X \quad \text{and} \quad Q: G(A) \rightarrow Y,$$

through

$$P(u, Au) := u \quad \text{and} \quad Q(u, Au) := Au,$$

for all $u \in X$. Obviously,

$$P(u, Au) = 0$$

implies $u = 0$ and $Au = 0$. Thus, the operator P is *bijective*. Since A is graph-closed, $G(A)$ is a closed linear subspace of the Banach space $X \times Y$. Hence $G(A)$ is also a Banach space. By the *continuous inverse theorem* in Section 3.5, the inverse operator

$$P^{-1}: X \rightarrow G(A)$$

is continuous. Obviously, the diagram

$$\begin{array}{ccc} X & \xrightarrow{A} & Y \\ & \searrow P^{-1} & \nearrow Q \\ & G(A) & \end{array}$$

is commutative (i.e., $A = QP^{-1}$). Therefore, A is continuous. \square

Standard Example 3. Let $A: X \rightarrow X$ be a linear *self-adjoint* operator on the Hilbert space X over \mathbb{K} . Then A is continuous.

Proof. Let $u_n \rightarrow u$ and $Au_n \rightarrow v$ in X as $n \rightarrow \infty$. It follows from

$$(Au_n | w) = (u_n | Aw) \quad \text{for all } w \in X$$

that

$$(v | w) = (u | Aw) = (Au | w) \quad \text{for all } w \in X,$$

that is, $Au = v$. Thus, A is graph-closed, and hence continuous, by Theorem 3.D. \square

3.8 Applications to Factor Spaces

The following results on *factor spaces* and *direct sums* represent important auxiliary tools for the investigation of linear and nonlinear operator equations in Section 3.12 and in Chapters 4 and 5. The proofs will be

based on the continuous inverse theorem, the closed graph theorem, the Hahn–Banach theorem, and the Zorn lemma.

Let L be a *linear subspace* of the linear space X over \mathbb{K} . For all $u, v \in X$, we define

$$u \equiv v \pmod{L} \quad \text{iff } u - v \in L. \quad (22)$$

This is an *equivalence relation*. In fact, for all $u, v, w, z \in X$ and $\alpha \in \mathbb{K}$, we have the following:

$$\begin{aligned} u &\equiv u \pmod{L}; \\ u \equiv v \pmod{L} &\Rightarrow v \equiv u \pmod{L}; \\ u \equiv v \pmod{L}, v \equiv w \pmod{L} &\Rightarrow u \equiv w \pmod{L}. \end{aligned} \quad (23)$$

This equivalence relation is compatible with the linear structure of L :

$$\begin{aligned} u \equiv v \pmod{L} &\Rightarrow \alpha u = \alpha v \pmod{L}; \\ u \equiv w \pmod{L}, v \equiv z \pmod{L} &\Rightarrow u + v \equiv w + z \pmod{L}. \end{aligned} \quad (24)$$

Definition 1. The *factor space* X/L consists of all the equivalence classes $[u]$ with respect to (22), that is,

$$v \in [u] \quad \text{iff } u \equiv v \pmod{L}.$$

Explicitly, this means that

$$[u] = u + L.$$

The elements v of the class $[u]$ are called the *representatives* of $[u]$. Obviously,

$$[u] = [v] \Leftrightarrow u \equiv v \pmod{L}. \quad (25)$$

If we introduce the linear operations

$$\alpha[u] := [\alpha u],$$

$$[u] + [v] := [u + v], \quad (26)$$

the factor space X/L becomes a *linear space*. The operations in (26) are well defined, namely, they are *independent* of the chosen representatives. This follows from (24) and (25). For example, if $[u] = [v]$, then $u \equiv v \pmod{L}$, and hence $\alpha u \equiv \alpha v \pmod{L}$, that is, $[\alpha u] = [\alpha v]$.

In other words, the factor space X/L consists of all the different sets

$$u + L, \quad \text{where } u \in X,$$

and the linear operations on X/L are given through

$$\begin{aligned} (u + L) + (v + L) &= (u + v) + L, \\ \alpha(u + L) &= \alpha u + L, \end{aligned}$$

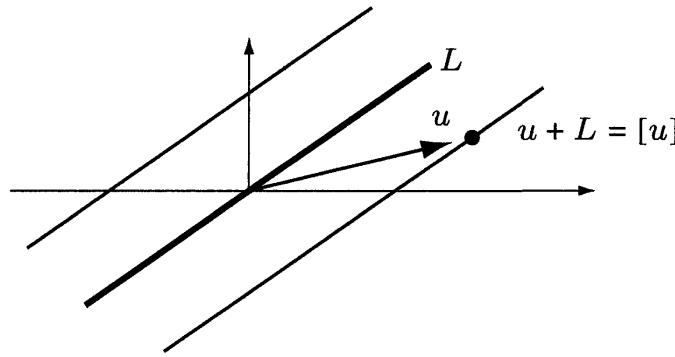


FIGURE 3.4.

which corresponds to the usual operations $A + B$ and αA for subsets A and B of linear spaces as defined in Section 1.1 of AMS Vol. 108.

Proposition 2. *Let L be a closed linear subspace of the normed space X over \mathbb{K} . Then the following are true:*

- (i) *The factor space X/L becomes a normed space over \mathbb{K} with respect to the norm*

$$\| [u] \| = \inf_{v \in [u]} \|v\|. \quad (27)$$

- (ii) *If X is a Banach space, then so is X/L .*

Since $[u] = u + L$, we get

$$\| [u] \| = \text{dist}(0, u + L) = \text{dist}(u, L).$$

Example 3. Let $X = \mathbb{R}^2$ with the Euclidean norm $\|\cdot\|$. In Figure 3.4, the factor space X/L consists of all the straight lines $[u] = u + L$ parallel to L , and the norm $\|u\|$ is equal to the distance from the origin to the straight line $[u]$.

Proof of Proposition 2. Ad (i). We first show that

$$\| [u] \| = 0 \Leftrightarrow [u] = 0.$$

This is identical to

$$\| [u] \| = 0 \Leftrightarrow u \in L.$$

In fact, if $u \in L$, then $[u] = L$. Hence $0 \in [u]$ and $\| [u] \| = 0$, by (27).

Conversely, let $\| [u] \| = 0$. Since L is closed, so is the set $[u] = u + L$. By (27), $0 \in u + L$. Hence $u \in L$.

Let $\alpha \in \mathbb{K}$. Since $\|\alpha v\| = |\alpha| \|v\|$, we have

$$\begin{aligned} \|\alpha [u]\| &= \inf_{w \in [u]} \|\alpha w\| \\ &= |\alpha| \inf_{w \in [u]} \|w\| = |\alpha| \| [u] \|. \end{aligned}$$

Finally, it follows from $\|w_1 + w_2\| \leq \|w_1\| + \|w_2\|$ that

$$\begin{aligned}\|[u] + [v]\| &= \inf_{w_1 \in [u], w_2 \in [v]} \|w_1 + w_2\| \\ &\leq \inf_{w_1 \in [u]} \|w_1\| + \inf_{w_2 \in [v]} \|w_2\| = \|[u]\| + \|[v]\|.\end{aligned}$$

Ad (ii). It follows from (27) that each class $[u]$ contains a point v such that

$$\|v\| \leq 2\|[u]\|. \quad (28)$$

Now let $([u_n])$ be a Cauchy sequence in X/L . Using a simple induction argument based on (28), we obtain a sequence (v_n) in X such that $v_n \in [u_n]$ and

$$\|v_n - v_{n+1}\| \leq 2\|[u_n] - [u_{n+1}]\| \quad \text{for all } n. \quad (29)$$

First suppose that

$$\|[u_n] - [u_{n+1}]\| \leq 2^{-n} \quad \text{for all } n. \quad (30)$$

It follows from (29) and the triangle inequality that

$$\|v_{n+m} - v_n\| \leq 2^{-n}(1 + 2^{-1} + 2^{-2} + \dots),$$

that is, (v_n) is Cauchy in X . Since X is a Banach space, we have

$$v_n \rightarrow v \quad \text{in } X \quad \text{as } n \rightarrow \infty.$$

By (27),

$$\|[u_n] - [v]\| \leq \|v_n - v\| \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

Hence $([u_n])$ is convergent in X/L .

In the general case, there exists a subsequence, again denoted by $([u_n])$, such that (30) holds. Then the assertion follows from Proposition 7 in Section 1.3 of AMS Vol. 108. \square

Definition 4. Let L be a linear subspace of the linear space X over \mathbb{K} . Then, the *canonical mapping*

$$\pi: X \rightarrow X/L$$

is defined through

$$\pi(u) := [u], \quad \text{for all } u \in X,$$

where $[u] = u + L$.

Proposition 5. If L is a closed linear subspace of the normed space X over \mathbb{K} , then the canonical mapping $\pi: X \rightarrow X/L$ is linear, continuous, and surjective.

Proof. For all $u \in X$, $\|\pi(u)\| = \| [u] \| \leq \|u\|$. □

Let $A: X \rightarrow Y$ be a *linear continuous operator*, where X and Y are *Banach spaces* over \mathbb{K} . We define the operator

$$[A]: X/N(A) \rightarrow R(A) \quad (31)$$

through

$$[A][u] := Au.$$

This definition is *independent* of the selected representative. In fact, let $[u] = [v]$. Then $u - v \in N(A)$, that is, $A(u - v) = 0$, and hence $Au = Av$. □

Proposition 6. *Let the range $R(A)$ of the operator A be closed.*

- (i) *The operator $[A]$ from (31) is a linear homeomorphism.*
- (ii) *There exists a number $c > 0$ such that*

$$c \cdot \text{dist}(u, N(A)) \leq \|Au\| \quad \text{for all } u \in X. \quad (32)$$

Proof. Ad (i). The null space $N(A) = \{u \in X: Au = 0\}$ is closed. In fact, if

$$Au_n = 0 \quad \text{and} \quad u_n \rightarrow u \quad \text{as } n \rightarrow \infty,$$

then $Au = 0$. Thus, $X/N(A)$ is a *Banach space*. Obviously, the operator $[A]$ is linear. Since

$$\| [A][u] \| = \|Av\| \leq \|A\| \|v\| \quad \text{for all } v \in [u],$$

we have $\| [A][u] \| \leq \|A\| \| [u] \|$, and thus $[A]$ is continuous.

Furthermore, the operator $[A]$ is *bijective*. In fact, if $[A][u] = 0$, then $u \in N(A)$, and hence $[u] = 0$.

Since $R(A)$ is a closed linear subspace of the Banach space Y , the range $R(A)$ is also a Banach space. The *continuous inverse theorem* from Section 3.5 now tells us that the inverse operator $[A]^{-1}: R(A) \rightarrow X/N(A)$ is continuous.

Ad (ii). By (i), there is a constant $d > 0$ such that

$$\| [A]^{-1}[u] \| \leq d \| [u] \| \quad \text{for all } [u] \in X/N(A).$$

Hence

$$\| [v] \| \leq d \| [A][v] \| \quad \text{for all } [v] \in X/N(A).$$

This is (32) with $c = d^{-1}$. □

3.9 Applications to Direct Sums and Projections

3.9.1 Projections

Definition 1. Let X be a linear space over \mathbb{K} , and let L_1 and L_2 be linear subspaces of X .

- (i) We write

$$X = L_1 \oplus L_2 \quad (33)$$

iff each $u \in X$ allows the following *unique* representation:

$$u = u_1 + u_2, \quad \text{where } u_1 \in L_1 \text{ and } u_2 \in L_2. \quad (34)$$

We say that X is the *direct sum* of L_1 and L_2 , and that L_2 is an *algebraic complement* of L_1 in X .

- (ii) The operator $P: X \rightarrow X$ is called an *algebraic projection* iff P is linear and $P^2 = P$.
- (iii) If X is a normed space, then the operator $P: X \rightarrow X$ is called a *continuous projection* iff P is a continuous algebraic projection.

Obviously,

$$X = L_1 \oplus L_2 \quad \text{iff } X = L_2 \oplus L_1.$$

Moreover, let $X = L_1 \oplus L_2$. Then

$$u \in L_1 \cap L_2 \quad \text{implies} \quad u = 0.$$

This follows from $u = u + 0 = 0 + u$ and from the uniqueness of the decomposition in (34).

Using the Zorn lemma, we will prove in Proposition 8 ahead that

Each linear subspace L_1 of the linear space X has an algebraic complement L_2 in X .

Proposition 2. *Let X be a linear space. Then the following statements hold true:*

- (i) *Suppose that $X = L_1 \oplus L_2$. If we set*

$$Pu := u_1$$

in (34), then $P: X \rightarrow X$ is an algebraic projection onto the linear subspace L_1 . Moreover,

$$L_1 = P(X) \quad \text{and} \quad L_2 = (I - P)(X) = N(P). \quad (35)$$

We call P the projection onto L_1 along L_2 .

- (ii) *Conversely, if $P: X \rightarrow X$ is an algebraic projection, then $X = L_1 \oplus L_2$ with (35).*

Proof. Ad (i). Since the decomposition in (34) is unique, and since

$$u_1 = u_1 + 0 \quad \text{with} \quad u_1 \in L_1 \quad \text{and} \quad 0 \in L_2,$$

we obtain $Pu_1 = u_1$, and hence $P^2u = Pu_1 = u_1 = Pu$. That is, $P^2 = P$.

By (34), $u_2 = u - u_1 = (I - P)u$. Hence $L_2 = (I - P)(X)$. Finally, it follows from (34) that

$$Pu = 0 \Leftrightarrow u \in L_2,$$

that is, $N(P) = L_2$.

Ad (ii). Let $u \in X$. Setting $u_1 := Pu$ and $u_2 := (I - P)u$, we obtain

$$u = u_1 + u_2, \quad \text{where } u_1 \in L_1 \quad \text{and} \quad u_2 \in L_2,$$

by (35). This decomposition is *unique*. In fact, let

$$u = v_1 + v_2, \quad \text{where } v_1 \in L_1 \quad \text{and} \quad v_2 \in L_2.$$

By (35), we get $v_1 = Pv$ and $v_2 = (I - P)v$ for some $v, w \in X$. Since $P^2 = P$, this implies $Pv_1 = v_1$ and $Pv_2 = 0$. Hence

$$u_1 = Pu = Pv_1 + Pv_2 = Pv_1 = v_1. \quad \square$$

This yields $u_1 = v_1$ and $u_2 = v_2$.

Definition 3. (i) The direct sum $X = L_1 \oplus L_2$ is called a *topological direct sum* iff the corresponding projection $P: X \rightarrow X$ onto L_1 along L_2 is *continuous*. Then we say that L_2 is a *topological complement* of L_1 in X .

(ii) The linear subspace L_1 *splits* the normed space X iff L_1 has a topological complement in X .

Example 4. Let $X = \mathbb{R}^2$. Then

$$\mathbb{R}^2 = L_1 \oplus L_2, \quad (36)$$

where L_1 and L_2 denote the two straight lines pictured in Figure 3.5. The projection $P: X \rightarrow X$ onto L_1 along L_2 corresponds to the usual *parallel projection* onto L_1 parallel to L_2 . Since P is continuous, (36) represents a topological direct sum.

Moreover, both L_1 and L_2 *split* \mathbb{R}^2 .

Proposition 5. *Let L be a linear subspace of the normed space X over \mathbb{K} . Then*

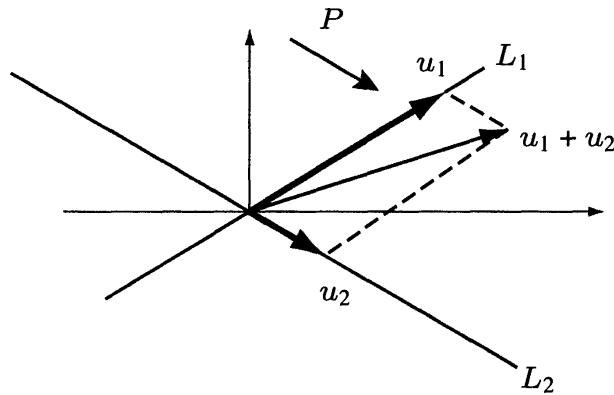


FIGURE 3.5.

- (i) L splits X iff there exists a continuous projection $P: X \rightarrow X$ onto L .
- (ii) If L splits X , then L is closed.

Unfortunately, the converse of (ii) is not true.⁴

Important classes of splitting closed linear subspaces will be considered later after some necessary preparations (cf. Standard Example 17).

Proof. Ad (i). This follows from Proposition 2.

Ad (ii). By (i), we have $L = P(X)$, where the projection $P: X \rightarrow L$ is continuous. Let (u_n) be a sequence in L such that $u_n \rightarrow u$ as $n \rightarrow \infty$. Letting $n \rightarrow \infty$, it follows from

$$u_n = Pu_n \quad \text{for all } n$$

that $u = Pu$, and hence $u \in L$. Thus, L is closed. \square

Proposition 6. Let $X = L_1 \oplus L_2$ be a direct sum, where L_1 and L_2 are linear subspaces of the Banach space X . Then the following two conditions are equivalent:

- (i) $X = L_1 \oplus L_2$ represents a topological direct sum.
- (ii) Both L_1 and L_2 are closed.

Proof. (i) \Rightarrow (ii). Both L_1 and L_2 split X , and hence L_1 and L_2 are closed.

(ii) \Rightarrow (i). Let $P: X \rightarrow X$ be the algebraic projection onto L_1 along L_2 . We have to show that P is continuous. To this end, let (u_n) be a sequence in X . Then,

$$u_n = u_{1n} + u_{2n}, \quad \text{where } u_{1n} \in L_1 \text{ and } u_{2n} \in L_2. \quad (37)$$

⁴Counterexamples were given by Murray, On complementary manifolds and projections in spaces L_p and l^p . *Transact. Amer. Math. Soc.* **41**(1937), 138–152.

Hence $u_{1n} = Pu_n$. Suppose that

$$u_n \rightarrow u \quad \text{and} \quad Pu_n \rightarrow v \quad \text{as } n \rightarrow \infty.$$

Letting $n \rightarrow \infty$ in (37), we get

$$u = v + w,$$

where $u_{2n} \rightarrow w$ as $n \rightarrow \infty$. Since L_1 and L_2 are closed, we have $v \in L_1$ and $w \in L_2$. Thus, $v = Pu$. Consequently, the operator P is graph-closed. The *closed graph theorem* in Section 3.7 tells us that P is *continuous*. \square

3.9.2 Codimension

Definition 7. Let L be a linear subspace of the linear space X over \mathbb{K} . Then, the *codimension* of L in X is defined as the dimension of the *factor space* X/L , denoted as

$$\text{codim } L := \dim(X/L).$$

Obviously, if $L = X$, then $X/L = \{0\}$, and hence $\text{codim } X = 0$. The following proposition explains the intuitive meaning of $\text{codim } L$.

Proposition 8. *Let L be a linear subspace of the linear space X over \mathbb{K} . Then the following statements hold true:*

- (i) *There exists a linear subspace M of X such that*

$$X = L \oplus M. \tag{38}$$

- (ii) *If M is any linear subspace of X such that (38) holds, then*

$$\text{codim } L = \dim M.$$

- (iii) *From (38) we get*

$$\dim X = \dim L + \dim M,$$

and hence

$$\dim X = \dim L + \text{codim } L.$$

It follows from (iii) that if $X = L \oplus M$ and $\dim X < \infty$, then

$$\text{codim } L = \dim X - \dim L. \tag{39}$$

Example 9. For an m -dimensional linear subspace L of \mathbb{R}^N , $N \geq 1$, we get

$$\text{codim } L = N - m. \tag{40}$$

For example, if L is a plane through the origin in \mathbb{R}^3 , then

$$\dim L = 2 \quad \text{and} \quad \operatorname{codim} L = 1.$$

Proof of Proposition 8. Ad (i). Let \mathcal{C} be the class of all the linear operators

$$P: D(P) \subseteq X \rightarrow L$$

such that $L \subseteq D(P)$ and $Pu = u$ on L . We write

$$P_1 \leq P_2 \quad \text{iff} \quad P_2 \text{ is an extension of } P_1.$$

By the *Zorn lemma* from the appendix in AMS Vol. 108, \mathcal{C} contains a maximal element P_0 . Then $D(P_0) = X$. Otherwise, there would exist a point $u_0 \in X - D(P_0)$. Set $N := D(P_0) + \operatorname{span}\{u_0\}$, and define the operator $P: N \rightarrow L$ through

$$P(u + \alpha u_0) := P_0(u) \quad \text{for all } u \in D(P_0), \alpha \in \mathbb{K}.$$

Then P is a proper extension of P_0 , contradicting the maximality of P_0 . In addition, we get

$$P_0^2 = P_0.$$

In fact, for each $v \in X$, it follows from $P_0v \in L$ that $P_0(P_0v) = P_0v$, by the construction of \mathcal{C} .

Therefore, the operator $P_0: X \rightarrow X$ is an algebraic projection onto L . Letting $M := (I - P_0)(X)$, we obtain

$$X = P_0(X) \oplus (I - P_0)(X) = L \oplus M.$$

Ad (ii). For each $u \in X$, we have

$$u = v + w, \quad \text{where } v \in L \text{ and } w \in M.$$

Define the map $\phi: M \rightarrow X/L$ through

$$\phi(w) := [w] \quad \text{for all } w \in M.$$

Then, ϕ is linear and surjective. Moreover, $\phi(w) = 0$ with $w \in M$ implies $w \in L$. It follows from $w \in L \cap M$ and $X = L \oplus M$ that $w = 0$. Thus, ϕ is a *bijection*. This yields

$$\dim M = \dim X/L,$$

and hence $\dim M = \operatorname{codim} L$.

Ad (iii). By (ii), it is sufficient to prove that $X = L \oplus M$ implies

$$\dim X = \dim L + \dim M. \tag{41}$$

First let $\dim L = \infty$. Then $L \subseteq X$ implies $\dim X = \infty$. Analogously, $\dim M = \infty$ yields $\dim X = \infty$.

Next suppose that $\dim L < \infty$ and $\dim M < \infty$. Then (41) follows from the fact that the union of a basis in L and a basis in M represents a basis in X . \square

Corollary 10. *Let L be a linear subspace of the linear space X over \mathbb{K} . Suppose that u_1, \dots, u_m are linearly independent elements of X such that*

$$L \cap \text{span}\{u_1, \dots, u_m\} = \{0\}. \quad (42)$$

Then, $m \leq \text{codim } L$.

Proof. It follows from (42) that $[u_1], \dots, [u_m]$ are linearly independent elements of X/L . Hence $m \leq \dim X/L$. \square

Corollary 11. *Let L be a closed linear subspace of the Banach space X over \mathbb{K} with $\text{codim } L < \infty$, and let S be a linear subspace of X such that*

$$L \subseteq S \subseteq X.$$

Then, S is closed and $\text{codim } S < \infty$.

Proof. Let us consider the canonical mapping

$$\pi: X \rightarrow X/L$$

from Section 3.8. Recall that $\pi(u) := [u] = u + L$ for all $u \in X$. The restriction of π to S is given by

$$\pi: S \rightarrow S/L.$$

The operator π is linear and continuous on X . Since $\text{codim } L < \infty$, we get $\dim X/L < \infty$, and hence $\dim S/L < \infty$. Consequently, the finite-dimensional subspace S/L of X/L is closed, and the preimage

$$S = \pi^{-1}(S/L)$$

is therefore also closed, by Lemma 12 which appears next.

Since $L \subseteq S$, it follows from

$$\alpha_1 u_1 + \cdots + \alpha_m u_m \equiv 0 \pmod{L}$$

that

$$\alpha_1 u_1 + \cdots + \alpha_m u_m \equiv 0 \pmod{S},$$

where $\alpha_1, \dots, \alpha_m \in \mathbb{K}$. Therefore, if u_1, \dots, u_m are linearly independent mod S , then they are also linearly independent mod L . Hence

$$\dim X/S \leq \dim X/L.$$

This yields $\text{codim } S \leq \text{codim } L$. □

Lemma 12. *Let $A: X \rightarrow Y$ be a continuous operator, where X and Y are normed spaces over \mathbb{K} . Let $W \subseteq Y$. The following conditions hold:*

- (i) *If W is open, then so is $A^{-1}(W)$.*
- (ii) *If W is closed, then so is $A^{-1}(W)$.*

Proof. This is a special case of a more general result about continuous maps on topological spaces (cf. Problem 1.13a). A direct proof resembles the following.

Ad (i). Let W be open, and let $u_0 \in A^{-1}(W)$. For each $\varepsilon > 0$, there is a $\delta(\varepsilon) > 0$ such that

$$\|u - u_0\| < \delta(\varepsilon) \quad \text{implies} \quad \|Au - Au_0\| < \varepsilon,$$

by the continuity of A . If we choose the number ε sufficiently small, then

$$\|u - u_0\| < \delta(\varepsilon) \quad \text{implies} \quad Au \in W,$$

and hence u_0 is an interior point of $A^{-1}(W)$. Thus, W is open.

Ad (ii). Use (i) and the fact that the complements of closed sets are open. □

3.9.3 Linear Operator Equations

Let us consider the linear operator equation

$$Au = b, \quad u \in X. \tag{43}$$

Proposition 13. *Suppose that the operator $A: X \rightarrow X$ is linear, where X and Y are linear spaces over \mathbb{K} . Let L be any fixed algebraic complement of the null space $N(A)$, namely, L is a linear subspace of X such that*

$$X = N(A) \oplus L. \tag{44}$$

Then the following statements are true:

- (i) *The restriction*

$$A: L \rightarrow R(A) \tag{45}$$

is linear and bijective. Hence

$$\text{codim } N(A) = \dim R(A). \tag{45*}$$

- (ii) In addition, suppose that X and Y are Banach spaces, L and $R(A)$ are closed, and the operator $A: X \rightarrow Y$ is continuous. Then the operator from (45) is a linear homeomorphism.

Recall that $R(A) = A(X)$. The number $\dim R(A)$ is called the *rank* of A . We denote this as

$$\operatorname{rank} A := \dim R(A).$$

Proof. It follows from $Au = 0$ with $u \in L$ that $u \in N(A) \cap L$. Hence $u = 0$, by (44).

Ad (ii). This follows from the *continuous inverse theorem* in Section 3.5. \square

Suppose that $\dim X < \infty$ and $\dim Y < \infty$. Let

$$B: L \rightarrow R(A)$$

denote the restriction of the operator $A: X \rightarrow R(A)$ to the linear subspace L of X . Then, for each given $b \in Y$, the solution set of the original equation (43) is given through

$$B^{-1}b + N(A),$$

where $\dim N(A) = \dim X - \operatorname{rank} A$, by (45*).

Proposition 14. Let $f_1, \dots, f_n, f: X \rightarrow \mathbb{K}$ be linear functionals on the linear space X over \mathbb{K} . Suppose that each solution $u \in X$ of the system

$$f_j(u) = 0, \quad j = 1, \dots, n,$$

is also a solution of the equation

$$f(u) = 0.$$

Then, there exist numbers $\alpha_1, \dots, \alpha_n \in \mathbb{K}$ such that

$$f = \alpha_1 f_1 + \cdots + \alpha_n f_n.$$

Proof. We may assume that f_1, \dots, f_n are linearly independent. The proof proceeds by induction.

Step 1: We prove the statement for $n = 1$. Since $f_1 \neq 0$, there exists a point $u_1 \in X$ such that $f_1(u_1) \neq 0$. Replacing u_1 with βu_1 , if necessary, we get

$$f_1(u_1) = 1.$$

Set

$$v := u - f_1(u)u_1.$$

Then $f_1(v) = 0$, and hence $f(v) = 0$, by hypothesis. This implies

$$0 = f(u) - f_1(u)f(u_1) \quad \text{for all } u \in X,$$

that is, $f = \alpha f_1$ for some $\alpha \in \mathbb{K}$.

Step 2: We prove the statement for $n = 2$. Since f_2 is linearly independent of f_1 , there exists a point $u_2 \in X$ such that

$$f_1(u_2) = 0 \quad \text{and} \quad f_2(u_2) \neq 0,$$

by Proposition 14 with $n = 1$ and $f = f_2$. Analogously, there exists a point $u_1 \in X$ such that

$$f_2(u_1) = 0 \quad \text{and} \quad f_1(u_1) \neq 0.$$

We may assume that

$$f_1(u_1) = f_2(u_2) = 1.$$

Set

$$v := u - f_1(u)u_1 - f_2(u)u_2.$$

Then, $f_1(v) = f_2(v) = 0$, and hence $f(v) = 0$ by hypothesis. This implies

$$0 = f(u) - f_1(u)f(u_1) - f_2(u)f(u_2) \quad \text{for all } u \in X,$$

that is, $f = \alpha_1 f_1 + \alpha_2 f_2$, where $\alpha_j := f(u_j)$.

Step 3: If the assertion is true for n , then a similar argument as in Step 2 shows that the statement is also true for $n + 1$. \square

3.9.4 Biorthogonal Systems and Splitting Subspaces

Definition 15. Let X be a normed space over \mathbb{K} . By an X -biorthogonal system $\{u_j, u_j^*\}_{j=1,\dots,n}$, we understand a system of points $u_1, \dots, u_n \in X$ and functionals $u_1^*, \dots, u_n^* \in X^*$ such that

$$\langle u_i^*, u_j \rangle = \delta_{ij} \quad \text{for all } i, j = 1, \dots, n.$$

Proposition 16. Let X be a normed space over \mathbb{K} .

- (i) Each system $u_1, \dots, u_n \in X$ of linearly independent points can be extended to an X -biorthogonal system.
- (ii) Each system $u_1^*, \dots, u_n^* \in X^*$ of linearly independent functionals can be extended to an X -biorthogonal system.

Proof. Ad (i). Let $L = \text{span}\{u_1, \dots, u_n\}$. Define the linear functional $u_i^*: L \rightarrow \mathbb{K}$ through

$$\langle u_i^*, \sum_{j=1}^n \alpha_j u_j \rangle := \alpha_i, \quad i = 1, \dots, n.$$

By the *Hahn–Banach theorem* in Section 1.1, u_i^* can be extended to a linear continuous functional $u_i^*: X \rightarrow \mathbb{K}$.

Ad (ii). Set $f_j := u_j^*$. Then the existence of points u_1, \dots, u_n with $\langle u_j^*, u_i \rangle = \delta_{ji}$ follows as in the proof of Proposition 14. \square

Standard Example 17. Let L be a linear subspace of the Banach space X over \mathbb{K} . Then L splits X if one of the following three conditions is met:

- (i) L is a closed linear subspace of the Hilbert space X .
- (ii) $\dim L < \infty$.
- (iii) L is closed and $\text{codim } L < \infty$.

Proof. Ad (i). By Proposition 12 in Section 5.1 of AMS Vol. 108, there exists a *continuous* orthogonal projection $P: X \rightarrow X$ onto L . Now use Proposition 5.

Ad (ii). Let $\{u_1, \dots, u_n\}$ be a basis of L . Extend this basis to an X -biorthogonal system $\{u_j, u_j^*\}$. Define

$$Pu := \sum_{j=1}^n \langle u_j^*, u \rangle u_j.$$

Then $Pu_k = u_k$ for all k , and hence $P^2 = P$. Thus, the operator $P: X \rightarrow X$ represents a *continuous* projection onto L . Now use Proposition 5.

Ad (iii). There exists a linear subspace M of X such that $X = L \oplus M$, where $\dim M = \text{codim } L < \infty$. Thus, L and M are *closed* subspaces of X . By Proposition 6, $X = L \oplus M$ is a topological direct sum, and hence L splits X . \square

3.9.5 Pseudo-Orthogonal Complements

Definition 18. Let L be a linear subspace of the normed space X over \mathbb{K} . The set

$$L^\perp := \{u^* \in X^*: \langle u^*, u \rangle = 0 \text{ for all } u \in L\} \tag{46}$$

is called the *pseudo-orthogonal complement* to L .

Let M be a linear subspace of X^* . Then we set

$${}^\perp M := \{u \in X: \langle u^*, u \rangle = 0 \text{ for all } u^* \in M\}.$$

These notions generalize orthogonal complements L^\perp in Hilbert spaces.⁵

Proposition 19. L^\perp and ${}^\perp M$ are closed linear subspaces of X and X^* , respectively.

Proof. Suppose that $u_n^* \in L^\perp$ for all n and

$$u_n^* \rightarrow u^* \text{ in } X^* \quad \text{as } n \rightarrow \infty.$$

If we let $n \rightarrow \infty$, it follows from $\langle u_n^*, v \rangle = 0$ for all n and all $v \in L$ that $\langle u^*, v \rangle = 0$ for all $v \in L$, and hence $u^* \in L^\perp$ (cf. Problem 3.5).

Suppose that $u_n \in {}^\perp M$ for all n and

$$u_n \rightarrow u \text{ in } X \quad \text{as } n \rightarrow \infty.$$

If we let $n \rightarrow \infty$, it follows from $\langle u^*, u_n \rangle = 0$ for all n and all $u^* \in M$ that $\langle u^*, u \rangle = 0$ for all $u^* \in M$, and hence $u \in {}^\perp M$. \square

Proposition 20. Let L be a linear subspace of the normed space X over \mathbb{K} .

$$\bar{L} = {}^\perp(L^\perp).$$

Proof. By (46), $(\bar{L})^\perp = L^\perp$. Therefore, it is sufficient to prove that

$$M = {}^\perp(M^\perp),$$

where M is a *closed* linear subspace of X . By Definition 18,

$$u \in {}^\perp(M^\perp) \text{ iff } \langle u^*, u \rangle = 0 \quad \text{for all } u^* \in M^\perp.$$

Hence $M \subseteq {}^\perp(M^\perp)$.

Conversely, we want to show that ${}^\perp(M^\perp) \subseteq M$. Let $v \in {}^\perp(M^\perp)$ and suppose that $v \notin M$. By Proposition 3 in Section 1.2, it follows from the *Hahn–Banach theorem* that there exists a functional $u^* \in X^*$ such that

$$u^* = 0 \text{ on } M \quad \text{and} \quad \langle u^*, v \rangle \neq 0.$$

Hence $u^* \in M^\perp$ and $v \notin {}^\perp(M^\perp)$. This is a contradiction. \square

The following result will be used in the theory of Fredholm operators, which will be studied in Chapter 5.

Proposition 21. Let X be a normed space over \mathbb{K} . Then

⁵In the following, the symbol L^\perp always corresponds to (46) if we do not state explicitly that L^\perp means an orthogonal complement in a Hilbert space.

(i) If L is a finite-dimensional linear subspace of X , then

$$\operatorname{codim} L^\perp = \dim L \quad \text{in } X^*.$$

(ii) If M is a finite-dimensional linear subspace of X^* , then

$$\operatorname{codim} {}^\perp M = \dim M \quad \text{in } X.$$

(iii) If L is a closed linear subspace of X such that L^\perp is finite-dimensional, then

$$\operatorname{codim} L = \dim L^\perp \quad \text{in } X.$$

Proof. Ad (i). If $L = \{0\}$, then $L^\perp = X^*$, and hence $\operatorname{codim} L^\perp = 0$.

Suppose now that $\dim L = n$, where $n > 0$. Let $\{u_1, \dots, u_n\}$ be a basis of L . Extend this to an X -biorthogonal system $\{u_j, u_j^*\}$. Define the continuous projection operator $P: X^* \rightarrow X^*$ through

$$Pu^* := u^* - \sum_{j=1}^n \langle u^*, u_j \rangle u_j^* \quad \text{for all } u^* \in X^*.$$

Obviously, $Pu^* = u^*$ iff $\langle u^*, u_j \rangle = 0$ for all j (i.e., $u^* \in L^\perp$). Thus, $P(X^*) = L^\perp$. Hence $\operatorname{codim} L^\perp = \dim(I - P)(X^*) = n$, by Proposition 8 along with $X^* = P(X^*) \oplus (I - P)(X^*)$.

Ad (ii). Use a similar argument as in the proof of (i).

Ad (iii). By Proposition 20, $L = {}^\perp(L^\perp)$. It follows from (ii) that $\operatorname{codim} L = \dim L^\perp$.

3.10 Dual Operators

The theory of linear operator equations in Banach spaces is essentially based on the concept of *duality*. To this end, we need dual operators.

The *key* relation for dual operators is given through

$$\langle A^T u^*, u \rangle = \langle u^*, Au \rangle \quad \text{for all } u \in X, u^* \in Y^*. \quad (47)$$

Proposition 1. Let

$$A: X \rightarrow Y$$

be a linear continuous operator, where X and Y are normed spaces over \mathbb{K} . Then there exists precisely one linear operator

$$A^T: Y^* \rightarrow X^*$$

such that relation (47) holds. In addition, A^T is continuous.

The operator A^T is called the transposed or *dual operator* to A . We will show in Example 3 ahead that, in finite-dimensional Hilbert spaces, the transposed operator A^T and the adjoint operator A^* correspond to the transposed matrix and the adjoint matrix, respectively.

Proof. Existence. Let $u^* \in Y^*$ be given. Set

$$f(u) := \langle u^*, Au \rangle \quad \text{for all } u \in X.$$

Then

$$|f(u)| \leq \|u^*\| \|Au\| \leq \|u^*\| \|Au\| \|u\| \quad \text{for all } u \in X. \quad (48)$$

Hence $f: X \rightarrow \mathbb{K}$ is a linear continuous functional, namely, $f \in X^*$. Define

$$A^T u^* := f.$$

Obviously, $\langle A^T u^*, u \rangle = \langle u^*, Au \rangle$ for all $u \in X$. This way, we obtain the linear operator $A^T: Y^* \rightarrow X^*$. By (48),

$$\|A^T u^*\| = \|f\| \leq \|A\| \|u^*\| \quad \text{for all } u^* \in Y^*,$$

and hence A^T is continuous.

Uniqueness. Let $u^* \in Y^*$ and $v^* \in X^*$ be given. Suppose that

$$\langle v^*, u \rangle = \langle u^*, Au \rangle \quad \text{for all } u \in X.$$

It follows from (47) that $\langle v^* - A^T u^*, u \rangle = 0$ for all $u \in X$, and hence $v^* = A^T u^*$. \square

Proposition 2. *Let $A: X \rightarrow X$ be a linear continuous operator on the Hilbert space X over \mathbb{K} . Then, the following diagram is commutative:*

$$\begin{array}{ccc} X^* & \xrightarrow{A^T} & X^* \\ J \uparrow & & \downarrow J^{-1} \\ X & \xrightarrow{A^*} & X \end{array}$$

Here, J denotes the duality map of X . Explicitly,

$$A^* = J^{-1} A^T J.$$

This result shows that there exists a simple relation between the *dual operator* A^T and the *adjoint operator* A^* .

Proof. Let $u, v \in X$. By Section 2.11 in AMS Vol. 108, we have

$$\langle Ju, v \rangle = (u | v).$$

Hence

$$(J^{-1}A^TJu \mid v) = \langle A^TJu, v \rangle = \langle Ju, Av \rangle = (u \mid Av). \quad \square$$

Example 3 (Matrix equations). Let $X \neq \{0\}$ be a finite-dimensional Hilbert space over \mathbb{K} with the orthonormal basis $\{e_1, \dots, e_N\}$ (e.g., $X = \mathbb{K}^N$, $N \geq 1$). Then, for each $u, b \in X$, we have the representations

$$u = \sum_{n=1}^N \xi_n e_n \quad \text{and} \quad b = \sum_{n=1}^N \beta_n e_n,$$

where $\xi_n, \beta_n \in \mathbb{K}$ for all n . Let

$$A: X \rightarrow X$$

be a linear operator. We are given $b \in X$ and $b^* \in X^*$. Then the following relations between operator equations and matrix equations hold true:

(i) The *original equation*

$$Au = b, \quad u \in X, \quad (49)$$

corresponds to the matrix equation

$$\mathcal{A}\xi = \beta, \quad (49^*)$$

where we set

$$\mathcal{A} := \begin{pmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & & \vdots \\ a_{N1} & \cdots & a_{NN} \end{pmatrix}, \quad \xi := \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_N \end{pmatrix}, \quad \beta := \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_N \end{pmatrix},$$

and the matrix elements a_{nm} of the operator A are given through

$$Ae_m = \sum_{n=1}^N a_{nm} e_n, \quad m = 1, \dots, N. \quad (50)$$

(ii) The *adjoint equation*

$$A^*u = b, \quad u \in X, \quad (51)$$

and the *dual equation*

$$A^T u^* = b^*, \quad u^* \in X^* \quad (52)$$

correspond to the matrix equations

$$\mathcal{A}^*\xi = \beta \quad (51^*)$$

and

$$\mathcal{A}^T \xi^* = \beta^*, \quad (52^*)$$

respectively. Here \mathcal{A}^* and \mathcal{A}^T denote the adjoint matrix and the transposed matrix to $\mathcal{A} = (a_{nm})$, respectively. Explicitly,

$$\mathcal{A}^* = (a_{nm}^*), \quad \text{where } a_{nm}^* := \bar{a}_{mn},$$

and

$$\mathcal{A}^T = (a_{nm}^T), \quad \text{where } a_{nm}^T := a_{mn},$$

for $n, m = 1, \dots, N$. The bar denotes the conjugate complex number. In addition,

$$u^* = \sum_{n=1}^N \xi_n^* e_n^* \quad \text{and} \quad b^* = \sum_{n=1}^N \beta_n^* e_n^*,$$

where $\xi_n^*, \beta_n^* \in \mathbb{K}$ and all n . The basis $\{e_1^*, \dots, e_N^*\}$ of the dual space X^* will be defined ahead.

If X is a real space (i.e., $\mathbb{K} = \mathbb{R}$), then $\mathcal{A}^* = \mathcal{A}^T$.

Proof. Ad (49). It follows from

$$b = Au = \sum_{n=1}^N \left(\sum_{m=1}^N a_{nm} \xi_m \right) e_n$$

that $\beta_n = \sum_{m=1}^N a_{nm} \xi_m$. This is (49*).

Ad (51). Noting that $(e_n | e_m) = \delta_{nm}$, from (50) we get

$$a_{nm} = (e_n | Ae_m).$$

Thus, the matrix elements of the adjoint operator A^* are given through

$$a_{nm}^* = (e_n | A^* e_m) = (Ae_n | e_m) = \bar{a}_{mn}.$$

Ad (52). For $n = 1, \dots, N$, define

$$e_n^* \left(\sum_{m=1}^N \xi_m e_m \right) := \xi_n.$$

Then, $e_n^*: X \rightarrow \mathbb{K}$ is a linear functional (i.e., $e_n^* \in X^*$). Let $u^* \in X^*$. Then

$$u^* \left(\sum_{n=1}^N \xi_n e_n \right) = \sum_{n=1}^N \xi_n \xi_n^*,$$

where $\xi_n^* := u^*(e_n)$. Hence

$$u^* = \sum_{n=1}^N \xi_n^* e_n^*, \quad \xi_1^*, \dots, \xi_N^* \in \mathbb{K}. \quad (53)$$

Conversely, each u^* from (53) is a linear functional on X . Thus, $\{e_1^*, \dots, e_N^*\}$ forms a basis of X^* .

According to (50), the matrix elements a_{nm}^T of the dual operator A^T are given through

$$A^T e_m^* = \sum_{n=1}^N a_{nm}^T e_n^*, \quad m = 1, \dots, N.$$

Since $\langle e_m^*, e_n \rangle = \delta_{mn}$, we get

$$\begin{aligned} a_{nm}^T &= \langle A^T e_m^*, e_n \rangle = \langle e_m^*, A e_n \rangle \\ &= \langle e_m^*, \sum_{k=1}^N a_{kn} e_k \rangle = a_{mn}. \end{aligned} \quad \square$$

The properties of dual operators can be described conveniently by using the so-called duality functor.

Definition 4. Let

$$A: X \rightarrow Y \tag{54}$$

be a linear continuous operator, where X and Y are normed spaces over \mathbb{K} . The *duality functor* \mathcal{D} assigns to (54) the dual operator

$$A^T: Y^* \rightarrow X^*. \tag{54*}$$

Proposition 5. Let X , Y , and Z be normed spaces over \mathbb{K} , and let $A: X \rightarrow Y$ and $B: Y \rightarrow Z$ be linear continuous operators.

Then, the duality functor \mathcal{D} is contravariant, that is, \mathcal{D} assigns to the sequence

$$X \xrightarrow{A} Y \xrightarrow{B} Z$$

the following sequence:

$$X^* \xleftarrow{A^T} Y^* \xleftarrow{B^T} Z^*.$$

Proof. We have to show that

$$(BA)^T = A^T B^T.$$

This follows immediately from

$$\langle v, BAu \rangle = \langle B^T v, Au \rangle = \langle A^T B^T v, u \rangle \quad \text{for all } u \in X, v \in Z^*. \quad \square$$

Corollary 6. If the operator $A: X \rightarrow Y$ is linear, continuous, and bijective, then so is the dual operator $A^T: Y^* \rightarrow X^*$. Moreover, we get

$$(A^T)^{-1} = (A^{-1})^T. \tag{55}$$

Proof. Let I_X denote the identity operator on X . It follows from

$$A^{-1}A = I_X \quad \text{and} \quad AA^{-1} = I_Y$$

that

$$A^T(A^{-1})^T = I_{X^*} \quad \text{and} \quad (A^{-1})^TA^T = I_{Y^*},$$

since $I_X^T = I_{X^*}$ and $I_Y^T = I_{Y^*}$. \square

Recall the following from Section 2.8. Let X be a Banach space over \mathbb{K} . If we set

$$j_X(u)(f) := \langle f, u \rangle \quad \text{for all } u \in X, f \in X^*,$$

then the linear continuous operator $j_X: X \rightarrow X^{**}$ preserves the norm, that is, $\|j_X(u)\| = \|u\|$ for all $u \in X$. Set

$$A^{TT} := (A^T)^T.$$

Proposition 7. *Let X and Y be Banach spaces over \mathbb{K} .*

(i) *The following diagram*

$$\begin{array}{ccc} X & \xrightarrow{j_X} & X^{**} \\ A \downarrow & & \downarrow A^{TT} \\ Y & \xrightarrow{j_Y} & Y^{**} \end{array}$$

is commutative for all operators $A \in L(X, Y)$.

(ii) *The duality functor \mathcal{D} is norm-preserving, that is,*

$$\|A^T\| = \|A\| \quad \text{for all } A \in L(X, Y).$$

(iii) *The duality functor \mathcal{D} is compact, that is, if $A \in L(X, Y)$ is compact, then so is $A^T \in L(Y^*, X^*)$.*

Proof. Ad (i). For all $u \in X$ and $v \in Y^*$,

$$\begin{aligned} \langle j_Y(Au), v \rangle &= \langle v, Au \rangle = \langle A^T v, u \rangle \\ &= \langle j_X(u), A^T v \rangle = \langle A^{TT} j_X(u), v \rangle, \end{aligned}$$

and hence $j_Y A = A^{TT} j_X$.

Ad (ii). Let $A \in L(X, Y)$. By the proof of Proposition 1,

$$\|A^T\| \leq \|A\|.$$

This implies

$$\|A^{TT}\| \leq \|A^T\|.$$

Since j_X and j_Y are norm-preserving, we get

$$\begin{aligned} \|A\| &= \sup_{\|u\|\leq 1} \|Au\| = \sup_{\|u\|\leq 1} \|j_Y(Au)\| \\ &= \sup_{\|u\|\leq 1} \|A^{TT}j_X(u)\| \leq \sup_{\|u\|\leq 1} \|A^{TT}\| \|j_X\| \|u\| \leq \|A^{TT}\|. \end{aligned}$$

Hence $\|A^T\| \leq \|A\| \leq \|A^{TT}\| \leq \|A^T\|$, showing that $\|A^T\| = \|A\|$.

Ad (iii). This will be proved in Section 5.1. \square

3.11 The Exactness of the Duality Functor

Riemann has shown us that proofs are better achieved through ideas than through long calculations.

David Hilbert

The *language* of exact sequences plays a fundamental role in modern mathematics (e.g., in algebraic topology). We want to show that this language allows us to give elegant proofs in linear operator theory (cf. Figure 3.6).

Definition 1. Let X_1, \dots, X_n be linear spaces over \mathbb{K} , and let $A_j: X_j \rightarrow X_{j+1}$, $j = 1, \dots, n - 1$, be linear operators. Then the sequence

$$X_1 \xrightarrow{A_1} X_2 \xrightarrow{A_2} X_3 \xrightarrow{X_3} \dots \xrightarrow{A_{n-2}} X_{n-1} \xrightarrow{A_{n-1}} X_n \quad (56)$$

is called *exact* iff $R(A_j) = N(A_{j+1})$ for all $j = 1, \dots, n - 2$.

The sequence (56) is called an *exact Banach sequence* iff it is exact and all the operators

$$A_j: X_j \rightarrow X_{j+1}, \quad j = 1, \dots, n - 1,$$

are linear and continuous, where X_1, \dots, X_n are Banach spaces over \mathbb{K} and the range $R(A_{n-1})$ is closed.⁶

In particular, the exactness of

$$X \xrightarrow{A} Y \xrightarrow{B} Z$$

⁶This implies that all the ranges $R(A_j)$, $j = 1, \dots, n - 1$, are closed. In fact, we have $R(A_j) = N(A_{j+1})$, and the null space $N(A_{j+1})$ is closed for all $j = 1, \dots, n - 2$, since A_{j+1} is continuous.

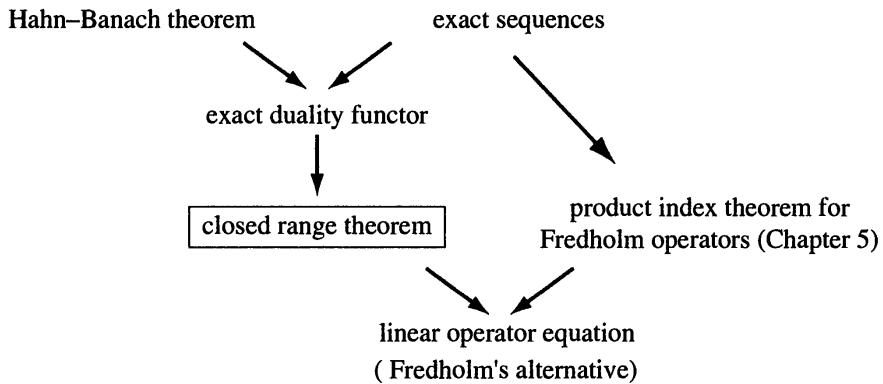


FIGURE 3.6.

means that $R(A) = N(B)$.

The following example shows that important operator properties can be translated into the language of exact sequences.

Example 2. Let $A: X \rightarrow Y$ be a linear operator, where X and Y are linear spaces over \mathbb{K} . Then

(i) A is *injective* iff the sequence

$$0 \longrightarrow X \xrightarrow{A} Y \quad (57)$$

is exact.

(ii) A is *surjective* iff the sequence

$$X \xrightarrow{A} Y \longrightarrow 0 \quad (58)$$

is exact.

(iii) A is *bijective* iff the sequence

$$0 \longrightarrow X \xrightarrow{A} Y \longrightarrow 0 \quad (59)$$

is exact.

Here, $0 \rightarrow X$ and $Y \rightarrow 0$ denote the trivial maps $0 \mapsto 0$ and $u \mapsto 0$, respectively.

Proof. Ad (i). The exactness of (57) means that $N(A) = \{0\}$.

Ad (ii). The exactness of (58) means that $R(A) = Y$.

Ad (iii). The exactness of (59) is equivalent to the exactness of (57) and (58). \square

Proposition 3. Let

$$0 \longrightarrow X \xrightarrow{A} Y \xrightarrow{B} Z \longrightarrow 0$$

be an exact sequence, where X , Y , and Z are finite-dimensional linear spaces over \mathbb{K} . Then

$$\dim X - \dim Y + \dim Z = 0.$$

Proof. The operator A is injective. Hence $\dim R(A) = \dim X$. Let W denote an algebraic complement of $N(B)$ in Y :

$$Y = N(B) \oplus W. \quad (60)$$

The operator B is surjective. Thus, the restriction $B: W \rightarrow Z$ is bijective, and hence $\dim Z = \dim W$. Since $N(B) = R(A)$, it follows from (60) that

$$\dim Y = \dim R(A) + \dim W = \dim X + \dim Z. \quad \square$$

Proposition 4. *The duality functor \mathcal{D} is exact, that is, \mathcal{D} sends exact Banach sequences to exact sequences.*

In Section 3.12 we will use Proposition 4 in order to prove the fundamental closed range theorem. This theorem tells us that the closedness of the range $R(A_1)$ in (56) implies $R(A_1^T) = N(A_1)^\perp$; thus the range $R(A_1^T)$ is also closed. Consequently, we get the following stronger result: the duality functor \mathcal{D} sends exact Banach sequences to exact Banach sequences.

Proof. Step 1: Let us first consider the *short* exact Banach sequence

$$X \xrightarrow{A} Y \xrightarrow{B} Z.$$

That is, $R(A) = N(B)$, and $R(B)$ is a closed linear subspace of Z . We have to show that

$$X^* \xleftarrow{A^T} Y^* \xleftarrow{B^T} Z^*$$

is an exact sequence, that is, $R(B^T) = N(A^T)$.

Since $R(A) = N(B)$, we get

$$BA = 0,$$

and hence $A^T B^T = (BA)^T = 0$. This implies $R(B^T) \subseteq N(A^T)$.

Conversely, let us show that $N(A^T) \subseteq R(B^T)$. To this end, choose $u^* \in N(A^T)$. Hence $u^* \in Y^*$ and

$$\langle u^*, Au \rangle = \langle A^T u^*, u \rangle = 0 \quad \text{for all } u \in X.$$

This yields $u^*(v) \equiv \langle u^*, v \rangle = 0$ for all $v \in R(A)$. Define

$$[u^*](v + R(A)) := u^*(v) \quad \text{for all } v \in Y.$$

It follows as in the proof of Proposition 6 in Section 3.8 that the linear functional

$$[u^*]: Y/R(A) \rightarrow \mathbb{K}$$

is *continuous*. Letting $[B](v + N(B)) := Bv$ for all $v \in Y$, we get the linear *homeomorphism*

$$[B]: Y/N(B) \rightarrow R(B),$$

by Proposition 6 in Section 3.8. Observe that the range $R(B)$ is closed. The decisive *trick* of our proof consists in introducing the linear functional v^* through the commutative diagram

$$\begin{array}{ccc} R(B) & \xrightarrow{v^*} & \mathbb{K} \\ [B]^{-1} \searrow & & \swarrow [u^*] \\ & Y/N(B) & \end{array}$$

that is, we set

$$v^* := [u^*][B]^{-1}. \quad (61)$$

Recall that $R(A) = N(B)$. The functional v^* is *continuous* on $R(B)$. Hence

$$|v^*(w)| \leq \text{const } \|w\| \quad \text{for all } w \in R(B),$$

where $R(B) \subseteq Z$. By the *Hahn–Banach theorem* (Theorem 1.B in Section 1.1), there exists a linear continuous extension

$$v^*: Z \rightarrow \mathbb{K}.$$

Relation (61) tells us that, for all $v \in Y$,

$$\begin{aligned} v^*(Bv) &= [u^*][B]^{-1}(Bv) \\ &= [u^*](v + N(B)) = [u^*](v + R(A)) = u^*(v). \end{aligned}$$

This yields

$$\langle v^*, Bv \rangle = \langle u^*, v \rangle \quad \text{for all } v \in Y,$$

and hence $u^* = B^T v^*$, which means $u^* \in R(B^T)$. Therefore, $N(A^T) \subseteq R(B^T)$.

Step 2: The general case can easily be reduced to Step 1. In fact, the sequence in (56) is an exact Banach sequence iff all the possible short sequences

$$X_j \xrightarrow{A_j} X_{j+1} \xrightarrow{A_{j+1}} X_{j+2}, \quad j = 1, \dots, n-2,$$

are exact Banach sequences. □

In the following two examples, let us apply the language of exact sequences to embeddings and projections.

Example 5. Let X be a closed linear subspace of the Banach space Y over \mathbb{K} , and let $j: X \rightarrow Y$ denote the trivial *embedding map* defined through $j(u) := u$ for all $u \in X$. Then, j is *injective*, that is, the sequence

$$0 \longrightarrow X \xrightarrow{j} Y .$$

is an *exact Banach sequence*. By Proposition 4, the dual sequence

$$0 \longleftarrow X^* \xleftarrow{j^T} Y^*$$

is also exact (i.e., the dual operator j^T is *surjective*).

Moreover, $N(j^T) = X^\perp$.

Proof. For all $u \in X$ and $u^* \in Y^*$,

$$\langle j^T(u^*), u \rangle_X = \langle u^*, j(u) \rangle_Y = \langle u^*, u \rangle_Y .$$

Therefore, the functional $j^T(u^*)$ represents the *restriction* of the functional $u^*: Y \rightarrow \mathbb{K}$ to the subspace X . Obviously,

$$j^T(u^*) = 0 \quad \text{iff} \quad u^* = 0 \text{ on } X$$

(i.e., $u^* \in X^\perp$). Hence $N(j^T) = X^\perp$. □

Example 6. Let X be a closed linear subspace of the Banach space Y over \mathbb{K} , and let

$$\pi: Y \rightarrow Y/X$$

be the canonical mapping from Section 3.8 defined through $\pi(u) := u + X$ for all $u \in Y$. Obviously, $N(\pi) = X$. Since π is linear, continuous, and *surjective*, the sequence

$$Y \xrightarrow{\pi} Y/X \longrightarrow 0$$

is exact. By Example 5,

$$0 \longrightarrow X \xrightarrow{j} Y \xrightarrow{\pi} Y/X \longrightarrow 0$$

is an *exact Banach sequence*. It follows from Proposition 4 that the dual sequence

$$0 \longleftarrow X^* \xleftarrow{j^T} Y^* \xleftarrow{\pi^T} (Y/X)^* \longleftarrow 0$$

is exact. Hence the dual operator π^T is *injective*, and $R(\pi^T) = N(j^T) = X^\perp$.

3.12 Applications to the Closed Range Theorem and to Fredholm Alternatives

The following result represents the most important theorem on linear operator equations.

Theorem 3.E (Banach's closed range theorem). *Let $A: X \rightarrow Y$ be a linear continuous operator where X and Y are Banach spaces over \mathbb{K} . Then the following three conditions are equivalent:*

- (i) Fredholm alternative. $R(A) = N(A^T)^\perp$ and $R(A^T) = N(A)^\perp$.
- (ii) Closed range. $R(A)$ is closed.
- (iii) A priori estimate. There is a constant $c > 0$ such that

$$c \cdot \text{dist}(u, N(A)) \leq \|Au\| \quad \text{for all } u \in X. \quad (62)$$

In terms of the operator equation

$$Au = b, \quad u \in X, \quad (\text{E})$$

and the dual equation

$$A^T u^* = b^*, \quad u^* \in Y^*, \quad (\text{E}^*)$$

Theorem 3.E(i) means the following.⁷ Let the range $R(A)$ be closed.

- (a) For given $b \in Y$, the original equation (E) has a solution iff

$$\langle u^*, b \rangle = 0 \quad (63)$$

for all solutions u^* of the homogeneous dual equation (E*).

- (b) Conversely, for given $b^* \in X^*$, the dual equation (E*) has a solution iff

$$\langle b^*, u \rangle = 0 \quad (64)$$

for all solutions u of the homogeneous original equation (E).

Observe that condition (63) is quite natural. In fact, if $Au = b$ and $A^T u^* = 0$, then

$$\langle u^*, b \rangle = \langle u^*, Au \rangle = \langle A^T u^*, u \rangle = 0. \quad (65)$$

⁷By definition, the homogeneous original equation and the homogeneous dual equation correspond to (E) with $b = 0$ and (E*) with $b^* = 0$, respectively.

Thus, (63) represents a simple *necessary* solvability condition for (E). The closed range theorem tells us that this condition is also a *sufficient* solvability condition provided that the range $R(A)$ is *closed*.

Furthermore, if $A^T u^* = b^*$ and $Au = 0$, then

$$\langle b^*, u \rangle = \langle A^T u^*, u \rangle = \langle u^*, Au \rangle = 0. \quad (66)$$

This is the solvability condition (64) for the dual equation (E^*) .

If X and Y are finite-dimensional spaces, then $R(A)$ is closed automatically. In this case, statements (a) and (b) correspond to classic results on finite linear systems.

Proof of Theorem 3.E. (i) \Rightarrow (ii). By Proposition 19 in Section 3.9, the set ${}^\perp N(A^T)$ is closed.

(ii) \Rightarrow (i). Let $R(A)$ be closed. According to Examples 5 and 6 in Section 3.11,

$$0 \longrightarrow N(A) \xrightarrow{j} X \xrightarrow{A} Y \xrightarrow{\pi} Y/R(A) \longrightarrow 0$$

represents an *exact Banach sequence*. By Proposition 4 in Section 3.11, the dual sequence

$$0 \longleftarrow N(A)^* \xleftarrow{j^T} X^* \xleftarrow{A^T} Y^* \xleftarrow{\pi^T} (Y/R(A))^* \longleftarrow 0$$

is *exact*. This implies

$$N(A^T) = R(\pi^T) \quad \text{and} \quad R(A^T) = N(j^T).$$

By Examples 5 and 6 in Section 3.11,

$$R(\pi^T) = R(A)^\perp \quad \text{and} \quad N(j^T) = N(A)^\perp.$$

Since $R(A)$ is *closed*, it follows from Proposition 20 in Section 3.9 that

$$R(A) = \overline{R(A)} = {}^\perp(R(A)^\perp).$$

Hence

$$R(A^T) = N(A)^\perp \quad \text{and} \quad R(A) = {}^\perp N(A^T).$$

(ii) \Rightarrow (iii). This is Proposition 6 in Section 3.8.

(iii) \Rightarrow (ii). First let $N(A) = \{0\}$. Then

$$c\|u\| \leq \|Au\| \quad \text{for all } u \in X.$$

This implies that $R(A)$ is closed. In fact, if $Au_n \rightarrow v$ as $n \rightarrow \infty$, then (Au_n) is Cauchy, and $c\|u_n - u_m\| \leq \|Au_n - Au_m\|$ shows that (u_n) is also Cauchy. Hence $u_n \rightarrow u$ as $n \rightarrow \infty$, that is, $Au = v$.

If $N(A) \neq \{0\}$, then we use the operator

$$[A]: X/N(A) \longrightarrow Y$$

from Proposition 6 in Section 3.8. Recall that $[A][u] := Au$ for all $u \in X$. Thus, the a priori estimate in (62) is equivalent to

$$c\|[u]\| \leq \| [A][u] \| \quad \text{for all } [u] \in X/N(A).$$

The same argument as the preceding one shows that $R([A])$ is closed. Since $R(A) = R([A])$, the range $R(A)$ is also closed. \square

Corollary 1 (Closed range theorem for Hilbert spaces). *Let $A: X \rightarrow X$ be a linear continuous operator on the Hilbert space X over \mathbb{K} . Then the following two conditions are equivalent:⁸*

- (i) $R(A) = N(A^*)^\perp$ and $R(A^*) = N(A)^\perp$.
- (ii) $R(A)$ is closed.

In terms of the operator equation

$$Au = b, \quad u \in X, \tag{E}$$

and the adjoint equation

$$A^*v = c, \quad v \in X, \tag{E_a}$$

this means the following. Let the range $R(A)$ be closed.

- (a) For given $b \in X$, the original equation (E) has a solution iff

$$(v \mid b) = 0$$

for all solutions v of the homogeneous adjoint equation (E_a).

- (b) For given $c \in X$, the adjoint equation (E_a) has a solution iff

$$(c \mid u) = 0$$

for all solutions u of the homogeneous original equation (E).

Proof of Corollary 1. By Proposition 2 in Section 3.10, we have

$$A^* = J^{-1}A^TJ,$$

and $\langle Ju, v \rangle = (u \mid v)$ for all $u, v \in X$. The assertion follows now from Theorem 3.E.

⁸Here, the symbol \perp denotes the *orthogonal* complement (cf. Section 2.9 in AMS Vol. 108).

In fact, by Theorem 3.E, the original equation (E) has a solution iff

$$\langle u^*, b \rangle = 0 \quad \text{for all } u^* \text{ with } A^T u^* = 0.$$

If we let $v := J^{-1}u^*$, this is equivalent to

$$(v \mid b) = 0 \quad \text{for all } v \text{ with } J^{-1}A^T Jv = 0,$$

that is, $b \in N(A^*)^\perp$. Hence $R(A) = N(A^*)^\perp$.

Moreover, the adjoint equation (E_a) can be written as

$$J^{-1}A^T(Jv) = J^{-1}c.$$

By Theorem 3.E, this equation has a solution iff

$$\langle J^{-1}c, u \rangle = 0 \quad \text{for all } u \text{ with } Au = 0.$$

This is equivalent to

$$(c \mid u) = 0 \quad \text{for all } u \text{ with } Au = 0$$

(i.e., $c \in N(A)^\perp$). Hence $R(A^*) = N(A)^\perp$. \square

Standard Example 2. Let $A: X \rightarrow Y$ be a linear continuous operator, where X and Y are Banach spaces over \mathbb{K} .

If $\text{codim } R(A) < \infty$, then the range $R(A)$ is closed.

Proof. Choose a linear subspace Z of Y such that

$$Y = R(A) \oplus Z.$$

As in Proposition 6 of Section 3.8, define the linear continuous injective operator $[A]: X/N(A) \rightarrow Y$ through

$$[A][u] := Au \quad \text{for all } [u] \in X/N(A).$$

Finally, set

$$B([u], z) := [A][u] + z \quad \text{for all } [u] \in X/N(A), z \in Z.$$

Then, the operator

$$B: (X/N(A)) \times Z \rightarrow Y$$

is linear, continuous, and bijective. In fact, $[A][u] + z = 0$ implies $[A][u] = 0$ along with $z = 0$ (i.e., $[u] = 0$). Since $\dim Z < \infty$, both Z and $X/N(A)$ are Banach spaces. Thus, B is a linear *homeomorphism*, by the *continuous inverse theorem* in Section 3.5. The set

$$W := \{([u], 0) : [u] \in X/N(A)\}$$

is closed in the product space $(X/N(A)) \times Z$. Hence the set $B(W)$ is also *closed*, by Lemma 12 in Section 3.9. Finally, observe that

$$R(A) = R([A]) = B(W).$$

Thus, the range $R(A)$ is closed. \square

Standard Example 3. Let $A: X \rightarrow Y$ be a linear continuous operator, where X and Y are Banach spaces over \mathbb{K} . Then the following two conditions are equivalent:

- (i) *A priori estimate.* There is a constant $c > 0$ such that

$$c\|u\| \leq \|Au\| \quad \text{for all } u \in X.$$

- (ii) The range $R(A)$ is closed and $Au = 0$ implies $u = 0$.

Proof. Observe that $\text{dist}(u, N(A)) = \|u\|$ if $N(A) = \{0\}$, and use Theorem 3.E. \square

Standard Example 4. Let $A: X \rightarrow Y$ be a linear continuous operator, where X and Y are Banach spaces over \mathbb{K} . Furthermore, let Z be a Banach space over \mathbb{K} such that the embedding

$$X \subseteq Z$$

is *compact*. Then the following two statements are equivalent:

- (i) *A priori estimate.* There is a constant $c > 0$ such that

$$c\|u\|_X \leq \|Au\|_Y + \|u\|_Z \quad \text{for all } u \in X. \quad (67)$$

- (ii) The range $R(A)$ is closed and $\dim N(A) < \infty$.

This result plays an important role in the theory of elliptic-type linear partial differential equations (cf. Lions and Magenes (1972), Vol. 1, Chapter 2, Section 5.2).

Proof. (i) \Rightarrow (ii). Since A is continuous, the null space $N(A)$ is closed, and hence $N(A)$ is a Banach space with respect to the norm $\|\cdot\|_X$. Let B be the closed unit ball in $N(A)$. We want to show that B is *compact*. Then $\dim N(A) < \infty$, by Section 2.3.

In fact, let (u_n) be a sequence in B . Since the embedding $X \subseteq Z$ is compact, the set B is relatively compact in Z . Thus, there exists a subsequence,

again denoted by (u_n) , such that $u_n \rightarrow u$ in Z as $n \rightarrow \infty$. Since $Au_n = 0$ for all n , it follows from (67) that

$$c\|u_n - u_m\|_X \leq \|u_n - u_m\|_Z \quad \text{for all } n, m.$$

Hence (u_n) is Cauchy in X . Thus, B is compact.

We now prove that the range $R(A)$ is *closed*. Since $\dim N(A) < \infty$, there exists a continuous projection $P: X \rightarrow X$ onto $N(A)$. Set $L := (I - P)(X)$. Then

$$X = N(A) \oplus L.$$

The operator $A: L \rightarrow R(A)$ is bijective on the closed linear subspace L of X . This implies the existence of a constant $d > 0$ such that

$$\|u\|_X \leq d\|Au\|_Y \quad \text{for all } u \in L. \quad (68)$$

Otherwise, there would exist a sequence (u_n) in L such that

$$\|u_n\|_X = 1 \quad \text{for all } n, \quad (69)$$

and $Au_n \rightarrow 0$ in Y as $n \rightarrow \infty$. Since the embedding $X \subseteq Z$ is compact, there is a subsequence, again denoted by (u_n) , such that $u_n \rightarrow v$ in Z as $n \rightarrow \infty$. By (67),

$$c\|u_n - u_m\|_X \leq \|Au_n - Au_m\|_Y + \|u_n - u_m\|_Z$$

(i.e., (u_n) is Cauchy in X), and hence

$$u_n \rightarrow u \quad \text{in } X \quad \text{as } n \rightarrow \infty.$$

This implies $u \in L$ and $Au = 0$. Hence $u = 0$ because $X = N(A) \oplus L$. From (69) we get $\|u\|_X = 1$. This contradicts $u = 0$.

(ii) \Rightarrow (i). From $X = N(A) \oplus L$ we obtain the decomposition

$$u = u_1 + u_2, \quad u_1 \in N(A), \quad u_2 \in L,$$

for all $u \in X$. Hence

$$\|u\|_X \leq \|u_1\|_X + \|u_2\|_X \quad \text{for all } u \in X. \quad (70)$$

All the norms are equivalent on the *finite-dimensional space* $N(A)$ (cf. Section 1.12 in AMS Vol. 108). Hence

$$\|u_1\|_X \leq \text{const}\|u_1\|_Z \quad \text{for all } u_1 \in N(A).$$

If we use $u_1 = u - u_2$, this implies

$$\|u_1\|_X \leq \text{const}(\|u\|_Z + \|u_2\|_Z).$$

Since the embedding $X \subseteq Z$ is continuous, we have $\|v\|_Z \leq \text{const}\|v\|_X$ for all $v \in X$, and hence

$$\|u_1\|_X \leq \text{const}(\|u\|_Z + \|u_2\|_X).$$

Thus, it follows from (70) that

$$\|u\|_X \leq \text{const}(\|u\|_Z + \|u_2\|_X). \quad (71)$$

Finally, by the *continuous inverse theorem* from Section 3.5, the operator $A: L \rightarrow R(A)$ is a linear homeomorphism. This implies (68), namely,

$$\|u_2\|_X \leq \text{const} \|Au_2\|_Y \quad \text{for all } u_2 \in L. \quad (72)$$

From (71) and (72) we get the desired inequality (67), since $Au_2 = Au$. \square

Problems

3.1. The Baire category. Let M be a set of the first Baire category in the Banach space X over \mathbb{K} , and let N be a nonempty open subset of X .

Show that $N - M$ is dense in N .

Solution: We have to prove that $N \subseteq \overline{(N - M)}$. If this is not true, then the set

$$S := N - \overline{(N - M)}$$

is nonempty and open. Hence S is of the second Baire category in X , by Theorem 3.A. Since

$$N - \overline{(N - M)} \subseteq N - (N - M) = M,$$

the set S is of the first Baire category. This is a contradiction.

3.2. Examples. Determine the Baire category of the following sets M in X :

- (i) $X := \mathbb{R}$ and $M := \{\xi \in X: \sin \xi = 0\}$.
- (ii) $X := \mathbb{R}$ and $M := \{\xi \in X: 0 \leq \sin \xi \leq 1\}$.
- (iii) $X := \mathbb{R}^2$ and $M := \{(\xi, \eta) \in X: \xi^2 + \eta^2 < 1\}$.
- (iv) $X := \mathbb{R}^2$ and $M := \{(\xi, \eta) \in X: \xi^2 + \eta^2 = 1\}$.
- (v) $X := \mathbb{R}^2$ and $M := \{(\xi, \eta) \in X: \xi + \eta = \rho, \rho = \text{rational number}\}$.
- (vi) $X := \mathbb{R}^2$ and $M := \{(\xi, \eta) \in X: 0 \leq \xi, \eta \leq 1\}$.

Solution: Cf. Problem 3.20.

3.3. Topological direct sum. Let

$$X = X_1 \oplus X_2$$

be a direct sum, where X is a Banach space over \mathbb{K} . Let

$$u = u_1 + u_2, \quad u_1 \in X_1, u_2 \in X_2,$$

be the corresponding decomposition for each $u \in X$. Show that the following three statements are mutually equivalent:

- (i) $X = X_1 \oplus X_2$ is a topological direct sum.
- (ii) The map $u \mapsto (u_1, u_2)$ is a linear homeomorphism from X onto the product space $X_1 \times X_2$.
- (iii) The norm $\|u\|_* := \|u_1\| + \|u_2\|$ is equivalent to the original norm $\|u\|$ on X .

Hint: Use the continuous inverse mapping theorem.

3.4. A variant of the Banach–Steinhaus theorem. Let (L_n) be a sequence of linear continuous operators $L_n: X \rightarrow Y$, where X is a Banach space over \mathbb{K} and Y is a normed space over \mathbb{K} . Suppose that the limit

$$Lu := \lim_{n \rightarrow \infty} L_n u$$

exists for all $u \in X$. Show that

$$\|L\| \leq \varliminf_{n \rightarrow \infty} \|L_n\| < \infty.$$

Solution: By the Banach–Steinhaus theorem (Corollary 1 in Section 3.3), $\sup_n \|L_n\| < \infty$. It follows from $\|L_n u\| \leq \|L_n\| \|u\|$ that

$$\|Lu\| = \varliminf_{n \rightarrow \infty} \|L_n u\| \leq \varliminf_{n \rightarrow \infty} \|L_n\| \|u\| \quad \text{for all } u \in X.$$

3.5. Weak convergence. Let X be a normed space X over \mathbb{K} . Show that

- (a) If $u_n^* \rightarrow u^*$ in X^* and $u_n \rightharpoonup u$ in X as $n \rightarrow \infty$, then

$$\langle u_n^*, u_n \rangle \rightarrow \langle u^*, u \rangle \quad \text{as } n \rightarrow \infty.$$

- (b) If $u_n \rightharpoonup u$ in X as $n \rightarrow \infty$, then

$$\|u\| \leq \varliminf_{n \rightarrow \infty} \|u_n\|.$$

- (c) If X is reflexive, then $u_n^* \rightharpoonup u^*$ in X^* as $n \rightarrow \infty$ is equivalent to

$$\langle u_n^*, u \rangle \rightarrow \langle u^*, u \rangle \quad \text{as } n \rightarrow \infty \text{ for all } u \in X.$$

- (d) If X is reflexive, then $u_n^* \rightharpoonup u^*$ in X^* and $u_n \rightharpoonup u$ in X as $n \rightarrow \infty$ imply

$$\langle u_n^*, u_n \rangle \rightarrow \langle u^*, u \rangle \quad \text{as } n \rightarrow \infty.$$

- (e) If X is a Hilbert space, then $u_n \rightharpoonup u$ in X and $\|u_n\| \rightarrow \|u\|$ as $n \rightarrow \infty$ imply $u_n \rightarrow u$ as $n \rightarrow \infty$.

Solution: Recall that $\langle v^*, v \rangle := v^*(v)$ and hence

$$|\langle v^*, v \rangle| \leq \|v^*\| \|v\| \quad \text{for all } v \in X, v^* \in X^*.$$

Ad (a). Since (u_n) is bounded, we get

$$\begin{aligned} |\langle u_n^*, u_n \rangle - \langle u^*, u \rangle| &= |\langle u_n^* - u^*, u_n \rangle + \langle u^*, u_n - u \rangle| \\ &\leq \|u_n^* - u^*\| \sup_n \|u_n\| + |\langle u^*, u_n - u \rangle| \rightarrow 0 \quad \text{as } n \rightarrow \infty. \end{aligned}$$

Ad (b). Use Problem 3.4 and the fact that X^* is a Banach space (cf. the proof of Example 3 in Section 3.3).

Ad (c). Use the definition of reflexive normed spaces in Section 2.8.

Ad (d). Use (c) and an analogous argument as in the proof of (a).

Ad (e). Since $(u_n | u) \rightarrow (u | u)$ and $(u | u_n) \rightarrow (u | u)$ as $n \rightarrow \infty$, we get

$$\|u - u_n\|^2 = \|u\|^2 - (u_n | u) - (u | u_n) - \|u_n\|^2 \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Cf. Zeidler (1986), Vol. 2B, Proposition 21.23.

3.6. Weak* convergence. Let X be a Banach space over \mathbb{K} , and let (u_n^*) be a sequence in the dual space X^* . We write

$$u_n^* \xrightarrow{*} u^* \quad \text{in } X^* \quad \text{as } n \rightarrow \infty$$

iff $\langle u_n^*, u \rangle \rightarrow \langle u^*, u \rangle$ as $n \rightarrow \infty$ for all $u \in X$. This is the so-called weak* convergence. Show that the following are true:

(a) If $u_n^* \xrightarrow{*} u^*$ in X^* as $n \rightarrow \infty$, then (u_n^*) is bounded in X^* and

$$\|u^*\| \leq \liminf_{n \rightarrow \infty} \|u_n^*\|.$$

(b) If $u_n^* \rightharpoonup u^*$ in X^* and $u_n \rightarrow u$ in X as $n \rightarrow \infty$, then

$$\langle u_n^*, u_n \rangle \rightarrow \langle u^*, u \rangle \quad \text{as } n \rightarrow \infty.$$

(c) If X is *reflexive*, then $u_n^* \xrightarrow{*} u^*$ in X^* iff $u_n^* \rightharpoonup u$ in X^* as $n \rightarrow \infty$.

(d) If X is *separable*, then each bounded sequence (u_n^*) in X^* has a subsequence $(u_{n'}^*)$ such that $u_{n'}^* \xrightarrow{*} u^*$ in X^* as $n' \rightarrow \infty$.

(e) Let (u_n^*) be a sequence in X^* . Then $u_n^* \xrightarrow{*} u^*$ in X^* as $n \rightarrow \infty$ iff (u_n^*) is bounded and there exists a dense subset D of X such that

$$\langle u_n^*, v \rangle \rightarrow \langle u^*, v \rangle \quad \text{as } n \rightarrow \infty \text{ for all } v \in D \text{ and fixed } u^* \in X^*.$$

Solution: Ad (a). Use Problem 3.4.

Ad (b), (c). Use similar arguments as in Problem 3.5.

Ad (d). Use a similar argument as in the proof of Proposition 6 in Section 2.8. To this end, let $\{v_k\}$ be a countable dense subset of X . By a diagonal procedure, we obtain a subsequence (w_n^*) of (u_n^*) such that

$$\langle w_n^*, v_k \rangle \rightarrow a_k \quad \text{as } n \rightarrow \infty \text{ for all } k.$$

Since $\{v_k\}$ is dense in X and (w_n^*) is bounded in X^* , it follows that, as $n \rightarrow \infty$, the limit $\langle w_n^*, v \rangle \rightarrow a(v)$ exists for all $v \in X$. From

$$|\langle w_n^*, v \rangle| \leq \sup_n \|w_n^*\| \|v\|,$$

we get $|a(v)| \leq \text{const} \|v\|$ for all $v \in X$, and hence $a \in X^*$. Thus, $\langle w_n^*, v \rangle \rightarrow \langle a, v \rangle$ for all $v \in X$.

Ad (e). Use the same argument as in the proof of Example 3 in Section 3.3.

Cf. Zeidler (1986), Vol. 2B, Proposition 21.26.

3.7. Subsequences. Show that a sequence (u_n) in a Banach space X over \mathbb{K} has the following convergence properties:

- (i) *Strong convergence.* Let u be a fixed element of X . If every subsequence of (u_n) has, in turn, a subsequence that converges strongly to u in X , then the original sequence (u_n) converges strongly to u (i.e., $u_n \rightarrow u$ in X as $n \rightarrow \infty$).
- (ii) *Weak convergence.* Let u be a fixed element in X . If every subsequence of (u_n) has, in turn, a subsequence that converges weakly to u , then the original sequence converges weakly to u (i.e., $u_n \rightharpoonup u$ in X as $n \rightarrow \infty$).
- (iii) *Bounded sequences.* Let (u_n) be a bounded sequence in the *reflexive* Banach space X . If all the weakly convergent subsequences of (u_n) have the same limit u , then (u_n) converges weakly to u (i.e., $u_n \rightharpoonup u$ in X as $n \rightarrow \infty$).

Hint: Cf. Zeidler (1986), Vol. 1, Section 10.5.

3.8. Compact operators and weak convergence. Let $A: X \rightarrow Y$ be a linear operator, where X and Y are Banach spaces over \mathbb{K} . Show that

- (i) If A is compact, then A is *strongly continuous*, that is, as $n \rightarrow \infty$,

$$u_n \rightarrow u \quad \text{implies} \quad Au_n \rightarrow Au.$$

- (ii) Conversely, if A is strongly continuous and X is reflexive, then A is compact.

Hint: Cf. Zeidler (1986), Vol. 2A, Proposition 21.29.

The following problems summarize important properties of reflexive Banach spaces. The statement in Problem 3.17 represents a deep result of functional analysis (the Eberlein–Šmuljan theorem).

3.9. Invariance of reflexivity under normisomorphisms. Let X and Y be normed spaces over \mathbb{K} with

$$X \cong Y$$

(i.e., X is normisomorphic to Y). Show that X is reflexive iff Y is reflexive.

3.10. Reflexivity of closed linear subspaces. Let L be a closed linear subspace of the reflexive Banach space X over \mathbb{K} . Then L is also reflexive.

This has been proved in Section 2.8.

3.11. Reflexivity of the dual space. Show that a Banach space X over \mathbb{K} is reflexive iff the dual space X^* is reflexive.

Hint: If X is reflexive, then the reflexivity of X^* follows by using a simple argument based on the surjectivity of the operator j from Section 2.8 (cf. Holmes (1975), p. 126).

Conversely, if X^* is reflexive, then so is X^{**} , by the preceding argument. Using the map $j: X \rightarrow X^{**}$, we obtain that $j(X)$ is a closed linear subspace of X^{**} with $X \cong j(X)$. Now use Problems 3.9 and 3.10.

3.12. Reflexivity of product spaces. Let X and Y be normed spaces over \mathbb{K} . Set

$$X \times Y := \{(u, v) : u \in X, v \in Y\}, \quad \|(u, v)\| := \|u\| + \|v\|,$$

$$X^* \times Y^* = \{(u^*, v^*) : u^* \in X^*, v^* \in Y^*\}, \quad \|(u^*, v^*)\|_* := \max\{\|u^*\|, \|v^*\|\}.$$

Show that

- (i) $X^* \times Y^*$ is a normed space over \mathbb{K} equipped with the norm $\|\cdot\|_*$.
- (ii) $(X \times Y)^* \cong X^* \times Y^*$.
- (iii) If X and Y are reflexive, then so is the product space $X \times Y$.

More precisely, there exists a *normisomorphism* $\mathcal{J}: X^* \times Y^* \rightarrow (X \times Y)^*$ given through

$$\mathcal{J}(u^*, v^*)(u, v) := u^*(u) + v^*(v),$$

for all $(u, v) \in X \times Y$ and $(u^*, v^*) \in X^* \times Y^*$.

3.13. Reflexivity of factor spaces. Let X be a Banach space over \mathbb{K} , and let L be a closed linear subspace of X . Recall that

$$L^\perp := \{u^* \in X^* : \langle u^*, u \rangle = 0 \text{ for all } u \in L\}.$$

Show that

- (i) L^\perp is a closed linear subspace of X^* .
- (ii) There exists a functional $u^* \in L^\perp$ with $\|u^*\| = 1$ and $\langle u^*, u \rangle = 1$ for some $u \in X$, provided $L \neq X$.
- (iii) $(X/L)^* \cong L^\perp$.
- (iv) If X is reflexive, then so is the factor space X/L .

More precisely, there exists a *normisomorphism* $\mathcal{J}: (X/L)^* \rightarrow L^\perp$ given through

$$\mathcal{J}(u^*)([u]) := u^*(u) \quad \text{for all } u \in X.$$

Recall that the elements $[u]$ of X/L are the sets $[u] := u + L$.

Hint: Use the Hahn–Banach theorem in (ii). The proof of (iii) is based on (ii). In order to prove (iv), use (iii) along with Problems 3.9 and 3.11.

3.14. Dual operators. Let $A: X \rightarrow X$ be a linear continuous operator on the *reflexive* Banach space X over \mathbb{K} . Show that

$$A^{TT} = A.$$

This relation corresponds to $A^{**} = A$ in Hilbert spaces.

3.15.* Embeddings. Let X and Y be Banach spaces over \mathbb{K} such that the embedding

$$X \subseteq Y$$

is continuous, and X is *dense* in Y . Show that

- (i) The embedding $Y^* \subseteq X^*$ is continuous.
- (ii) If X is reflexive, then Y^* is dense in X^* .

Hint: Use the Hahn–Banach theorem. Cf. Zeidler (1986), Vol. 2A, p. 98.

3.16. Weak topology. Let X be a normed space over \mathbb{K} . A subset W of X is called *weakly open* iff, for each point $u_0 \in W$, there is a number $\varepsilon > 0$ and there are finitely many functionals $f_1, \dots, f_n \in X^*$ such that the set

$$\{u \in X : |f_j(u - u_0)| < \varepsilon, \quad j = 1, \dots, n\}$$

is contained in W . Show that

- (i) All the weakly open subsets of X form a separated *topology* (cf. Problem 1.12).

This is called the *weak topology* of X . By weak closedness, weak compactness, and so forth, we understand closedness, compactness, and so on with respect to the *weak topology*.

- (ii) The weak convergence is identical to the convergence with respect to the weak topology.
- (iii) In a finite-dimensional normed space over \mathbb{K} , the weak topology is identical to the usual topology induced by the norm.

Important Remark (The shortcoming of classic sequences in general topological spaces). We have shown in Problems 1.15 and 1.18 that if X is a metric space (e.g., X is a normed space), then a subset M of X is compact iff it is *sequentially* compact.

Unfortunately, this result is *not* valid in general topological spaces (e.g., normed spaces equipped with the weak topology). In order to characterize compact sets by means of convergence in general topological spaces, one needs *generalized sequences* (Moore–Smith sequences).

Important results are summarized in the appendix to Zeidler (1986), Vol. 1, pp. 758ff.

3.17. Weak compactness and reflexivity.** Let X be a Banach space over \mathbb{K} . Then the following three fundamental statements are mutually equivalent:

- (i) X is reflexive.
- (ii) Each bounded sequence in X has a weakly convergent subsequence.
- (iii) The closed unit ball B in X is *weakly compact*.

Study the proof in Rolewicz (9172), Chapter 5. Also see Holmes (1975), pp. 126 and 149, and Dunford and Schwarz (1958), Vol. 1.

Recall that

$$B \text{ is } \textit{compact} \text{ iff } \dim X < \infty,$$

by Section 2.3. Thus, the closed unit ball B of X carries important information about the structure of the Banach space X .

3.18. Reflexivity of finite-dimensional normed spaces. Show that each finite-dimensional normed space over \mathbb{K} is reflexive.

Hint: An elementary proof follows from Proposition 4 in Section 1.21 of AMS Vol. 108.

The statement is also an immediate consequence of Problems 3.16(iii) and 3.17.

3.19. Locally convex spaces. By a *seminorm* p on the linear space X over \mathbb{K} , we understand a function $p: X \rightarrow [0, \infty[$ such that

$$p(u + v) \leq p(u) + p(v) \quad \text{and} \quad p(\alpha u) = |\alpha|p(u)$$

for all $u, v \in X$ and $\alpha \in \mathbb{K}$. Obviously, each norm is a seminorm. Conversely, a seminorm is a norm iff $p(u) = 0$ implies $u = 0$.

By definition, a *locally convex space* consists of a linear space X over \mathbb{K} together with a system of seminorms $\{p_j\}_{j \in J}$ on X such that, for $u \in X$,

$$u = 0 \quad \text{iff} \quad p_j(u) = 0 \quad \text{for all } j \in J.$$

A subset W of X is called *open* iff, for each point $u_0 \in W$, there is a number $\varepsilon > 0$ and finitely many seminorms p_{j_1}, \dots, p_{j_n} such that the set

$$\{u \in X : p_{j_k}(u - u_0) < \varepsilon, \quad k = 1, \dots, n\}$$

is contained in W . Show that

- (i) These open sets form a separated *topology* on X (cf. Problem 1.12f).
- (ii) Each normed space X equipped with the *weak topology* is a *locally convex space* with respect to the system of seminorms $\{|f|\}_{f \in X^*}$.

Historical Remark. The theory of locally convex spaces was developed in the 1950s, motivated by the observation that *spaces of generalized functions* are locally convex spaces but *not* normed spaces (cf. the appendix to Zeidler (1986), Vol. 2B, pp. 1056ff).

3.20. Solution to Problem 3.2. The sets M in (i), (iv), and (v) are of the first Baire category in X , whereas the remaining sets are of the second Baire category in X .

:

4

The Implicit Function Theorem

Data aequatione quotcunque fluentes quantitae involvente fluxiones invenire et vice versa.¹

Isaac Newton to Leibniz, 1676

It is worth noting that the notation facilitates discovery. This, in a most wonderful way, reduces the mind's labor.

Gottfried Wilhelm Leibniz

In this chapter let us consider some basic facts about the differential calculus for operators. The main strategy encompasses the following:

- (i) *Differentiation means linearization.*
- (ii) *Higher derivatives correspond to multilinearization.*

¹ “It is useful to differentiate functions and to solve differential equations.” More precisely, Newton communicated his discovery to Leibniz in the following form:

6a cc d ae 13e ff 7i 3l 9n 4o 4q rr 4s 9t 12v x.

This decodes into the Latin sentence above, which must have been incomprehensible to Leibniz, although Leibniz too discovered differential calculus at about the same time. It is said that more ingenuity is required to decode this anagram than to discover differential calculus.

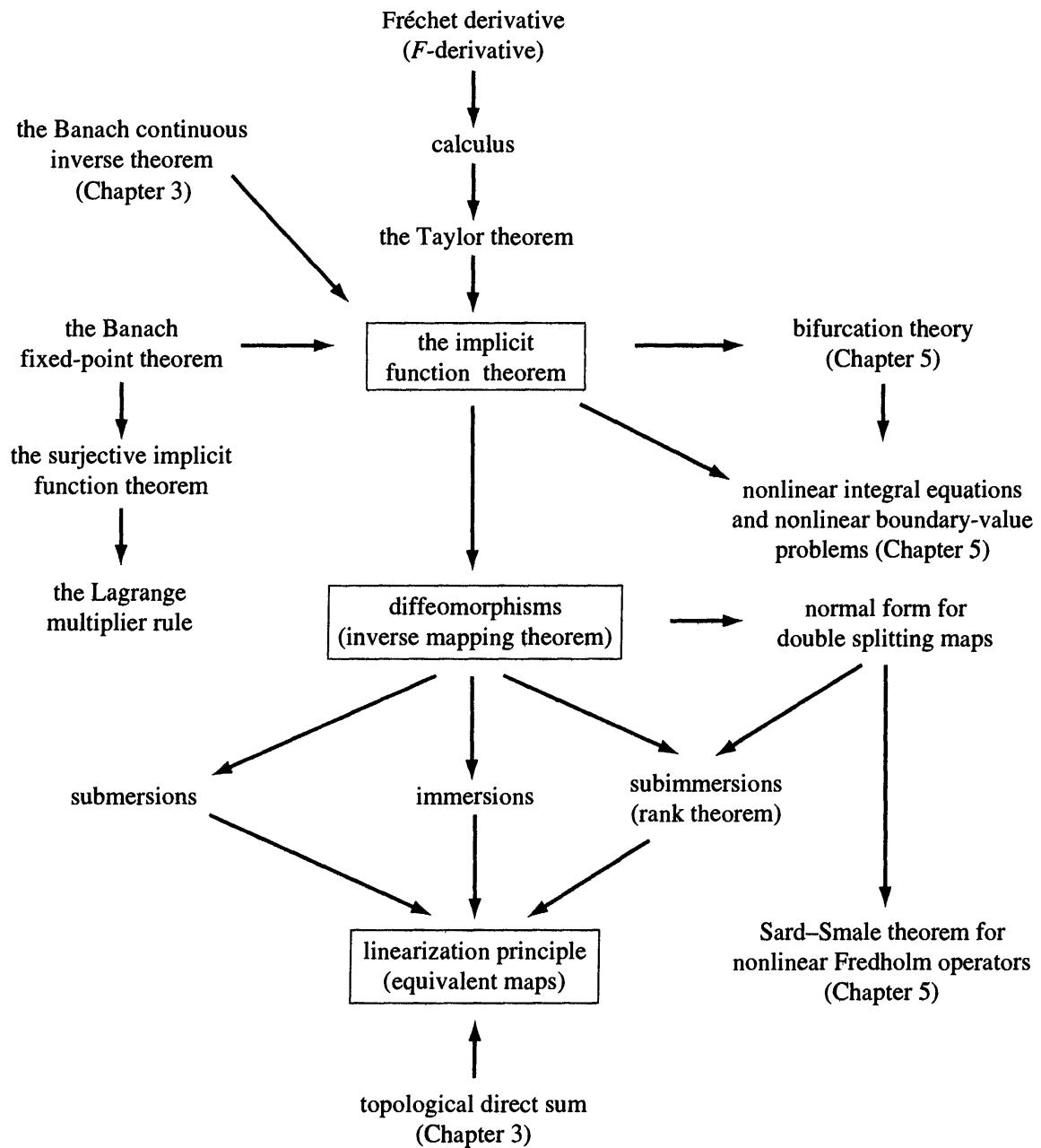


FIGURE 4.1.

The fundamental implicit function theorem on the unique local solvability of parameter-dependent operator equations is a consequence of the *Banach fixed-point theorem* combined with calculus. In fact, the implicit function theorem represents a *cornerstone* of nonlinear analysis. Important applications are displayed in Figure 4.1.²

We will introduce a notation that produces formulas that are as simple as possible.

²Many other applications of the implicit function theorem are studied in Zeidler (1986), Vol. 1, Chapter 4, and Vol. 4, Chapter 73 (Applications to Banach Manifolds).

4.1 m -Linear Bounded Operators

Definition 1. Let X_1, \dots, X_m and Y be Banach spaces over \mathbb{K} . The mapping

$$M: X_1 \times \cdots \times X_m \rightarrow Y$$

is called *m -linear and bounded* iff M is linear in each argument and there is a constant $C \geq 0$ such that

$$\|M(u_1, \dots, u_m)\| \leq C\|u_1\| \|u_2\| \cdots \|u_m\| \quad (1)$$

for all $u_j \in X_j$, $j = 1, \dots, m$.

The *norm* of M is defined by

$$\|M\| := \sup_{\|u_j\| \leq 1, j=1, \dots, m} \|M(u_1, \dots, u_m)\|$$

so that

$$\|M(u_1, \dots, u_m)\| \leq \|M\| \|u_1\| \|u_2\| \cdots \|u_m\|$$

for all $u_j \in X_j$, $j = 1, \dots, m$.

Some examples will be considered in Section 4.3.

Proposition 2. *Each m -linear bounded operator is continuous.*

Proof. For example, let $m = 2$. Suppose that

$$(u_n, v_n) \rightarrow (u, v) \text{ in } X_1 \times X_2 \quad \text{as } n \rightarrow \infty.$$

Then $u_n \rightarrow u$ and $v_n \rightarrow v$ as $n \rightarrow \infty$, and hence (u_n) and (v_n) are bounded. Thus,

$$\begin{aligned} \|M(u_n, v_n) - M(u, v)\| &= \|M(u_n - u, v_n) + M(u, v_n - v)\| \\ &\leq \|M\| \|u_n - u\| \|v_n\| + \|M\| \|u\| \|v_n - v\| \rightarrow 0 \end{aligned}$$

as $n \rightarrow \infty$. □

In the following, we write

$$r(h) = o(\|h\|^m), \quad h \rightarrow 0$$

iff $\frac{\|r(h)\|}{\|h\|^m} \rightarrow 0$ as $h \rightarrow 0$. In order to simplify notation, we set

$$Mu_1 u_2 \dots u_m \equiv M(u_1, \dots, u_m).$$

4.2 The Differential of Operators and the Fréchet Derivative

The *key formulas* are given by the decomposition

$$f(u + h) - f(u) = df(u)h + o(\|h\|), \quad h \rightarrow 0 \quad (2)$$

and

$$f'(u) \equiv df(u),$$

as well as

$$df(u + h)k - df(u)k = d^2 f(u)hk + r \quad (3)$$

with the “small” *remainder* r , namely,

$$\sup_{\|k\| \leq 1} \|r(u; h, k)\| = o(\|h\|), \quad h \rightarrow 0 \quad (4)$$

and

$$f''(u) \equiv d^2 f(u).$$

Definition 1. Let $f: U(U) \subseteq X \rightarrow Y$ be a given operator defined on an open neighborhood of the point u , where X and Y are Banach spaces over \mathbb{K} .

- (i) The *differential* $df(u)$ of f at the point u exists iff there is a *linear bounded* operator denoted by

$$df(u): X \rightarrow Y$$

such that (2) holds for all $h \in X$ in some open neighborhood of $h = 0$ in X .

Synonymously, we also write $f'(u)$ instead of $df(u)$ and we call $f'(u)$ the *F-derivative*³ of f at the point u .

- (ii) The *second differential* $d^2 f(u)$ of f at the point u exists iff there is a *bilinear bounded operator* denoted by

$$d^2 f(u): X \times X \rightarrow Y$$

such that (3) and (4) hold for all $k \in X$ and all h in some open neighborhood of $h = 0$ in X .

³*F*-derivative stands for “Fréchet derivative.”

Synonymously, we also write $f''(u)$ instead of $d^2 f(u)$ and we call $f''(u)$ the *second F-derivative* of f at u .

Roughly speaking, $df(u) \equiv f'(u)$ represents a *linearization* of the operator f at the point u .

Moreover, the linearization of $df(u)k \equiv f'(u)k$ at the point u leads to the bilinear operator $d^2 f(u)hk \equiv f''(u)hk$.

It follows easily that $df(u)$ is *uniquely determined* by (2). In fact,

$$df(u)h = \lim_{t \rightarrow 0} t^{-1}[f(u + th) - f(u)] \equiv \frac{d}{dt} f(u + th)|_{t=0}, \quad (5)$$

provided $df(u)$ exists. Analogously, the existence of $d^2 f(u)$ implies

$$d^2 f(u)hk = \frac{d}{dt} df(u + th)k|_{t=0}. \quad (6)$$

Remark 2. Formulas (5) and (6) are frequently used in the following way:

- (a) One formally computes $df(u)$ by means of (5).
- (b) One justifies this by verifying the decomposition in (2).

The same method works nicely in the case of $d^2 f(u)$. Here one has to use (6) along with (3).

Similarly to (3) and (4), the $(n+1)$ th *differential* $d^{n+1} f(u)$ at the point u is defined through *induction* by means of the following formula:

$$d^n f(u + h)h_1 \dots h_n - d^n f(u)h_1 \dots h_n = d^{n+1} f(u)hh_1 h_2 \dots h_n + r \quad (7)$$

with the “small” *remainder*

$$\sup_{\|h_j\| \leq 1, j=1, \dots, n} \|r(u; h, h_1, \dots, h_n)\| = o(\|h\|), \quad h \rightarrow 0. \quad (8)$$

Here we assume that

$$d^{n+1} f(u): X \times \dots \times X \rightarrow Y$$

is an $(n+1)$ -linear bounded map. Synonymously, we set

$$f^{(n+1)}(u) \equiv d^{n+1} f(u).$$

Here $f^{(n+1)}(u)$ is called the $(n+1)$ th *F-derivative* of f at the point u . Parallel to (5) and (6), we obtain that the existence of $d^{n+1} f(u)$ implies

$$d^{n+1} f(u)hh_1 \dots h_n = \frac{d}{dt} d^n f(u + th)h_1 \dots h_n|_{t=0} \quad (9)$$

for all $h_1, \dots, h_n \in X$, where $n = 1, 2, \dots$.

Definition 3. The differential $d^n f$ is said to be continuous at the point u iff for each $\varepsilon > 0$, there is a $\delta(\varepsilon) > 0$ such that

$$\|d^n f(u + h) - d^n f(u)\| \leq \varepsilon \quad \text{for all } h \in X \text{ with } \|h\| < \delta(\varepsilon),$$

where the norm is to be understood in the sense of Definition 1 from Section 9.1. Explicitly, this means that

$$\|d^n f(u + h)h_1 \dots h_n - d^n f(u)h_1 \dots h_n\| \leq \varepsilon \|h_1\| \cdots \|h_n\|$$

for all $h_1, \dots, h_n \in X$ and all $h \in X$ with $\|h\| < \delta(\varepsilon)$.

The map $f: U \subseteq X \rightarrow Y$ on the open subset U of X is called C^k ($k \geq 0$) iff $d^n f$ is continuous on U for $n = 0, 1, \dots, k$, where $d^0 f := f$.

Moreover, f is called C^∞ iff f is C^k for all k .

The relation between f'' and the *iterated derivative* $(f')'$ will be studied in Section 4.6. As we will explain there, with a view to concrete applications, it is easier to work with $f''(u)$, in the sense of Definition 1, than with $(f')'(u)$. Our definition of $f''(u)$ emphasizes the philosophy that higher derivatives correspond to *multilinearization*.

Classical Standard Example 4. Suppose that the real-valued function

$$f: U(u) \subseteq \mathbb{R}^N \rightarrow \mathbb{R}$$

of N real variables (ξ_1, \dots, ξ_N) has continuous partial derivatives up to the k th order on an open neighborhood $U(u)$ of the point u . Then the differential $d^n f(u)$ exists for all $n = 1, \dots, k$, where

$$d^k f(u)h_1 \dots h_k = \sum_{j_1, \dots, j_k=1}^N \partial_{j_1} \cdots \partial_{j_k} f(u)h_{1j_1}h_{2j_2} \cdots h_{kj_k} \quad (10)$$

for all $h_1, \dots, h_k \in \mathbb{R}^N$ with $h_j := (h_{j1}, \dots, h_{jN})$.

In addition, if f has continuous partial derivatives up to the k th order on the open subset U of \mathbb{R}^N , then f is C^k on U .

Formula (10) is identical to the well-known classic formula for differentials.

Proof. Let $k = 1$ and $N = 2$. We set $u := (\xi, \eta)$ and $h := (\alpha, \beta)$. The classic mean value theorem tells us that

$$\begin{aligned} & f(\xi + \alpha, \eta + \beta) - f(\xi, \eta) \\ &= f(\xi + \alpha, \eta + \beta) - f(\xi, \eta + \beta) + f(\xi, \eta + \beta) - f(\xi, \eta) \\ &= f_\xi(\xi + \vartheta\alpha, \eta + \beta)\alpha + f_\eta(\xi, \eta + \theta\beta)\beta, \quad \text{where } 0 < \vartheta, \theta < 1. \end{aligned}$$

Note that $|h| = (\|\alpha\|^2 + \|\beta\|^2)^{\frac{1}{2}}$; then the continuity of f_ξ and f_η at (ξ, η) implies that

$$f(u+h) - f(u) = f_\xi(u)\alpha + f_\eta(u)\beta + o(|h|), \quad h \rightarrow 0.$$

Hence

$$df(u)h = f_\xi(u)\alpha + f_\eta(u)\beta.$$

This is (10) for $k = 1$.

The remaining statements are proved similarly. \square

Example 5 (Differentiation of bilinear operators). Let

$$B: X_1 \times X_2 \rightarrow Y$$

be a bilinear bounded operator, where X_1 , X_2 , and Y are Banach spaces over \mathbb{K} . Set $X := X_1 \times X_2$ and $u := (u_1, u_2)$ for $u \in X$.

Then B is C^∞ . For all $u, h, k \in X$,

$$dB(u)h = B(u_1, u_2) + B(h_1, h_2), \quad (11)$$

$$d^2B(u)kh = B(k_1, h_2) + B(h_1, k_2), \quad (12)$$

and $d^nB(u) = 0$ if $n = 3, 4, \dots$

Proof. Let us use the strategy from Remark 2.

Ad (11). Formally,

$$dB(u)h = \frac{d}{dt}B(u + th)|_{t=0}.$$

Since

$$\begin{aligned} B(u + th) &= B(u_1 + th_1, u_2 + th_2) \\ &= B(u_1, u_2) + t[B(h_1, u_2) + B(u_1, h_2)] + t^2B(h_1, h_2), \end{aligned} \quad (13)$$

we get (11).

To justify this, we have to inspect the remainder. Note that

$$\|u\| = \|u_1\| + \|u_2\| \quad \text{for all } u \in X.$$

By (13) with $t = 1$,

$$B(u + h) = B(u, h) + dB(u)h + r,$$

where $r = B(h_1, h_2)$. Hence

$$\|r\| \leq \|B\| \|h_1\| \|h_2\| \leq \|B\| \|h\|^2 \quad \text{for all } h \in X,$$

that is, $r = o(\|h\|)$ as $h \rightarrow 0$.

It follows from (11) and the bilinearity of B that $dB(u): X \rightarrow Y$ is linear. Finally, we have to show that the operator $dB(u): X \rightarrow Y$ is bounded. But this follows from

$$\begin{aligned}\|dB(u)h\| &\leq \|B\| \|u_1\| \|h_2\| + \|B\| \|h_1\| \|u_2\| \\ &\leq 2\|B\| \|u\| \|h\| \quad \text{for all } h \in X.\end{aligned}$$

Ad (12). Relation (12) follows analogously to (11). Since $d^2B(u)$ does not depend on u , we also get $d^nB(u) = 0$ if $n \geq 3$.

By (11), the continuity of $u \mapsto dB(u)$ follows from

$$\begin{aligned}\|dB(u)h - dB(v)h\| &\leq \|B\| \|u_1 - v_1\| \|h_2\| + \|B\| \|h_1\| \|u_2 - v_2\| \\ &\leq 2\|B\| \|u - v\| \|h\|,\end{aligned}$$

and hence

$$\|dB(u) - dB(v)\| \leq 2\|B\| \|u - v\| \quad \text{for all } u, v \in X.$$

Similarly, we get the continuity of d^2B . □

Proposition 6 (The sum rule). *Let $f, g: U(u) \subseteq X \rightarrow Y$ be mappings on an open neighborhood of the point u , where X and Y are Banach spaces over \mathbb{K} . Let $n = 1, 2, \dots$. Then*

$$(\alpha f(u) + \beta g(u))^{(n)} = \alpha f^{(n)}(u) + \beta g^{(n)}(u) \quad \text{for all } \alpha, \beta \in \mathbb{K},$$

provided the F -derivatives $f^{(n)}(u)$ and $g^{(n)}(u)$ exist.

Proof. This follows immediately from the definition of the F -derivative. □

Partial F -derivatives are defined parallelly to the classical situation.

Definition 7. Let the map

$$f: U(u, v) \subseteq X \times Y \rightarrow Z$$

be given on an open neighborhood of the point (u, v) , where X , Y , and Z are Banach spaces over \mathbb{K} .

Let v be fixed and set $g(w) := f(w, v)$. If g has an F -derivative at the point u , then we define the *partial F -derivative* $f_u(u, v)$ through

$$f_u(u, v) := g'(u).$$

The partial F -derivative $f_v(u, v)$ is defined similarly. Instead of $f_u(u, v)$, $f_v(u, v)$, one also writes $D_1f(u, v)$, $D_2f(u, v)$, respectively.

Proposition 8. *Let $f: U(u, v) \subseteq X \times Y \rightarrow Z$ be given in Definition 7. If the F -derivative $f'(u, v)$ exists, then the partial derivatives $f_u(u, v)$, $f_v(u, v)$ also exist and*

$$f'(u, v)(h, k) = f_u(u, v)h + f_v(u, v)k \quad \text{for all } h \in X, k \in Y.$$

Proof. Note that $f_u(u, v)h = f'(u, v)(h, 0)$ and $f_v(u, v)k = f'(u)(0, k)$. \square

Further properties of partial F -derivatives will be proved in Problem 4.11.

4.3 Applications to Analytic Operators

Definition 1. Let X and Y be Banach spaces over \mathbb{K} . Let there be given a k -linear bounded operator

$$M: X \times X \times \cdots \times X \rightarrow Y,$$

which is *symmetric* in all variables. A *power operator* is created from M by setting

$$Mu^k := M(u, \dots, u) \quad (14a)$$

and

$$Mu^m v^n := M(\underbrace{u, \dots, u}_{m \text{ times}}; \underbrace{v, \dots, v}_{n \text{ times}}), \quad m + n = k, \quad (14b)$$

for any partition of k . For $k = 0$, Mu^0 is a fixed element w in X . Henceforth $\|M\| \|u\|^n$ with $n = 0$ will denote the norm $\|w\|$ of this element.

More precisely, in (14b) we need only that $(u_1, \dots, u_k) \mapsto M(u_1, \dots, u_k)$ be symmetric with respect to both (u_1, \dots, u_m) and (u_{m+1}, \dots, u_k) .

Example 2 (Integral operators). Let $X = Y = C[a, b]$, and let $\mathcal{A}: [a, b] \times [a, b] \rightarrow \mathbb{R}$ be continuous, where $-\infty < a < b < \infty$. Define

$$M(u, v, w)(y) := \int_a^b \mathcal{A}(y, x)u(x)v(x)w(x)dx \quad \text{for all } y \in [a, b]$$

and all $u, v, w \in X$. Then we obtain a *power operator* from $X \times X$ to X by setting

$$Muv^2 := M(u, v, v) \quad \text{for all } u, v \in X.$$

We have⁴

$$\|M(u, v, w)\| \leq (b - a) \max_{a \leq x, y \leq b} |\mathcal{A}(x, y)| \|u\| \|v\| \|w\| \quad \text{for all } u, v, w \in X,$$

and hence

$$\|M\| \leq (b - a) \max_{a \leq x, y \leq b} |\mathcal{A}(x, y)|.$$

Definition 3. Let $C^k[a, b]$ be the set of all continuous functions

$$u: [a, b] \rightarrow \mathbb{R}$$

⁴Recall that $\|u\| = \max_{a \leq x \leq b} |u(x)|$.

that have continuous derivatives up to order k on the compact interval $[a, b]$.

Then $C^k[a, b]$ is a real Banach space with the norm

$$\|u\| := \sum_{j=0}^k \max_{a \leq x \leq b} |u^{(j)}(x)|.$$

This is a special case of Example 6 in Section 4.4.

Example 4 (Differential operators). Let $X := C^1[a, b]$, $Y := C[a, b]$, and

$$M(u, v, w)(y) := \phi(y)u'(y)v'(y)w(y) \quad \text{for all } y \in [a, b],$$

all $u, v, w \in X$, and fixed function $\phi \in Y$. Then, letting

$$Mu^2v := \phi \cdot (u')^2v \quad \text{for all } u, v \in X,$$

we obtain a power operator from $X \times X$ to Y . We have

$$\|M(u, v, w)\|_Y \leq \|\phi\|_Y \|u\|_X \|v\|_X \|w\|_X \quad \text{for all } u, v, w \in X,$$

and hence $\|M\| \leq \|\phi\|_Y$.

Computation with power operators is the same as with ordinary powers. For all $u, v \in X$, $\alpha \in \mathbb{K}$, and $n, m, k \in \mathbb{N}$, we have

$$M(\alpha u)^k = \alpha^k Mu^k, \quad \|Mu^m v^n\| \leq \|M\| \|u\|^m \|v\|^n, \quad (15)$$

$$M(u + v)^k = \sum_{j=0}^k \binom{k}{j} Mu^{k-j}v^j \quad (\text{binomial formula}), \quad (16)$$

$$\|Mu^k - Mv^k\| \leq kR^{k-1}\|M\| \|u - v\| \quad \text{for } \|u\|, \|v\| \leq R. \quad (17)$$

The simple idea behind the proof of (17) can be seen by considering the decomposition

$$Mu^2 - Mv^2 = M(u, u) - M(v, v) = M(u - v, u) + M(v, u - v),$$

and so

$$\|Mu^2 - Mv^2\| \leq \|M\| \|u - v\| \|u\| + \|M\| \|v\| \|u - v\|.$$

Proposition 5. *Power operators have Lipschitz continuous F-derivatives of arbitrary order. The formulas for the F-derivatives are parallel to the corresponding classic formulas.*

To explain this, let the power operator $A: X \rightarrow Y$ be given, where

$$A(u) := Mu^k.$$

Then, for each $u \in X$,

$$A'(u)h = kMu^{k-1}h, \quad A''(u)h_1h_2 = k(k-1)Mu^{k-2}h_1h_2, \quad (18)$$

for all $h, h_1, h_2 \in X$, and so on.

The proof of (18) follows from (16). Let us explain the simple idea of the proof with the following special case.

Example 6. Let $Au := Mu^2$. Then, for all $u, v, h, h_1, h_2 \in X$,

$$A'(u)h = 2Muh, \quad (19)$$

$$\|A'(u) - A'(v)\| \leq 2\|M\| \|u - v\|, \quad (20)$$

$$A''(u)h_1h_2 = 2Mh_1h_2, \quad (21)$$

and $A^{(n)}(u) \equiv 0$ if $n \geq 3$.

Proof. From

$$\begin{aligned} A(u+h) - Au &= M(u+h, u+h) - M(u, u) \\ &= 2M(u, h) + M(h, h) = 2Muh + r(h), \end{aligned}$$

along with $\|r(h)\| \leq \|M\| \|h\|^2$, we get (19).

Relation (20) follows from

$$\|(A'(u) - A'(v))h\| = \|2M(u-v)h\| \leq 2\|M\| \|u-v\| \|h\|,$$

that is, $A': X \rightarrow Y$ is Lipschitz continuous. Furthermore,

$$A'(u+h_1)h_2 - A'(u)h_2 = 2Mh_1h_2.$$

This yields (21). □

The following definition is basic. Let us consider expressions of the form

$$A(u) := \sum_{k=0}^{\infty} M_k(u - u_0)^k, \quad u \in X, \quad (22)$$

together with the *majorant condition*

$$\sum_{k=0}^{\infty} \|M_k\| \|u - u_0\|^k < \infty. \quad (23)$$

Convergence of (23) ensures the absolute convergence of (22), and hence the convergence of (22), by Section 1.22 in AMS Vol. 108.

Definition 7. Let X and Y be Banach spaces over \mathbb{K} , and let $M_k: X \rightarrow Y$ be power operators for all $k = 0, 1, \dots$.

The operator $A: U(u_0) \subseteq X \rightarrow Y$ is called *analytic* at the point $u_0 \in X$ iff there is a number $\rho > 0$ such that series (23) converges for all $u \in X$ with $\|u - u_0\| \leq \rho$ and $A(u)$ allows representation (22) for all those points u .

The operator A is called analytic on the open set V iff A is analytic at each point of V .

Theorem 4.A. *If the operator A is analytic at the point u_0 , then A is C^∞ on some open neighborhood $V(u_0)$ of u_0 .*

Moreover, the formulas for the derivatives are obtained by an application of the corresponding formulas for power operators.

For example, from (22) and (23) we obtain

$$A'(u)h = \sum_{k=1}^{\infty} k M_k(u - u_0)^{k-1} h \quad \text{for all } h \in X \text{ and } u \in V(u_0).$$

Proof. By (23), the series

$$\sum_{k=0}^{\infty} \|M_k\| z^k$$

converges for all $z \in \mathbb{C}$ with $|z| \leq \rho$, and hence the differentiated series

$$\sum_{k=1}^{\infty} k \|M_k\| z^{k-1}$$

converges for all $z \in \mathbb{C}$ with $|z| \leq \frac{\rho}{2}$.

Step 1: We want to prove that the F -derivative $A'(u)$ exists for all $u \in X$ with $\|u - u_0\| \leq \frac{\rho}{2}$.

Let $\|h\| \leq \frac{\rho}{2}$ and $\|u - u_0\| \leq \frac{\rho}{2}$, and let $k = 2, 3, \dots$. By (16),

$$M_k(u + h - u_0)^k - M_k(u - u_0)^k = k M_k(u - u_0)^{k-1} h + r_k,$$

where

$$r_k := \sum_{j=2}^k \binom{k}{j} M_k(u - u_0)^{k-j} h^j.$$

Since $\sum_{j=0}^k \binom{k}{j} = 2^k$,

$$\|r_k\| \leq \sum_{j=2}^k \binom{k}{j} \|M_k\| \|u - u_0\|^{k-j} \|h\|^j \leq 4\|h\|^2 \|M_k\| \rho^{k-2}.$$

Hence

$$A(u + h) - A(u) = Lh + r,$$

where

$$Lh := \sum_{k=1}^{\infty} kM_k(u - u_0)^{k-1}h \quad \text{and} \quad r := \sum_{k=2}^{\infty} r_k.$$

This implies

$$\|Lh\| \leq \left(\sum_{k=1}^{\infty} k\|M_k\| \left(\frac{\rho}{2}\right)^{k-1} \right) \|h\|$$

and

$$\|r\| \leq \frac{4}{\rho^2} \left(\sum_{k=2}^{\infty} \|M_k\| \rho^k \right) \|h\|^2.$$

Therefore, $A'(u) = L$.

Step 3: The existence of higher derivatives can be proved analogously. \square

The explicit computation of higher F -derivatives can frequently be based on the following *special product rule*:

$$\frac{d}{dt} B(\phi(t), \psi(t)) \Big|_{t=s} = B(\phi'(s), \psi(s)) + B(\phi(s), \psi'(s)). \quad (24)$$

Proposition 8. Let $B: X \times Y \rightarrow Z$ be a bilinear bounded operator, where X , Y , and Z are Banach spaces over \mathbb{K} . Furthermore, let

$$\phi: U(s) \subseteq \mathbb{R} \rightarrow X \quad \text{and} \quad \psi: U(s) \subseteq \mathbb{R} \rightarrow Y$$

be functions defined on an open neighborhood of $s \in \mathbb{R}$ such that the derivatives $\phi'(s)$ and $\psi'(s)$ exist.

Then the derivative of the function $t \mapsto B(\phi(t), \psi(t))$ exists at the point s and formula (24) holds true.

Proof. We have

$$\phi(s+h) = \phi(s) + h\phi'(s) + h\alpha(h) \quad \text{and} \quad \psi(s+h) = \psi(s) + h\psi'(s) + h\beta(h),$$

where $\alpha(h) \rightarrow 0$ and $\beta(h) \rightarrow 0$ as $h \rightarrow 0$. Hence

$$B(\phi(s+h), \psi(s+h)) = B(\phi(s), \psi(s)) + hb + h\gamma(h),$$

where b denotes the right-hand side of (24) and $\gamma(h) \rightarrow 0$ as $h \rightarrow 0$, by the continuity of B . \square

Standard Example 9. Let X and Y be Banach spaces over \mathbb{K} with $X \neq \{0\}$ and $Y \neq \{0\}$. By Section 1.23 in AMS Vol. 108, there exists a maximal nonempty subset

$L_{\text{inv}}(X, Y)$ of $L(X, Y)$ such that $A^{-1} \in L(Y, X)$ for all $A \in L_{\text{inv}}(X, Y)$.

Define the operator $\Phi: L_{\text{inv}}(X, Y) \rightarrow L(Y, X)$ by

$$\Phi(A) := A^{-1}.$$

Then Φ is analytic and

$$\Phi'(A)B = -\Phi(A)B\Phi(A) \quad \text{for all } A \in L_{\text{inv}}(X, Y), B \in L(X, Y). \quad (25)$$

Proof. Set $H := -A^{-1}B$. The Neumann series from Section 1.23 in AMS Vol. 108 yields

$$\Phi(A+B) = (A(I-H))^{-1} = (I-H)^{-1}A^{-1} = (I+H+H^2+\cdots)A^{-1}. \quad (26)$$

For $\|H\| < 1$, this series has the geometric series

$$(1 + \|H\| + \|H\|^2 + \cdots) \|A^{-1}\|$$

as a majorant series, and Φ is hence analytic.

By Theorem 4.A, the operator Φ has F -derivatives of each order. According to (9) and (26),

$$\Phi'(A)B = \frac{d}{dt}\Phi(A+tB)\Big|_{t=0} = HA^{-1}.$$

This is (25). □

Using (9) and the special product rule (24), all the higher derivatives of Φ are obtained recursively. For example, to compute Φ'' , note that

$$(E, F) \mapsto -EBF$$

is a bilinear bounded operator from $L(Y, X) \times L(Y, X)$ to $L(X, Y)$, since $\|EBF\| \leq \|B\| \|E\| \|F\|$. By (9) and (25),

$$\Phi''(A)CB = \frac{d}{dt}\Phi'(A+tC)B\Big|_{t=0} = -\Phi'(A)CB\Phi(A) - \Phi(A)B\Phi'(A)C,$$

for all $A \in L_{\text{inv}}(X, Y)$ and $B, C \in L(X, Y)$.

4.4 Integration

Throughout this section let $-\infty < a < b < \infty$, and let X be a Banach space over \mathbb{K} with the norm $\|\cdot\|$. Furthermore, set

$$\|u\|_* := \sup_{a \leq t \leq b} \|u(t)\|.$$

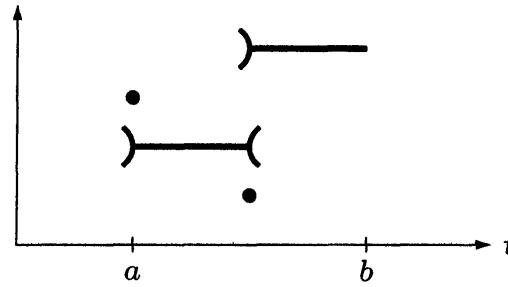


FIGURE 4.2.

Parallel to classical analysis, we will define the *integral* through

$$\int_a^b u(t)dt := \lim_{n \rightarrow \infty} \int_a^b u_n(t)dt, \quad (27)$$

where (u_n) is a sequence of regular step functions with

$$\|u - u_n\|_* \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (28)$$

Definition 1. Let the function $u: [a, b] \rightarrow X$ be given.

(a) u is called a regular⁵ *step function* iff there exists a decomposition $a = t_0 < t_1 < \dots < t_m = b$ of the interval $[a, b]$ such that the function $u = u(t)$ is *constant* on all the open subintervals $]t_{j-1}, t_j[$, that is,

$$u(t) = b_j \quad \text{for all } t \in]t_{j-1}, t_j[, \quad j = 1, \dots, m,$$

where $b_j \in X$ (see Figure 4.2). The integral of such a regular step function u is defined by

$$\int_a^b u(t)dt := \sum_{j=1}^m (t_j - t_{j-1})b_j.$$

Obviously, by the triangle inequality,

$$\left\| \int_a^b u(t)dt \right\| \leq (b - a)\|u\|_*. \quad (29)$$

(b) u is called *integrable* iff there exists a sequence (u_n) of regular step functions $u_n: [a, b] \rightarrow X$ such that (28) holds. Then the *integral* $\int_a^b u(t)dt$ is defined by (27).

The following proposition shows that definition (b) makes sense.

Proposition 2. (i) *The limit in (27) exists.*

⁵More general step functions are considered in the appendix to AMS Vol. 108 in order to define the more general Lebesgue integral.

(ii) *This limit is independent of the choice of (u_n) .*

Proof. Ad (i). If v and w are regular step functions on $[a, b]$, then so is the linear combination $\alpha v + \beta w$ for $\alpha, \beta \in \mathbb{K}$. This follows by using a refinement of the corresponding partitions of $[a, b]$.

Let

$$\|u - u_n\|_* \rightarrow 0 \quad \text{and} \quad \|u - v_n\|_* \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

where (u_n) and (v_n) are sequences of regular step functions on $[a, b]$. Then, for each $\varepsilon > 0$, there is an $n_0(\varepsilon)$ such that

$$\begin{aligned} \|u_n - u_m\|_* &\leq \|(u_n - u) + (u - u_m)\|_* \\ &\leq \|u_n - u\|_* + \|u - u_m\|_* < \varepsilon \quad \text{for all } n, m \geq n_0(\varepsilon), \end{aligned} \tag{30}$$

that is, (u_n) is *Cauchy* with respect to $\|\cdot\|_*$. Furthermore,

$$\|u_n - v_n\|_* \leq \|u_n - u\|_* + \|u - v_n\|_* \rightarrow 0 \quad \text{as } n \rightarrow \infty. \tag{31}$$

It follows from (29) and (30) that

$$\begin{aligned} \left\| \int_a^b u_n(t) dt - \int_a^b u_m(t) dt \right\| &= \left\| \int_a^b (u_n(t) - u_m(t)) dt \right\| \\ &\leq (b-a) \|u_n - u_m\|_* < (b-a)\varepsilon \end{aligned}$$

for all $n, m \geq n_0(\varepsilon)$. Thus, the sequence $(\int_a^b u_n(t) dt)$ is Cauchy in the Banach space X . Hence the limit in (27) exists.

Ad (ii). By (29) and (31),

$$\left\| \int_a^b u_n(t) dt - \int_a^b v_n(t) dt \right\| \leq (b-a) \|u_n - v_n\|_* \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad \square$$

Standard Example 3. Each continuous function $u: [a, b] \rightarrow X$ is integrable.

Proof. Let $n = 1, 2, \dots$. For $j = 1, \dots, n$, define $t_j := a + j\Delta t$, where $\Delta t := \frac{b-a}{n}$. Set

$$u_n(t) := u(t_j) \quad \text{for all } t \in [t_{j-1}, t_j[,$$

along with $u_n(b) := u(b)$. Then

$$\|u - u_n\|_* \leq \max_{1 \leq j \leq n} \sup_{t_{j-1} \leq t \leq t_j} \|u(t) - u(t_j)\| \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

since u is *uniformly* continuous on $[a, b]$ (cf. Proposition 9 in Section 1.11 of AMS Vol. 108). \square

Obviously, if $a < c < b$ and $u: [a, b] \rightarrow X$ is integrable over both subintervals $[a, c]$ and $[c, b]$, then u is also integrable over $[a, b]$ and

$$\int_a^b u(t)dt = \int_a^c u(t)dt + \int_c^b u(t)dt. \quad (32)$$

Proposition 4 (Properties of the integral). *Let $u, v: [a, b] \rightarrow X$ be integrable. Then the following properties exist:*

- (i) Linearity. *For all $\alpha, \beta \in \mathbb{K}$, the function $\alpha u + \beta v$ is integrable over $[a, b]$ and*

$$\int_a^b (\alpha u(t) + \beta v(t))dt = \alpha \int_a^b u(t)dt + \beta \int_a^b v(t)dt.$$

- (ii) Generalized triangle inequality. *The function $t \mapsto \|u(t)\|$ is integrable over $[a, b]$ and*

$$\left\| \int_a^b u(t)dt \right\| \leq \int_a^b \|u(t)\|dt. \quad (33)$$

- (iii) Functionals. *For each functional $u^* \in X^*$, the function $t \mapsto \langle u^*, u(t) \rangle$ is integrable over $[a, b]$ and*

$$\int_a^b \langle u^*, u(t) \rangle dt = \langle u^*, \int_a^b u(t)dt \rangle. \quad (34)$$

Proof. Ad (i), (iii). Use definition (27) and the continuity of u^* .

Ad (ii). By (27) and the triangle inequality,

$$\begin{aligned} \left\| \int_a^b u(t)dt \right\| &= \lim_{n \rightarrow \infty} \left\| \int_a^b u_n(t)dt \right\| \\ &\leq \lim_{n \rightarrow \infty} \int_a^b \|u_n(t)\|dt = \int_a^b \|u(t)\|dt. \end{aligned}$$

Note that if $t \mapsto u_n(t)$ is a regular step function on $[a, b]$, then so is $t \mapsto \|u_n(t)\|$. \square

Theorem 4.B (The fundamental theorem of calculus). *Let $u: [a, b] \rightarrow X$ be continuous. Then the function $v: [a, b] \rightarrow X$ defined by*

$$v(s) := \int_a^s u(t)dt, \quad a \leq s \leq b,$$

is differentiable on $[a, b]$ with⁶

$$v'(s) = u(s) \quad \text{for all } s \in [a, b].$$

Proof. Suppose that $h > 0$ and $s < b$. By (33),

$$\begin{aligned} \|h^{-1}(v(s+h) - v(s)) - u(s)\| &= \left\| h^{-1} \int_s^{s+h} (u(t) - u(s)) dt \right\| \\ &\leq h^{-1} \int_s^{s+h} \|u(t) - u(s)\| dt \\ &\leq \sup_{s \leq t \leq s+h} \|u(t) - u(s)\| \rightarrow 0 \quad \text{as } h \rightarrow 0. \end{aligned}$$

The proof for $h < 0$ proceeds similarly. \square

Corollary 5. If the function $u: [a, b] \rightarrow X$ is continuously differentiable, then

$$\int_a^s u'(t) dt = u(s) - u(a) \quad \text{for all } s \in [a, b].$$

Proof. Set $v(s) := \int_a^s u'(t) dt - u(s) + u(a)$. By Theorem 4.B,

$$v'(s) = 0 \quad \text{on } [a, b] \quad \text{and} \quad v(a) = 0.$$

Thus, for all $u^* \in X^*$,

$$\frac{d}{ds} \langle u^*, v(s) \rangle = \langle u^*, v'(s) \rangle = 0 \quad \text{on } [a, b]$$

and $\langle u^*, v(a) \rangle = 0$. By classic calculus, this implies

$$\langle u^*, v(s) \rangle = 0 \quad \text{on } [a, b] \quad \text{for all } u^* \in X^*.$$

Hence $v(s) = 0$ on $[a, b]$. \square

Example 6. Let $-\infty < a < b < \infty$, and let X be a Banach space over \mathbb{K} . For $k = 0, 1, \dots$, let

$$C^k([a, b], X)$$

denote the set of all continuous functions $u: [a, b] \rightarrow X$ that have continuous derivatives up to order k . Then $C^k([a, b], X)$ becomes a Banach space over \mathbb{K} equipped with the norm

$$\|u\|_k := \sum_{j=0}^k \max_{a \leq t \leq b} \|u^{(j)}(t)\|.$$

⁶The derivatives $v'(a)$ and $v'(b)$ are to be understood as one-sided derivatives.

Instead of $C^0([a, b], X)$, we simply write $C([a, b], X)$.

Proof. Obviously, $\|\cdot\|_k$ is a norm.

Step 1: First let $k = 0$. Then we can use the same proof as in Standard Example 6 from Section 1.3 in AMS Vol. 108.

Step 2: Let $k = 1$. Suppose that (u_n) is a Cauchy sequence with respect to the norm $\|\cdot\|_1$, that is, for each $\varepsilon > 0$, there is an $n_0(\varepsilon)$ such that

$$\|u_n - u_m\|_1 = \max_{a \leq t \leq b} \|u_n(t) - u_m(t)\| + \max_{a \leq t \leq b} \|u'_n(t) - u'_m(t)\| < \varepsilon$$

for all $n, m \geq n_0(\varepsilon)$. By Step 1, $C([a, b], X)$ is a Banach space. Thus, there exist functions $u, v \in C([a, b], X)$ such that

$$\|u_n - u\|_0 \rightarrow 0 \quad \text{and} \quad \|u'_n - v\|_0 \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Let $s \in [a, b]$. Then

$$\left\| \int_a^s (u'_n(t) - v(t)) dt \right\| \leq (b-a) \|u'_n - v\|_0 \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

and hence

$$u(s) - u(a) = \int_a^s v(t) dt.$$

By Theorem 4.B, $u'(s) = v(s)$. This implies $u \in C^1([a, b], X)$ and the convergence $\|u_n - u\|_1 \rightarrow 0$ as $n \rightarrow \infty$.

For $k \geq 2$, proceed by induction. \square

4.5 Applications to the Taylor Theorem

In classic analysis, the fundamental Taylor formula describes the approximation of functions by polynomials. The *generalized Taylor formula* on Banach spaces reads as follows:

$$f(u+h) = f(u) + \sum_{k=1}^{n-1} \frac{1}{k!} f^{(k)}(u) h^k + R_n, \quad (35)$$

where we set $f^{(k)}(u)h^k := f^{(k)}(u)h\dots h$, and the remainder has the form

$$R_n := \int_0^1 \frac{(1-\tau)^{n-1}}{(n-1)!} f^{(n)}(u + \tau h) h^n d\tau, \quad (36)$$

where $n = 1, 2, \dots$. For $n = 1$, the sum will be zero in (35), by definition.

Theorem 4.C. *Let the map $f: U \subseteq X \rightarrow Y$ be C^n on the open convex set U , where X and Y are Banach spaces over \mathbb{K} . Then the Taylor formula (35), (36) holds true for all $u \in U$ and all $h \in X$ with $u + h \in U$.*

In particular, since $\int_0^1 (1 - \tau)^{n-1} d\tau = \frac{1}{n}$, it follows from (36) that

$$\|R_n\| \leq \frac{1}{n!} \sup_{0 \leq \tau \leq 1} \|f^{(n)}(u + \tau h)h^n\|. \quad (36^*)$$

Proof. For given $v^* \in Y^*$, set

$$\phi(t) := \langle v^*, f(u + th) \rangle, \quad 0 \leq t \leq 1.$$

By (9), for $k = 1, \dots, n$,

$$\phi^{(k)}(t) = \langle v^*, f^{(k)}(u + th)h^k \rangle, \quad 0 \leq t \leq 1.$$

The classic Taylor theorem for real functions tells us that

$$\phi(1) - \phi(0) - \sum_{k=1}^{n-1} \frac{1}{k!} \phi^{(k)}(0) - \int_0^1 \frac{(1-\tau)^{n-1}}{(n-1)!} \phi^{(n)}(\tau) d\tau = 0.$$

Using (34), this means

$$\langle v^*, f(u + h) - f(u) - \sum_{k=1}^{n-1} \frac{1}{k!} f^{(k)}(u)h^k - R_n \rangle = 0 \quad \text{for all } v^* \in Y^*.$$

Hence we obtain (35). □

4.6 Iterated Derivatives

Proposition 1. Let $f: U(u) \subseteq X \rightarrow Y$ be defined on an open neighborhood of the point u , where X and Y are Banach spaces over \mathbb{K} .

- (i) The second F -derivative $f''(u)$ exists iff the iterated derivative $(f')'(u)$ exists. In this case, we get

$$f''(u)hk = (f')'(u)(h)(k) \quad \text{for all } h, k \in X. \quad (37)$$

- (ii) f'' is continuous at the point u iff $(f')'$ is continuous at u .

Proof. Ad (i). Step 1: Suppose that $(f')'(u)$ exists. We want to prove that $f''(u) \equiv d^2 f(u)$ exists along with (37).

The existence of $(f')'(u)$ means the following. The operator

$$f'(v): X \rightarrow Y$$

is linear and bounded for all $v \in V$, where V is some open neighborhood of the point u . This way, we get the operator

$$f': V \subseteq X \rightarrow L(X, Y).$$

Then the operator

$$(f')'(u): X \rightarrow L(X, Y)$$

is linear and bounded. Hence

$$(f')'(u)(h) \in L(X, Y) \quad \text{for all } h \in X,$$

which implies

$$(f')'(u)(h)(k) \in Y \quad \text{for all } k \in X.$$

To simplify notation, set $g(v) := f'(v)$. For all $h, k \in L(X, Y)$,

$$\|g'(u)(h)(k)\| \leq \|g'(u)h\| \|k\| \leq \|g'(u)\| \|h\| \|k\|,$$

meaning that the map $(h, k) \mapsto g'(u)(h)(k)$ is bilinear and bounded from $X \times X$ to Y . It remains to prove (37). By definition of $g'(u)$,

$$g(u + h) - g(u) = g'(u)h + R(h),$$

where $R(h) \in L(X, Y)$ and $R(h) = o(\|h\|)$ as $h \rightarrow 0$. This implies

$$f'(u + h)k - f'(u)k = g'(u)(h)k + r,$$

where $r(h, k) := R(h)k$. Hence $\|r(h, k)\| \leq \|R(h)\| \|k\|$, that is,

$$\sup_{\|k\| \leq 1} \|r(h, k)\| = o(\|h\|), \quad h \rightarrow 0.$$

This yields (37).

Step 2: Suppose that $f''(u)$ exists. We have to show that $(f')'(u)$ exists along with (37). This follows similarly to Step 1.

Ad (ii). The statement follows immediately from the formula

$$\|(f')'(u + h) - (f')'(u)\| = \|f''(u + h) - f''(u)\|, \quad (38)$$

which follows from (37) along with

$$\begin{aligned} \|(f')'(u + h) - (f')'(u)\| &= \sup_{\|h_1\| \leq 1} \|g'(u + h)h_1 - g'(u)h_1\| \\ &= \sup_{\|h_1\| \leq 1, \|h_2\| \leq 1} \|g'(u + h)(h_1)(h_2) - g'(u)(h_1)(h_2)\|. \end{aligned} \quad \square$$

Similarly, we get the following more general result. Let us write

$$Df(u) := f'(u), \quad D^2f(u) := D(Df)(u), \text{ and so forth.}$$

Corollary 2. *Let the map $f: U \subseteq X \rightarrow Y$ be given as in Proposition 1, and let $n = 2, 3, \dots$*

- (i) The n th F -derivative $f^{(n)}(u)$ exists iff the n th iterated derivative $D^n f(u)$ exists. In this case, we have

$$D^n f(u)(h_1 \dots h_n) = f^{(n)}(u)h_1 \dots h_n \text{ for all } h_j \in X, j = 1, \dots, n. \quad (39)$$

- (ii) If $f^{(n)}(v)$ and $f^{(n)}(w)$ exist, then

$$\|f^{(n)}(v) - f^{(n)}(w)\| = \|D^n f(v) - D^n f(w)\|.$$

- (iii) $f^{(n)}$ is continuous at the point u iff $D^n f$ is continuous at u .

Remark 3 (Different approaches to the F -derivative). Let $\text{dom } f$ denote the domain of definition of f . Note that the iterated derivatives

$$\begin{aligned} f: \text{dom } f &\subseteq X \rightarrow Y, \\ f': \text{dom } f' &\subseteq X \rightarrow L(X, Y), \\ (f')': \text{dom } (f')' &\subseteq X \rightarrow L(X, L(X, Y)), \text{ and so forth,} \end{aligned}$$

correspond to image spaces that acquire a more and more *complicated structure*. Our definition of higher derivatives in Section 4.2 avoids this.

In mathematical literature, higher derivatives are frequently defined as iterated derivatives. Corollary 2 tells us that there exists a natural relation between the two different approaches, which allows us to identify $f^{(n)}$ with $D^n f$. Roughly speaking, one observes the following:

- (a) Our definition of $f^{(n)}$ in Section 4.2 is convenient with a view to the *computation* of derivatives in *concrete situations* (e.g., see Problems 4.1 through 4.7).
- (b) The use of iterated derivatives $D^n f$ simplifies *theoretical* investigations (e.g., see the chain rule in the next section).

Let $f = (f_1, \dots, f_N)$ and $N, n = 1, 2, \dots$. Then the following two formulas are frequently used in connection with the *chain rule*:

$$D^n f(u) = (D^n f_1(u), \dots, D^n f_N(u)) \quad (40)$$

and

$$f^{(n)}(u)h_1 \dots h_n = (f_1^{(n)}(u)h_1 \dots h_n, \dots, f_N^{(n)}(u)h_1 \dots h_n) \quad (41)$$

for all $h_j \in X_j, j = 1, \dots, N$.

Proposition 4. Let X, X_1, \dots, X_N be Banach spaces over \mathbb{K} , and let the operator

$$f: U \subseteq X \rightarrow X_1 \times \dots \times X_N$$

be given on the open neighborhood U of u . Then the following are true:

- (i) The iterated F -derivative $D^n f(u)$ exists iff $D^n f_j(u)$ exists for all $j = 1, \dots, N$. Here formula (40) holds true.
- (ii) The n th F -derivative $f^{(n)}(u)$ exists iff $f_j^{(n)}(u)$ exists for all $j = 1, \dots, N$. Here formula (41) holds true.
- (iii) f is C^n on U iff f_1, \dots, f_N are C^n on U .

Proof. This follows immediately from the definition of the norm

$$\|v\| = \|v_1\| + \cdots + \|v_N\|$$

on $X_1 \times \cdots \times X_N$ and from the corresponding definitions of $D^n f(u)$ and $f^{(n)}(u)$. \square

4.7 The Chain Rule

The chain rule represents the *most important* rule of differential calculus. The key formula reads as follows:

$$(g \circ f)'(u) = g'(f(u)) \circ f'(u). \quad (42)$$

Recall that $(g \circ f)(u) := g(f(u))$. Formula (42) tells us that the operations of linearization and composition can be interchanged. Since $g'(f(u))$ and $f'(u)$ are linear operators, formula (42) can also be written as

$$(g \circ f)'(u) = g'(f(u))f'(u). \quad (42^*)$$

This shorter notation is convenient in order to avoid clumsy formulas for higher derivatives, as we will explain ahead.

Theorem 4.D (The chain rule). *Let X , Y , and Z be Banach spaces over \mathbb{K} , and let the two mappings*

$$f: U(u) \subseteq X \rightarrow Y \quad \text{and} \quad g: V(f(u)) \subseteq Y \rightarrow Z$$

be given, where U and V are open neighborhoods of the points u and $f(u)$, respectively. Let $m = 1, 2, \dots$ be fixed.

If the F -derivatives $f^{(m)}(u)$ and $g^{(m)}(f(u))$ exist, then the F -derivative $(g \circ f)^{(m)}(u)$ exists and formula (42) holds true for $m = 1$.

Corollary 1. *For fixed $m = 1, 2, \dots$, suppose that the mappings*

$$f: U \subseteq X \rightarrow Y \quad \text{and} \quad g: V \subseteq Y \rightarrow Z$$

are C^m , where U and V are open sets and $f(U) \subseteq V$.

Then the composite map $g \circ f$ is also C^m on U .

Proof of Theorem 4.D. *Step 1:* Let $m = 1$. Set $v := g(u)$. By hypothesis,

$$\begin{aligned} f(u + h) - f(u) &= f'(u)h + \|h\| a(h), \\ g(v + k) - g(v) &= g'(v)k + \|k\| b(k), \end{aligned}$$

where $a(h) \rightarrow 0$ as $h \rightarrow 0$ and $b(k) \rightarrow 0$ as $k \rightarrow 0$. Choose $v := f(u)$ and $k := f(u + h) - f(u)$. Then

$$g(f(u + h)) - g(f(u)) = g'(v)f'(u)h + r(h), \quad (43)$$

where

$$r(h) := g'(v)\|h\|a(h) + \|k\|b(k) = o(\|h\|), \quad h \rightarrow 0,$$

since $\frac{\|r(h)\|}{\|h\|} \leq \|g'(v)\| \|a(h)\| + \dots \rightarrow 0$ as $h \rightarrow 0$. From (43) we get (42).

Step 2: In order to proceed by induction, we assume that the statement has been proved for $m = n$.

Suppose that $f^{(n+1)}(u)$ and $g^{(n+1)}(v)$ exist. By Section 4.6, this is equivalent to the existence of the iterated derivatives $D^{n+1}f(u)$ and $D^{n+1}g(v)$. By (42),

$$D(g \circ f)(u) = Dg(v)Df(u). \quad (44)$$

Define the bilinear bounded map $B: L(Y, Z) \times L(X, Y) \rightarrow L(X, Z)$ by

$$B(R, S) := RS \quad \text{for all } R \in L(Y, Z), S \in L(X, Y).$$

In this connection, note that $\|RS\| \leq \|R\| \|S\|$. Then equation (42) can be written in the form

$$D(g \circ f)(u) = B(Dg(v), Df(u)). \quad (45)$$

Example 5 in Section 4.2 tells us that B is C^∞ . By assumption, there exist $D^n(Dg)(v)$ and $D^n(Df)(u)$. Applying the chain rule for $m = n$ to (45) along with (40), we obtain the existence of $D^n D(g \circ f)(u)$. By Section 4.6, this implies the existence of $(g \circ f)^{(n+1)}(u)$. \square

Proof of Corollary 1. Observe that the continuity of Dg and Df implies the continuity of $D(g \circ f)$, by means of (45) along with the continuity of B . Now use the same induction argument as in the proof of Theorem 4.D. \square

Let us compute the F -derivative $(g \circ f)^{(n)}$ explicitly. Since Theorem 4.D ensures the existence of the derivatives, we can use the convenient formula (9) along with the special product rule (24). For example, let $n = 2$. Differentiating

$$(g \circ f)'(u + tk)h = g'(f(u + tk))f'(u + tk)h$$

at the point $t = 0$, we obtain

$$(g \circ f)''(u)kh = g''(v)f'(u)kf'(u)h + g'(v)f''(u)kh \quad (46)$$

for all $h, k \in X$, where $v := f(u)$. In this connection, note that we may write

$$g'(f(u + tk))f'(u + tk)h = C(g'(f(u + tk)), f'(u + tk)h),$$

where the bilinear bounded operator $C: L(Y, Z) \times Y \rightarrow Z$ is defined by

$$C(R, y) := Ry \quad \text{for all } R \in L(Y, Z), y \in Y.$$

In fact, this bilinear operator is bounded, since $\|Ry\| \leq \|R\| \|y\|$.

Analogously, we get $(g \circ f)^{(n)}(u)$ for $n = 3, 4, \dots$

Remark 2 (Convenient notation). Recall that $f'(u): X \rightarrow Y$ and $g'(v): Y \rightarrow Z$ are linear bounded operators, whereas $f''(u): X \times X \rightarrow Y$ and $g''(v): Y \times Y \rightarrow Z$ are bilinear bounded operators. Recall also our convention $Mvw := M(v, w)$ for bilinear operators. Without this convention, formula (46) reads as follows:

$$(g \circ f)''(u)(k, h) = g''(v)(f'(u)k, f'(u)h) + g'(v)[f''(u)(k, h)].$$

In contrast to this, our notation in (46) avoids redundant symbols.

Standard Example 3 (The product rule). Set

$$H(u) := B(f(u), g(u)).$$

We want to justify the formula

$$H'(u)h = B(f'(u)h, g(u)) + B(f(u), g'(u)h) \quad \text{for all } h \in X. \quad (47)$$

Let X, Y, Z , and W be Banach spaces over \mathbb{K} . Suppose that $B: Y \times Z \rightarrow W$ is a bilinear bounded operator, and suppose that the operators

$$f: U \subseteq X \rightarrow Y \quad \text{and} \quad g: U \subseteq X \rightarrow Z$$

are defined on the open neighborhood U of u and that the n th *F-derivatives* $f^{(n)}(u)$ and $g^{(n)}(u)$ exist for fixed $n = 1, \dots$. Then the following are true:

(i) The n th *F-derivative* $H^{(n)}(u)$ exists.

(ii) Formula (47) holds true.

(iii) If f and g are C^n on U , then so is H .

Proof. Ad (i), (iii). By Example 5 in Section 4.2, B is C^∞ . Now use the chain rule along with Proposition 4 in Section 4.6 on the differentiability of the map $u \mapsto (f(u), g(u))$.

Ad (ii). By (9) and the special product rule (24), differentiation of

$$H(u + th) = B(f(u + th), g(u + th))$$

at the point $t = 0$ yields (47). \square

The explicit expression for higher derivatives can also be obtained by using (9) and the special product rule (24). For example, differentiation of

$$H'(u + tk)h = B(f'(u + tk)h, g(u + tk)) + B(f(u + tk), g'(u + tk)h)$$

at the point $t = 0$ yields

$$\begin{aligned} H''(u)kh &= B(f''(u)kh, g(u)) + B(f'(u)h, g'(u)k) \\ &\quad + B(f'(u)k, g'(u)h) + B(f(u), g''(u)kh). \end{aligned}$$

4.8 The Implicit Function Theorem

We want to solve the operator equation

$$F(u, v) = 0 \tag{48}$$

in a neighborhood of the point (u_0, v_0) , where we assume that

$$F(u_0, v_0) = 0. \tag{49}$$

In particular, we are interested in a *locally unique* solution (cf. Figure 4.3). Condition (50) is decisive. Set $U := \{u \in X : \|u - u_0\| < \rho\}$.

Theorem 4.E. *Let X , Y , and Z be Banach spaces over \mathbb{K} , and let*

$$F: U(u_0, v_0) \subseteq X \times Y \rightarrow Z$$

be a C^n -map on an open neighborhood of the point (u_0, v_0) such that (49) holds and $1 \leq n \leq \infty$. Suppose that the operator

$$F_v(u_0, v_0) : Y \rightarrow Z \text{ is bijective.} \tag{50}$$

Then the following statements hold true:

- (i) *There exist numbers $r > 0$ and $\rho > 0$ such that, for each given $u \in U$, the original equation (48) has a unique solution $v \in Y$ with $\|v - v_0\| \leq r$. Denote this solution by $v(u)$.*

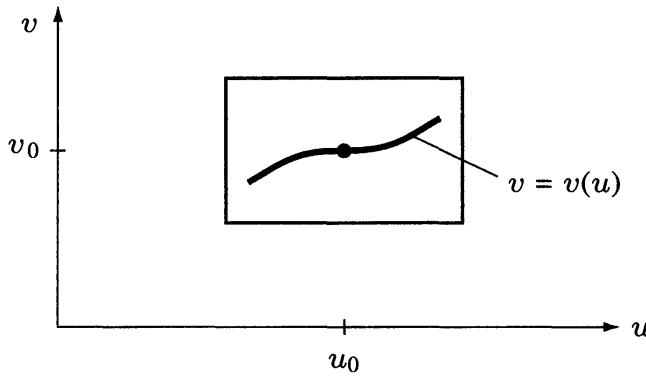


FIGURE 4.3.

(ii) *The function $u \mapsto v(u)$ is C^n on U . In particular,*

$$v'(u) = -F_v(u, v(u))^{-1} F_u(u, v(u)) \quad \text{for all } u \in U. \quad (51)$$

Since this theorem ensures the existence of $v^{(m)}(u)$ for $m = 1, 2, \dots, n$, these derivatives can be computed by means of (9) in the following way. For example, let $m = 2$. By (51),

$$F_u(u + tk, v(u + tk))h + F_v(u + tk, v(u + tk))v'(u + tk)h = 0.$$

Differentiating this at the point $t = 0$ by using the chain rule and the product rule, we get

$$\begin{aligned} & F_{uu}(u, v(u))kh + F_{vu}(u, v(u))v'(u)kh + F_{uv}(u, v(u))kv'(u)h \\ & + F_{vv}(u, v(u))v'(u)kv'(u)h + F_v(u, v(u))v''(u)kh = 0 \end{aligned}$$

for all $h, k \in X$. Applying the inverse operator $F_v(u, v(u))^{-1}$ to this, we obtain $v''(u)kh$. The same argument can be used for computing $v^{(m)}(u)$ with $m \geq 3$.

Proof. To simplify notation, assume that $u_0 = 0$ and $v_0 = 0$. Assume also that $X \neq \{0\}$, $Y \neq \{0\}$, and $Z \neq \{0\}$. Otherwise, the statements are trivial. Set

$$f(u, v) := F_v(0, 0)v - F(u, v)$$

and

$$A_u v := F_v(0, 0)^{-1} f(u, v). \quad (52)$$

To avoid confusion, note that u is an index in “ A_u ” but not a partial derivative. Then, the original problem (48) is equivalent to the *fixed-point problem* $F_v(0, 0)v = f(u, v)$, that is,

$$v = A_u v, \quad v \in X. \quad (53)$$

Step 1: We apply the *Banach fixed-point theorem* to (53).

By the *continuous inverse mapping theorem* from Section 3.5, the inverse operator $F_v^{-1}(0, 0): Z \rightarrow Y$ is linear and continuous. Furthermore, it follows from Proposition 7 in Section 1.23 of AMS Vol. 108 that the operator

$$F_v(u, v)^{-1}: Z \rightarrow Y \quad (54)$$

is linear and continuous for all (u, v) in a sufficiently small neighborhood W of $(0, 0)$ in $X \times Y$ and $\sup_{(u,v) \in W} \|F_v(u, v)^{-1}\| < \infty$.

Let $\|u\|, \|v\|, \|w\| \leq r$. Observe

$$f_v(u, v) = F_v(0, 0) - F_v(u, v).$$

Since f_v is continuous at $(0, 0)$ and $f_v(0, 0) = 0$, *Taylor's theorem* in Section 4.5 implies that⁷

$$\begin{aligned} \|f(u, v) - f(u, w)\| &\leq \sup_{0 \leq \tau \leq 1} \|f_v(u, v + \tau(w - v))\| \|v - w\| \\ &= o(1)\|v - w\|, \quad r \rightarrow 0. \end{aligned}$$

Since $f(0, 0) = 0$ and f is continuous at $(0, 0)$, we get

$$\begin{aligned} \|f(u, v)\| &\leq \|f(u, v) - f(u, 0)\| + \|f(u, 0)\| \\ &\leq o(1)\|v\| + \|f(u, 0)\|, \quad r \rightarrow 0, \end{aligned}$$

where $\|f(u, 0)\| \rightarrow 0$ as $\|u\| \rightarrow 0$. Finally, note that

$$\begin{aligned} \|A_u v\| &\leq \|F_v(0, 0)^{-1}\| \|f(u, v)\|, \\ \|A_u v - A_u w\| &\leq o(1)\|F_v(0, 0)^{-1}\| \|v - w\|, \quad r \rightarrow 0. \end{aligned}$$

Let $M := \{v \in Y: \|v\| \leq r\}$ and $U := \{u \in X: \|u\| < \rho\}$. Then, for sufficiently small $r > 0$ and $\rho > 0$, we obtain

- (a) $\|A_u v\| \leq r$ and
- (b) $\|A_u v - A_u w\| \leq \frac{1}{2}\|v - w\|$

for each given $u \in U$ and all $v, w \in M$.

By the *Banach fixed-point theorem* from Section 1.6 of AMS Vol. 108, the operator $A_u: M \rightarrow M$ has a unique fixed point $v(u)$, that is, for each $u \in U$, equation (53) has a unique solution $v(u) \in M$.

This is statement (i) of Theorem 4.E.

Step 2: We choose the numbers $r > 0$ and $\rho > 0$ to be so small that

$$\|F_v(u, v)^{-1}\| \leq \text{const} \quad \text{for all } u \in U, v \in M, \quad (55)$$

⁷Recall that $a(r) = o(1)$ as $r \rightarrow 0$ means that $a(r) \rightarrow 0$ as $r \rightarrow 0$.

and

$$(u, v) \mapsto F_v(u, v)^{-1} \text{ is continuous from } U \times M \text{ to } L(Z, Y). \quad (56)$$

This is possible by (54). Furthermore, we also may assume that F is C^n on some open subset W of $X \times Y$ with $U \times M \subset W$.

Step 3: We show that $u \mapsto v(u)$ is continuous on U . Let $u, z \in U$. Then

$$\begin{aligned} \|v(u) - v(z)\| &= \|A_u v(u) - A_z v(z)\| \\ &\leq \|A_u v(u) - A_u v(z)\| + \|A_u v(z) - A_z v(z)\| \\ &\leq \frac{1}{2} \|v(u) - v(z)\| + R(u), \end{aligned}$$

where $R(u) \rightarrow 0$ as $u \rightarrow z$, by (52). Hence $\|v(u) - v(z)\| \leq 2R(u) \rightarrow 0$ as $u \rightarrow z$.

Step 4: We show that the F -derivative $v'(u)$ exists for each $u \in U$. Let $h \in X$ and set

$$k := v(u + h) - v(u)$$

along with $v := v(u)$, where $\|h\|$ is sufficiently small. Since F is C^1 on W , it follows from

$$0 = F(u + h, v + k) - F(u, v)$$

that

$$0 = F_u(u, v)h + F_v(u, v)k + o(\|h\| + \|k\|), \quad \|h\| + \|k\| \rightarrow 0.$$

Hence

$$k = -F_v(u, v)^{-1}(F_u(u, v)h + o(\|h\| + \|k\|)). \quad (57)$$

Since $k \rightarrow 0$ as $h \rightarrow 0$, this implies

$$\|k\| \leq \text{const}\|h\| + 2^{-1}\|k\|$$

if $\|h\|$ is sufficiently small. Therefore, $\|k\| \leq \text{const}\|h\|$, and hence

$$v(u + h) - v(u) = -F_v(u, v)^{-1}F_u(u, v)h + o(\|h\|), \quad h \rightarrow 0.$$

Thus, the F -derivative $v'(u)$ exists, where

$$v'(u) = -F_v(u, v(u))^{-1}F_u(u, v(u)) \quad \text{for all } u \in U. \quad (58)$$

By (56) and Step 3, $u \mapsto v'(u)$ is *continuous* on U .

Step 5: Suppose that F is C^2 on W . Then, applying the *chain rule* to (58), it follows that the second F -derivative $u \mapsto v''(u)$ is *continuous* on U . In this connection, note the following:

(a) Since $u \rightarrow v(u)$ is C^1 on U , the map

$$u \rightarrow (u, v(u))$$

is also C^1 from U to $X \times Y$, by Proposition 4 in Section 4.6.

(b) The map

$$A \mapsto A^{-1}$$

is C^∞ from $L_{\text{inv}}(Y, Z)$ to $L(Z, Y)$, by Standard Example 9 in Section 4.3. Moreover, observe that $F_v(u, v(u)) \in L_{\text{inv}}(Y, Z)$ for all $u \in U$, by (56).

(c) The bilinear bounded map

$$(R, S) \mapsto -RS$$

is C^∞ from $L(Z, Y) \times L(X, Z)$, by Example 5 in Section 4.2. Moreover, observe that

$$F_v(u, v(u))^{-1} \in L(Z, Y) \quad \text{and} \quad F_u(u, v(u)) \in L(X, Z).$$

Step 6: Suppose that F is C^n on W , $n \geq 3$. Using the same argument as in Step 5, it follows that $v = v(u)$ is C^n on U , by induction. \square

Applications of the implicit function theorem to integral equations can be found in Section 5.13.

4.9 Applications to Differential Equations

Let us study the following initial-value problem:

$$\begin{aligned} x'(t) &= f(t, x(t), p) && \text{for all } t \in]t_0 - a, t_0 + a[, \\ x(t_0) &= y. \end{aligned} \tag{59}$$

For given $t_0 \in \mathbb{R}$ and $y \in X$, we are looking for a solution $x(\cdot)$ with $x(t) \in X$ for all $t \in]t_0 - a, t_0 + a[$, where X is a Banach space over \mathbb{K} . Here p denotes a parameter living in the Banach space P over \mathbb{K} (e.g., $P = \mathbb{R}$).

In the special case where $X := \mathbb{R}^N$, $N \geq 1$, problem (59) corresponds to a system

$$x'_j(t) = f_j(t, x(t), p(t)), \quad j = 1, \dots, N,$$

of N real differential equations.

The following result is fundamental to the theory of ordinary differential equations.

Proposition 1. *Suppose that the mapping*

$$f: U \subseteq \mathbb{R} \times X \times P \rightarrow X$$

is C^n for fixed n ($1 \leq n < \infty$), where U is an open set containing the given point (t_0, x_0, p_0) . Then the following are true:

- (i) There exist a number $a > 0$ and an open neighborhood $V(x_0, p_0)$ of (x_0, p_0) in $X \times P$ such that the original initial-value problem (59) has a unique solution $x = x(t; y, p)$ for each $(y, p) \in V(x_0, p_0)$.
- (ii) The solution depends smoothly on the initial value y and the parameter p , that is, the mapping

$$(t, y, p) \mapsto x(t; y, p)$$

is C^n from $[t_0 - a, t_0 + a] \times V(x_0, p_0)$ to X .

The original system (59) is equivalent to the following system:

$$\begin{aligned} x'(t) &= f(t, x(t), p(t)), \\ p'(t) &= 0 \quad \text{on } [t_0 - a, t_0 + a], \\ x(t_0) &= y, \quad p(t_0) = p, \end{aligned} \tag{59*}$$

where the values of the unknown function $t \mapsto (x(t), p(t))$ live in the product space $X \times P$. Observe that (59*) does not contain any additional parameters. Thus, it is sufficient to prove Proposition 1 in the case where f is independent of the parameter p .

Our proof will be based on the *implicit function theorem*. The idea is the following:

- (α) Existence. We apply the implicit function theorem to the *rescaled* problem (60).
- (β) Smoothness. We differentiate the original problem (59). This way, we obtain the *new* problem (70), which tells us that the solution of (59) is C^n .

Proof. To simplify the notation only, we suppose that $U = \mathbb{R} \times X$.

Ad (i). *Step 1: Rescaling.* Set $J := [-1, 1]$. The decisive *trick* of the proof is the following rescaling:

$$t = t_0 + sa, \quad z(s; \tau, y) := x(\tau + as; y) - y \quad \text{for all } s \in J.$$

Then the original problem (59) corresponds to the *new* problem:

$$\begin{aligned} z'(s) - af(\tau + as, z(s) + y) &= 0 \quad \text{for all } s \in J, \\ z(0) &= 0 \end{aligned} \tag{60}$$

with fixed $\tau = t_0$. We write this as an operator equation⁸

$$F(z, a, \tau, y) = 0 \tag{61}$$

⁸In the following, let $\tau \in \mathbb{R}$ be a free parameter. This is important in order to prove *continuity* properties of the solution $x = x(t)$.

with the operator $F: Z \times \mathbb{R} \times \mathbb{R} \times X \rightarrow W$ and the real Banach spaces

$$Z := \{z \in C^1(J, X): z(0) = 0\}, \quad W := C(J, X).$$

The norm on Z and W is given by

$$\|z\|_Z := \max_{s \in J} \|z(s)\|_X + \max_{s \in J} \|z'(s)\|_X$$

and

$$\|z\|_W := \max_{s \in J} \|z(s)\|_X.$$

Step 2: The implicit function theorem. Set $Q := (0, 0, t_0, x_0)$, namely, $z = 0$, $a = 0$, $\tau = t_0$, $y = x_0$. Obviously, $F(Q) = 0$, and the *linearization* of (60) at Q yields

$$F_z(Q)z = z'.$$

The *crucial point* is that, for every $w \in W$, there exists exactly one $z \in Z$ with $z' = w$, namely, $z(s) = \int_0^s w(t)dt$. Hence $F_z(Q): Z \rightarrow W$ is *bijective*.

Thus, by the *implicit function theorem* from Section 4.8, there exist numbers $\rho > 0$ and $r > 0$ such that, for given $a > 0$, $\tau \in \mathbb{R}$, and $y \in X$ with

$$a, |\tau - t_0|, \|y - x_0\|_X < \rho, \quad (62)$$

the operator equation (61) has a unique solution $z \in Z$ with $\|z\|_Z < r$. In addition, the map

$$(a, \tau, y) \mapsto z \quad (63)$$

is C^1 from the open subset (62) in $\mathbb{R} \times \mathbb{R} \times X$ to Z .

The remaining part of the proof is routine.

However, for the convenience of the reader, let us discuss this in detail.

Step 3: Uniqueness for the original problem (59). It follows from Step 2 that each solution of (59) is *locally unique*. In fact, consider a solution $x = x(t)$ of (59), say at the point t_0 . Since the function $x(\cdot)$ is differentiable, it is continuous, and hence $x'(\cdot)$ is continuous, by (59). Thus, $x(\cdot)$ is C^1 . Moreover, if we choose the number $a > 0$ sufficiently small, then $\|z\|_Z < r$, and Step 2 yields the local uniqueness. That is, $x(\cdot)$ is unique on a small neighborhood of t_0 .

Furthermore, observe that local uniqueness implies *global uniqueness* by using the following standard argument. Let $x = x(t)$ and $\chi = \chi(t)$ be two solutions of (59) for some fixed $a > 0$. Let $I :=]t_0 - c, t_0 - b[$ be the largest open interval such that

$$x(t) = \chi(t) \quad \text{on } I.$$

For example, suppose that $b < a$. By continuity, $x(b) = \chi(b)$. Applying the local uniqueness to the point b , we obtain that $x(t) = \chi(t)$ on some small neighborhood of $t = b$. This contradicts the maximality of I .

Ad (ii). Let $x(\cdot)$ be the solution of (59) from Step 2. Introduce the partial derivatives

$$v(t, y) := x_t(t, y) \quad \text{and} \quad w(t, y) := x_y(t, y) \quad \text{on } \mathcal{U},$$

where $\mathcal{U} := \{(t, y) \in \mathbb{R} \times X : |t - t_0| < a, \|y - x_0\|_X < \rho\}$. Let us show that

$$v \text{ and } w \text{ are } C^{n-1} \text{ on } \mathcal{U} \quad (64)$$

provided a and ρ are sufficiently small depending on n . By Problem 4.11, this implies that $x = x(t, y)$ is C^n on \mathcal{U} . Let us prove (64) by induction.

Step A: Let $n = 1$. We want to show the following:

- (a) $(t, y) \mapsto x(t, y)$ is continuous on \mathcal{U} .
- (b) $(t, y) \mapsto x_t(t, y)$ is continuous on \mathcal{U} .
- (c) $(t, y) \mapsto x_y(t, y)$ is continuous on \mathcal{U} .
- (d) For each $(t, y) \in \mathcal{U}$, there exists $x_{ty}(t, y) = x_{yt}(t, y)$.

Ad (a). The continuity of the mapping from (63) implies that

$$\|z(t_0 + \sigma, y + h) - z(t_0, y)\|_Z < \varepsilon$$

provided $|\sigma|$ and $\|h\|_X$ are sufficiently small. This means that

$$\max_{s \in J} |z(s; t_0 + \sigma, y + h) - z(s; t_0, y)| < \varepsilon.$$

Thus, $(t, y) \mapsto x(t, y)$ is continuous on \mathcal{U} , by the definition of the function z in Step 1.

Ad (b). Use the original equation (59) and (a).

Ad (c), (d). Since the mapping from (63) is C^1 , there exists the continuous derivative $A := z_y$. That is, the operator $A: X \rightarrow Z$ is linear and

$$\|z(y + h) - z(y) - Ah\|_Z = o(\|h\|), \quad h \rightarrow 0 \text{ in } X. \quad (65)$$

The continuity of z_y means that

$$\|A(t_0 + \sigma, y + k)h - A(t_0, y)h\|_Z \leq \varepsilon\|h\|_X \quad \text{for all } h \in X \quad (66)$$

if $|\sigma|$ and $\|k\|_X$ are sufficiently small. Relation (65) implies

$$\begin{aligned} & \max_{s \in J} |z(s; \tau, y + h) - z(s; \tau, y) - (Ah)(s)| \\ & + \max_{s \in J} |z_s(s; \tau, y + h) - z_s(s; \tau, y) - (Ah)'(s)| = o(\|h\|), \quad h \rightarrow 0 \text{ in } X. \end{aligned}$$

By the definition of the derivative, this yields

$$z_y(s; \tau, y)h = (Ah)(s) \quad (67)$$

and $z_{ys}(s; \tau, y) = (Ah)'(s)$, that is,

$$z_{ys}(s; \tau, y) = z_{sy}(s; \tau, y), \quad (68)$$

for all $h \in X$ and all $s \in J$, $(\tau, y) \in \mathcal{U}$. Moreover, for $s \in J$, it follows from (66) that

$$\|z_y(s; t_0 + \sigma, y + k)h - z_y(s; t_0, y)h\|_X \leq \varepsilon \|h\|_X \quad \text{for all } h \in X,$$

if $|\sigma|$ and $\|k\|_X$ are sufficiently small. Hence

$$\|z_y(s; t_0 + \sigma, y + k) - z_y(s; t_0, y)\| \leq \varepsilon. \quad (69)$$

Using the definition of the function z in Step 1, we get (d) and (c) from (68) and (69), respectively.

Step B: Using (59*), we obtain Proposition 1 for $n = 1$ in the case where f depends on a parameter.

Step C: Let $n \geq 2$. Suppose that Proposition 1 has been proved for $n - 1$. Let $x = x(t, y)$ be the solution of the original problem (59) from Step A. Assume that $(t, y) \mapsto x(t, y)$ is C^{n-1} on \mathcal{U} .

Differentiating the original equation (59) with respect to y and t , and using $x_{ty}(t, y) = x_{yt}(t, y)$, we get⁹ the following *new* linear initial-value problem:

$$v'(t) = f_t(t, x(t, y)) + f_x(t, x(t, y))v(t), \quad w'(t) = f_x(t, x(t, y))w(t) \quad (70a)$$

$$v(t_0) = \alpha \quad \text{and} \quad w(t_0) = I, \quad (70b)$$

where $\alpha := x_t(t_0, y) = f(t_0, y)$. For $y = y_0$, we set $\alpha_0 := f(t_0, y_0)$. Since $(t, y) \mapsto x(t, y)$ is C^{n-1} on \mathcal{U} , the right-hand side of (70a) is also C^{n-1} as a function of (t, y, v, w) . Using Proposition 1 for $n - 1$, we obtain that the map

$$(t, \alpha) \mapsto (v(t; \alpha), w(t; \alpha))$$

is C^{n-1} on a small neighborhood of the point (t_0, α_0) . Since

$$y \mapsto \alpha \equiv f(t_0, y)$$

is C^{n-1} on small neighborhood of y_0 , we obtain the desired result (64). \square

4.10 Diffeomorphisms and the Local Inverse Mapping Theorem

Let us apply the *implicit function theorem* to the study of the *local* behavior of nonlinear mappings. The results of this section and the next are of great importance for nonlinear analysis.

⁹To simplify notation, we write $v(t)$ and $w(t)$ instead of $v(t, y) = x_t(t, y)$ and $w(t, y) = x_y(t, y)$, respectively.

Definition 1. Let U and V be nonempty open sets in the Banach spaces X and Y over \mathbb{K} . Let $0 \leq r \leq \infty$.

The mapping $f: U \rightarrow V$ is called a C^r -diffeomorphism iff f is bijective and both f and f^{-1} are C^r -maps.

A local C^r -diffeomorphism at the point u_0 is a C^r -diffeomorphism from some open neighborhood $U(u_0)$ in X onto some open neighborhood¹⁰ $V(f(u_0))$ in Y .

Obviously, C^0 diffeomorphisms are homeomorphisms.

Theorem 4.F (The local inverse mapping theorem). *Let $f: U(u_0) \subseteq X \rightarrow Y$ be a C^r -map on some open neighborhood of the point u_0 , where X and Y are Banach spaces over \mathbb{K} and $1 \leq r \leq \infty$.*

Then f is a local C^r -diffeomorphism at u_0 iff $f'(u_0): X \rightarrow Y$ is bijective.

A global version of the inverse mapping theorem will be considered in Problem 4.12.

Proof. Step 1: Let $f'(u_0): X \rightarrow Y$ be bijective and set $v_0 := f(u_0)$. Furthermore, set

$$F(u, v) := f(u) - v.$$

Then the equation

$$F(u, v) = 0, \quad u \in X, v \in Y, \quad (71)$$

can be solved for u by the *implicit function theorem* in Section 4.8, because $F(u_0, v_0) = 0$, and the map $F_u(u_0, v_0) = f'(u_0)$ is bijective from X to Y . Thus, there exist open neighborhoods $\mathcal{U}(u_0)$ and $\mathcal{V}(v_0)$ such that, for each $v \in \mathcal{V}(v_0)$, equation (71) has a unique solution $u = u(v)$ in $\mathcal{U}(u_0)$. The map

$$v \mapsto u(v)$$

is C^r . Obviously, $u(v) = f^{-1}(v)$ for all $v \in \mathcal{V}(v_0)$.

If we set $g(v) := f^{-1}(v)$, then

$$f(g(v)) = v \quad \text{and} \quad g(f(u)) = u. \quad (72^*)$$

Differentiation by the *chain rule* gives

$$f'(u)g'(v) = I \quad \text{and} \quad g'(v)f'(u) = I, \quad (72)$$

where $v := f(u)$. Hence $g'(v) = f'(u)^{-1}$, that is,

$$(f^{-1})'(v) = f'(u)^{-1} \quad \text{for all } u \in \mathcal{U}(u_0). \quad (73)$$

¹⁰We also speak of a local diffeomorphism between the points u_0 and $f(u_0)$.

Step 2: Let f be a local C^r -diffeomorphism at u_0 . Then (72*) implies equation (72) for all $u \in \mathcal{U}(u_0)$. Hence $f'(u): X \rightarrow Y$ is bijective for all $u \in \mathcal{U}(u_0)$. \square

Corollary 2. Let $f: U \rightarrow V$ and $g: V \rightarrow W$ be C^r -diffeomorphisms, where U , V , and W are nonempty open sets in Banach spaces over \mathbb{K} and $1 \leq r \leq \infty$.

Then the composite map $g \circ f: U \rightarrow W$ is also a C^r -diffeomorphism.

Proof. Obviously, $g \circ f$ is bijective. By the chain rule,

$$(g \circ f)'(u) = g'(f(u))f'(u) \quad \text{for all } u \in U.$$

According to the *inverse mapping theorem* (Theorem 4.F), $f'(u)$ and $g'(f(u))$ are *bijective*, and hence so is $(g \circ f)'(u)$. Thus, again using the inverse mapping theorem, we prove the assertion in Corollary 2. \square

4.11 Equivalent Maps and the Linearization Principle

Let us now study the case where $f'(u_0)$ is *not* bijective, that is, where f is *not* a local diffeomorphism at u_0 . Our point of departure is the following commutative diagram:

$$\begin{array}{ccc} U & \xrightarrow{f} & V \\ \phi \uparrow & & \downarrow \psi \\ W & \xrightarrow{g} & Z \end{array} \tag{74}$$

Definition 1. Let $f: U \rightarrow V$ and $g: W \rightarrow Z$ be two C^r -maps where U , V , W , and Z are nonempty open subsets of Banach spaces over \mathbb{K} . Let $0 \leq r \leq \infty$.

We say that the map f at the point u_0 is C^r -equivalent to the map g at the point w_0 iff there exist a local C^r -diffeomorphism ϕ between w_0 and u_0 and a local C^r -diffeomorphism ψ between $f(u_0)$ and $g(w_0)$ such that the diagram (74) is locally commutative. By definition, this means that we have

$$g = \psi \circ f \circ \phi \quad \text{on some open neighborhood of } u_0.$$

We write

$$f \overset{r}{\sim} g \text{ at } (u_0, w_0)$$

iff f at u_0 is C^r -equivalent to g at w_0 . To discuss this, let

$$v = f(u) \text{ on some open neighborhood of } u_0.$$

If we introduce the *new local coordinates* $u = \phi(w)$ and $z = \psi(v)$, then we get

$$z = g(u) \text{ on some open neighborhood of } w_0.$$

Roughly speaking:

Equivalent maps locally possess the same structure.

Since local diffeomorphisms are invariant under composition of maps and inverse formation, we find that the equivalence of maps represents an *equivalence relation*. That is,

- (i) $f \sim g$ at (u_0, w_0) implies $g \sim f$ at (w_0, u_0) .
- (ii) $f \sim g$ at (u_0, w_0) and $g \sim h$ at (w_0, z_0) imply $f \sim h$ at (u_0, z_0) .

Example 2. Let $f: U(u_0) \subseteq \mathbb{R} \rightarrow \mathbb{R}$ be C^r on some open neighborhood of u_0 ($1 \leq r \leq \infty$). Then f at u_0 is C^r -equivalent to both the *linearizations*

$$u \mapsto f(u_0) + f'(u_0)(u - u_0) \quad \text{at } u_0 \quad (75)$$

and

$$u \mapsto f'(u_0)u \quad \text{at } u = 0. \quad (75^*)$$

Relation (75*) is a direct consequence of Theorem 4.G, which follows the next definition. Use the translations $u \mapsto u - u_0$ and $v \mapsto v - f(u_0)$ in order to obtain (75) from (75*).

If $f'(u_0) \neq 0$, then (75*) is also C^∞ -equivalent to $u \mapsto u$ at $u = 0$. In fact, if we set $w := f'(u_0)^{-1}v$, then $v = f'(u_0)u$ is transformed into the new equation $w = u$.

Definition 3. Let $f: U(u_0) \subseteq X \rightarrow Y$ be a C^1 -map on an open neighborhood of u_0 , where X and Y are Banach spaces over \mathbb{K} . Then

- (i) f is called a *submersion* at u_0 iff $f'(u_0): X \rightarrow Y$ is *surjective* and the null space $N(f'(u_0))$ splits X .
- (ii) f is called an *immersion* at u_0 iff $f'(u_0): X \rightarrow Y$ is *injective* and the range $R(f'(u_0))$ splits Y .
- (iii) f is called a *subimmersion* at u_0 iff either

$$X \text{ and } Y \text{ have finite dimensions} \quad (76)$$

and rank $f'(u)$ is *constant* on some open neighborhood of u_0 or condition (76) is not satisfied and $N(f'(u_0))$ splits X , $R(f'(u_0))$ splits Y , as well as

$$f'(u)(N_c) = f'(u)(X)$$

for all u on some open neighborhood of u_0 , where N_c is given in such a way that $X = N(f'(u_0)) \oplus N_c$ is a fixed topological sum.

Theorem 4.G (The linearization principle). *Let $f: U(u_0) \subseteq X \rightarrow Y$ be a C^r -map on an open neighborhood of u_0 , where X and Y are Banach spaces over \mathbb{K} and $1 \leq r \leq \infty$. Suppose that f is a submersion, an immersion, or a subimmersion at u_0 .*

Then f at u_0 is C^r -equivalent to the linearization $f'(u_0): X \rightarrow Y$ at $u = 0$.

Corollary 4. *In addition, the following hold:*

- (i) *If f is a submersion at u_0 , then f is locally surjective at u_0 , that is, there exists an open neighborhood $\mathcal{U}(u_0)$ in X and a number $\rho > 0$ such that the equation*

$$f(u) = v, \quad u \in \mathcal{U}(u_0) \quad (77)$$

has a solution for each $v \in Y$ with $\|v - f(u_0)\| < \rho$.

- (ii) *If f is an immersion at u_0 , then f is locally injective at u_0 , that is, there exists an open neighborhood $\mathcal{U}(u_0)$ in X such that, for each $v \in Y$, equation (77) has at most one solution.*

Theorem 4.G is also called the *rank theorem* since subimmersions on finite-dimensional spaces have locally constant rank.

Important applications of this theorem to the theory of manifolds can be found in Zeidler (1986), Vol. 4, Section 73.11.

Proof of Corollary 4. Use Theorem 4.G and the observation that the linearized equation

$$f'(u_0)(u - u_0) = v, \quad u \in \mathcal{U}(u_0), \quad (77^*)$$

has the corresponding properties. In fact, if f is a submersion at u_0 , then the solutions of (77*) are given through $u = u_0 + A^{-1}v + w$, where $w \in N(f'(u_0))$ and

$$A: N_c \rightarrow Y$$

denotes the *restriction* of $f'(u_0)$ to N_c . Note that A is a linear homeomorphism, by Proposition 13 in Section 4.9.

If f is an immersion at u_0 , then $f'(u_0)$ is injective, and hence equation (77*) has at most one solution. \square

Set $N := N(f'(u_0))$ and $R := R(f'(u_0))$. Choose fixed topological direct sums

$$X = N \oplus N_c \quad \text{and} \quad Y = R \oplus R_c, \quad (78)$$

along with the corresponding continuous projections

$$P: X \rightarrow N \quad \text{and} \quad Q: Y \rightarrow R.$$

Proof of Theorem 4.G. *Step 1:* Let f be a *submersion* at u_0 . Let the operator A be given as in the proof of Corollary 4. Define

$$\phi(u) := Pu + A^{-1}f(u). \quad (79)$$

Since $f'(u_0)h = f'(u_0)(I - P)h$ and hence $A^{-1}f'(u_0)(I - P)h = A^{-1}A(I - P)h = (I - P)h$, we get

$$\phi'(u_0)h = Ph + A^{-1}f'(u_0)h = h \quad \text{for all } h \in X.$$

By the *inverse mapping theorem* (Theorem 4.F), ϕ is a local C^r -diffeomorphism at u_0 . Multiplying (79) by $f'(u_0)$, we obtain

$$f'(u_0)\phi(u) = f(u) \quad \text{on some open neighborhood of } u_0.$$

This way, we get the commutative diagram:

$$\begin{array}{ccc} \mathcal{U}(u_0) \subseteq X & \xrightarrow{f} & Y \\ \phi \downarrow & & \nearrow f'(u_0) \\ \mathcal{U}(0) \subseteq X. & & \end{array}$$

Hence f at u_0 is C^r -equivalent to $f'(u_0)$ at $u = 0$.

Step 2: Let f be an *immersion* at u_0 . After a translation we may assume that $u_0 = 0$ and $f(0) = 0$. Define

$$\phi(v) := f(f'(0)^{-1}Qv) + (I - Q)v. \quad (80)$$

Then

$$\phi'(0)k = Qk + (I - Q)k = k \quad \text{for all } k \in Y.$$

By the *inverse mapping theorem*, ϕ is a local C^r -diffeomorphism at $v = 0$ with $\phi(0) = f(0)$. Since $Qf'(0) = f'(0)$, we get

$$\phi(f'(0)v) = f(v)$$

for all v on some open neighborhood of $v = 0$. Thus, the following diagram is commutative:

$$\begin{array}{ccc} \mathcal{U}(0) \subseteq X & \xrightarrow{f} & Y \\ f'(0) \searrow & & \uparrow \phi \\ & & \mathcal{V}(0) \subseteq Y \end{array}$$

Hence f at $u_0 = 0$ is C^r -equivalent to $f'(0)$ at $u = 0$.

Step 3: Let f be a *subimmersion* at u_0 . The proof will be given in the next section, by using a normal form for double splitting maps. \square

4.12 The Local Normal Form for Nonlinear Double Splitting Maps

We will use the notation from (78). Our goals are the *first normal form*

$$f(\phi(n, r)) = f(u_0) + r + g(n, r) \quad \text{for all } (r, n) \in \mathcal{U}(0, 0) \subseteq N \times R, \quad (81a)$$

where

$$g(n, r) \in R_c \text{ on } \mathcal{U}(0, 0), \quad g(0, 0) = 0, \quad g'(0, 0) = 0, \quad (81b)$$

and the *second normal form*

$$f(\psi(u)) = f(u_0) + f'(u_0)(u - u_0) + a(u) \quad \text{for all } u \in \mathcal{U}(u_0), \quad (82a)$$

where

$$a(u) \in R_c \text{ on } \mathcal{U}(u_0), \quad a(u_0) = 0, \quad a'(u_0) = 0. \quad (82b)$$

This can be regarded as a variant of the Taylor theorem.

Proposition 1. *Let $f: U(u_0) \subseteq X \rightarrow Y$ be a C^r -map on an open neighborhood of u_0 , where X and Y are Banach spaces over \mathbb{K} and $1 \leq r \leq \infty$. Suppose that the null space $N := N(f'(u_0))$ splits X and the range $R := R(f'(u_0))$ splits Y . Then the following are true:*

- (i) *There exist an open neighborhood $\mathcal{U}(0, 0)$ in $N \times R$ and a local C^r -diffeomorphism $\phi: \mathcal{U}(0, 0) \rightarrow X$ between $(0, 0)$ and u_0 such that (81) holds.*

- (ii) There exist an open neighborhood $\mathcal{U}(u_0)$ in X and a local C^r -diffeomorphism $\psi: \mathcal{U}(u_0) \rightarrow X$ between u_0 and u_0 such that (82) holds.

Proof. Without loss of generality, let $u_0 = 0$ and $f(u_0) = 0$.

Ad (i). The *proof idea* is to apply the *inverse mapping theorem* to the mapping

$$F(u) := (u_1, f_1(u))$$

and to let $\phi := F^{-1}$.

The topological direct sums $X = N \oplus N_c$ and $Y = R \oplus R_c$ yield the decompositions

$$u = u_1 + u_2 \quad \text{and} \quad f(u) = f_1(u) + f_2(u),$$

for $u \in X$ and $f(u) \in Y$, respectively (i.e., $u_1 \in N$, $u_2 \in N_c$, and $f_1(u) \in R$, $f_2(u) \in R_c$).

Since $f(0) = 0$ and

$$f'(0)h = f'_1(0)h + f'_2(0)h \quad \text{for all } h \in X$$

with $f'(0)h \in R$, we obtain

$$f_1(0) = f_2(0) = 0 \quad \text{and} \quad f'_2(0) = 0.$$

The map $F: \mathcal{U}(0) \subseteq X \rightarrow N \times R$, as defined above, is C^r with $F(0) = (0, 0)$ and

$$F'(0)h = (h_1, f'_1(0)h) = (h_1, f'(0)h) \quad \text{for all } h \in X.$$

Since $f'(0): N_c \rightarrow R$ is bijective, it follows from $F'(0)h = 0$ that $h_1 = 0$ and hence $h_2 = 0$ (i.e., $h = 0$). Thus, $F'(0): X \rightarrow N \times R$ is bijective, and the *inverse mapping theorem* (Theorem 4.F) implies that F is a local C^r -diffeomorphism at $u_0 = 0$.

Letting $\phi := F^{-1}$, we get $\phi(n, r) = u$, where $n = u_1$ and $r = f_1(u)$. Thus

$$f(\phi(n, r)) = f_1(u) + f_2(u) = r + f_2(\phi(n, r)).$$

This is (81) with $g(n, r) := f_2(\phi(n, r))$.

Finally, we obtain $g(0, 0) = 0$ from $f_2(0) = 0$ and $\phi(0, 0) = 0$. Moreover, $f'_2(0) = 0$ implies

$$g'(0, 0) = f'_2(0)\phi'(0, 0) = 0.$$

Ad (ii). Define $\chi(u) := (n, r)$ by

$$n := Pu \quad \text{and} \quad r := f'(0)u. \tag{83}$$

If $(n, r) = 0$, then $u = 0$, since $f'(0): N_c \rightarrow R$ is bijective. Thus, the map $\chi: X \rightarrow N \times R$ is a C^∞ -diffeomorphism, by the inverse mapping theorem. Letting $\psi(u) := \phi(n, r)$ along with $(n, r) := \chi(u)$, we get

$$\begin{aligned} f(\psi(u)) &= f(\phi(n, r)) = r + g(n, r) \\ &= f'(0)u + a(u), \end{aligned}$$

where $a(u) := g(n, r)$. This is (82). \square

We are now able to finish the proof of Theorem 4.G in the case where f is a *subimmersion* at u . Without loss of generality, we again assume that $u_0 := 0$ and $f(u_0) := 0$. Define

$$H(n, r) := f(\phi(n, r)).$$

Then

(a) $H_n(n, r) = 0$ for all $(n, r) \in N \times R$ on some open neighborhood $\mathcal{U}(0, 0)$. This will be proved ahead.

(b) $H(n, r)$ is independent of n on $\mathcal{U}(0, 0)$. In fact, we may assume that $\mathcal{U}(0, 0)$ is convex. By the *Taylor theorem*,

$$\|H(n_1, r) - H(n_2, r)\| \leq \sup_{0 \leq \tau \leq 1} \|H_n(n_1 + \tau(n_2 - n_1), r)\| = 0,$$

for all $(n_1, r), (n_2, r) \in \mathcal{U}(0, 0)$.

(c) Set $G(r) := H(n, r)$. Due to (b) and (81),

$$G(r) = H(0, r) = r + g(0, r).$$

Since $G'(0)h = h + g'(0, 0)h = h$ for all $h \in R$, the operator $G'(0): R \rightarrow Y$ is *injective*. By Theorem 4.G for immersions, G at $r = 0$ is equivalent to $G'(0)$ at $r = 0$. Thus, Step 2 of the proof of Theorem 4.G tells us that there exists a local C^r -diffeomorphism $\psi: \mathcal{U}(0) \subseteq Y \rightarrow R$ between $v = 0$ and $r = 0$ such that $(\psi \circ G)(r) = G'(0)r$, that is,

$$(\psi \circ f \circ \phi)(n, r) = r$$

on some open neighborhood of $(0, 0)$ in $N \times R$. Consequently, f at $u_0 = 0$ is equivalent to the mapping

$$(n, r) \mapsto r \quad . \quad (84)$$

from $N \times R$ to R at the point $(0, 0)$.

(d) Obviously, the mapping from (84) at $(0, 0)$ is equivalent to $f'(0)$ at $u = 0$, by means of (83). In fact,

$$f'(0)\chi^{-1}(n, r) = r \quad \text{for all } (n, r) \in N \times R.$$

(e) By (c) and (d), f at $u_0 = 0$ is equivalent to $f'(0)$ at $u = 0$. This is the *statement of Theorem 4.G for subimmersions*.

It remains to prove (a). In the following, let $(n, r) \in \mathcal{U}(0, 0)$. For all $(h, k) \in N \times R$, we have

$$H'(n, r) = f'(\phi(n, r))\phi'(n, r) \quad (85)$$

and

$$H'(n, r)(h, k) = k + g'(n, r)(h, k).$$

Since $k \in R$ and $g'(n, r)(h, k) \in R_c$, it follows from

$$H'(n, r)(h_1, k_1) = H'(n, r)(h_2, k_2), \quad (86)$$

along with $(h_j, k_j) \in N \times R$, $j = 1, 2$, that $k_1 = k_2$. Thus, the map

$$H'(n, r): \{0\} \times R \rightarrow H'(n, r)(N \times R), \quad (n, r) \in \mathcal{U}(0, 0), \quad (87)$$

is injective. We want to show that

$$\text{The map from (87) is surjective,} \quad (88)$$

provided $\mathcal{U}(0, 0)$ is sufficiently small. In fact, since ϕ is a local C^r -diffeomorphism, the linearization

$$\phi'(n, r): N \times R \rightarrow X$$

is bijective, by the inverse mapping theorem (Theorem 4.F). Thus, by (85),

$$H'(n, r)(N \times R) = R(f'(\phi(n, r))). \quad (89)$$

Case 1: Let $\dim X < \infty$ and $\dim Y < \infty$. Since f is a *subimmersion* at $u_0 = 0$,

$$\dim R(f'(\phi(n, r))) = \text{const} = \dim R(f'(0)) = \dim R.$$

By (89), $\dim H'(n, r)(N \times R) = \dim R$. Thus, the injectivity of (87) implies (88).¹¹

Case 2: Let $\dim X = \infty$ or $\dim Y = \infty$. Since f is a *subimmersion* at $u_0 = 0$,

$$f'(u)(N_c) = f'(u)(X) \quad \text{for all } u \in \mathcal{U}(0).$$

Let $v \in H'(n, r)(N \times R)$ be given. Then $v \in f'(u)(X)$, by (89). Here, $u := \phi(n, r)$. Thus, there is an $n_c \in N_c$ such that $v = f'(u)n_c$. By the proof of Proposition 1, $\phi = F^{-1}$ and

$$F'(u)n_c = (0, f'_1(u)n_c).$$

Thus, we get $\phi'(n, r)(0, f'_1(u)n_c) = n_c$ along with $w := f'_1(u)n_c \in R$. By (85),

$$H'(n, r)(0, w) = f'(u)n_c = v.$$

This yields (88).

¹¹Observe that $\dim(\{0\} \times R) = \dim R < \infty$ in (87).

To finish the proof, let $(h, k) \in N \times R$ be given. Since the map from (87) is surjective, there exists a $\tilde{k} \in R$ such that

$$H'(n, r)(0, \tilde{k}) = H'(n, r)(h, k).$$

By (86), $k = \tilde{k}$. Therefore,

$$H_r(n, r)k = H_n(n, r)h + H_r(n, r)k \quad \text{for all } h \in N,$$

and hence $H_n(n, r) = 0$. This proves part (a).

The proof of Theorem 4.G is now complete. \square

4.13 The Surjective Implicit Function Theorem

We want to solve the equation

$$F(u, v) = 0 \tag{90}$$

under the assumption that¹²

$$F_v(u_0, v_0): Y \rightarrow Z \text{ is surjective.} \tag{91}$$

Theorem 4.H. *Let X , Y , and Z be Banach spaces over \mathbb{K} , and let*

$$F: U(u_0, v_0) \subseteq X \times Y \rightarrow Z$$

be a C^1 -map on an open neighborhood of the point (u_0, v_0) such that $F(u_0, v_0) = 0$ and (91) holds. Then the following are true:

- (i) *Let $r > 0$. There is a number $\rho > 0$ such that, for each given $u \in X$ with $\|u - u_0\| < \rho$, the original equation (90) has a solution v , denoted by $v(u)$, such that*

$$\|v - v_0\| < r.$$

In particular, the limit $u \rightarrow u_0$ in X implies $v(u) \rightarrow v_0$.

- (ii) *There is a number $d > 0$ such that $\|v(u)\| \leq d\|F_v(u_0, v_0)v(u)\|$.*

Proof. Without loss of generality, we may assume that $u_0 = 0$ and $v_0 = 0$.

Step 1: Let $N := \{v \in Y : F_v(0, 0)v = 0\}$. There is a number $d > 0$ such that, for each $z \in Z$, there is a point $w(z) \in Y$ with

$$F_v(0, 0)w(z) = z \quad \text{and} \quad \|w(z)\| \leq d\|z\|. \tag{92}$$

¹²Observe that we do *not* need the null space $N(F_v(u_0, v_0))$ to split Y .

In fact, by the *closed range theorem* in Section 3.12, there is a number $c > 0$ such that

$$\|F_v(0, 0)(v - n)\| \geq c \cdot \text{dist}(v, N) \quad \text{for all } v \in Y, n \in N.$$

Since $F_v(0, 0): Y \rightarrow Z$ is surjective, there exists a $v \in Y$ such that $z = F_v(0, 0)(v - n)$ for all $n \in N$. Moreover, there is some $n \in N$ such that

$$\|v - n\| \leq 2 \cdot \text{dist}(v, N).$$

Step 2: The original equation (90) is equivalent to

$$F_v(0, 0)v = f(u, v) \quad \text{with } f(u, v) := F_v(0, 0)v - F(u, v). \quad (93)$$

Let $\|u\| \leq \rho$ and $\|v\|, \|w\| \leq r$. Then, by Step 1 of the proof of the implicit function theorem (Theorem 4.E),

$$\|f(u, v)\| \leq o(1)\|v\| + \|f(u, 0)\|, \quad r \rightarrow 0, \quad (94)$$

and

$$\|f(u, w) - f(u, v)\| \leq o(1)\|w - v\|, \quad r \rightarrow 0. \quad (95)$$

Step 3: For a given $u \in X$ with $\|u\| \leq \rho$, let us consider the following iterative method:

$$F_v(0, 0)v_{m+1} = f(u, v_m), \quad m = 0, 1, 2, \dots, \quad (96)$$

where $v_0 := 0$ and v_{m+1} is chosen according to (92), that is,

$$v_{m+1} := w(f(u, v_m)).$$

Since $\|v_{m+1}\| \leq d\|f(u, v_m)\|$, it follows from (94) and (95) that, for sufficiently small ρ and r , we get

$$\|v_m\| \leq r \quad \text{and} \quad \|v_{m+2} - v_{m+1}\| \leq 2^{-1}\|v_{m+1} - v_m\| \quad \text{for all } m = 0, 1, \dots.$$

It follows, as in the proof of Theorem 1.A in AMS Vol. 108, that (v_m) is a Cauchy sequence, and hence

$$v_m \rightarrow v \quad \text{as } m \rightarrow \infty$$

for some v . This implies $\|v\| \leq r$ and $F_v(0, 0)v = f(u, v)$, by (96). Hence $F(u, v) = 0$.

Finally, if we let $m \rightarrow \infty$, it follows from

$$\|v_{m+1}\| \leq d\|F_v(0, 0)v_{m+1} - F(u, v_m)\|$$

that $\|v\| \leq d\|F_v(0, 0)v\|$. □

4.14 Applications to the Lagrange Multiplier Rule

Let us consider the minimum problem

$$f(u) = \min !, \quad (97a)$$

along with the side condition

$$G(u) = 0. \quad (97b)$$

Our goal is to justify the *necessary* solvability condition

$$f'(u) + \lambda G'(u) = 0, \quad (98)$$

where λ is called a *Lagrange multiplier*.

Proposition 1. *Let $f: U(u) \subseteq X \rightarrow \mathbb{R}$ and $G: U(u) \subseteq X \rightarrow Y$ be C^1 on an open neighborhood of u , where X and Y are real Banach spaces. Suppose that u is a solution of (97a), (97b), where*

$$G'(u): X \rightarrow Y \text{ is surjective.}$$

Then there exists a functional $\lambda \in Y^$ such that (98) holds true.*

Sufficient solvability conditions can be found in Zeidler (1986), Vol. 3, Section 43.8.

In the special case where $Y := \mathbb{R}^n$ and $G(u) = (g_1(u), \dots, g_n(u))$, the *surjectivity* of $G'(u): X \rightarrow Y$ is equivalent to the fact that, for each $(w_1, \dots, w_n) \in \mathbb{R}^n$, the system

$$g'_j(u)h = w_j, \quad j = 1, \dots, n,$$

has a solution $h \in X$. Then condition (98) is equivalent to

$$f'(u) + \sum_{j=1}^n \lambda_j g'_j(u) = 0,$$

where $\lambda = (\lambda_1, \dots, \lambda_n) \in \mathbb{R}^n$.

The following proof will be based on the *surjective implicit function theorem* and the *closed range theorem*.

We have shown in Section 5.17.3 of AMS Vol. 108 that in terms of statistical physics, absolute temperature is nothing other than a Lagrange multiplier. Moreover, applications of the Lagrange multiplier rule to capillary surfaces can be found in Problem 2.12.

Proof. Let u be a solution of (97). Note that $G(u) = 0$. We want to show that

$$G'(u)h = 0 \quad \text{implies} \quad f'(u)h = 0.$$

To this end, let $h \in X$ be given such that $G'(u)h = 0$. Set

$$F(\varepsilon, v) := G(u + \varepsilon h + v),$$

where (ε, v) lives in some neighborhood of $(0, 0)$ in $\mathbb{R} \times X$. Since $F(0, 0) = 0$ and

$$F_v(0, 0)v = G'(u)v \quad \text{for all } v \in X,$$

it follows from the *surjective implicit function theorem* (Theorem 4.H) applied to F that there are numbers $\rho > 0$ and $r > 0$ such that, for each $\varepsilon \in \mathbb{R}$ and $|\varepsilon| \leq \rho$, there is a $v(\varepsilon) \in X$ with $\|v(\varepsilon)\| \leq r$ such that

$$G(u + \varepsilon h + v(\varepsilon)) = 0 \tag{99}$$

and

$$\|v(\varepsilon)\| \leq \text{const} \|F_v(0, 0)v(\varepsilon)\| \leq \text{const} \|G'(u)v(\varepsilon)\| \tag{100}$$

along with $\|v(\varepsilon)\| \rightarrow 0$ as $\varepsilon \rightarrow 0$. By the definition of the F -derivative,

$$G(u + k) = G'(u)k + o(\|k\|), \quad k \rightarrow 0.$$

Thus, according to (99),

$$G'(u)v(\varepsilon) + o(1)\|\varepsilon h + v(\varepsilon)\| = 0, \quad \varepsilon \rightarrow 0,$$

since $G'(u)h = 0$. By (100), $\|v(\varepsilon)\| \leq o(1)(\|\varepsilon h\| + \|v(\varepsilon)\|)$, that is,

$$\|v(\varepsilon)\| = o(\varepsilon), \quad \varepsilon \rightarrow 0.$$

Since u is a solution to (97), it follows from (99) that

$$f(u + \varepsilon h + v(\varepsilon)) \geq f(u).$$

This yields

$$f'(u)(\varepsilon h + v(\varepsilon)) + o(1)\|\varepsilon h + v(\varepsilon)\| \geq 0, \quad \varepsilon \rightarrow 0.$$

Dividing by ε and letting $\varepsilon \rightarrow \pm 0$, we get $f'(u)h \geq 0$ and $f'(u)h \leq 0$, which means that

$$f'(u)h = 0.$$

Step 2: It follows from Step 1 that $G'(u)h = 0$, $h \in X$, implies $f'(u)h = 0$, that is, we get the *key condition*

$$f'(u) \in N(G'(u))^\perp.$$

By the *closed range theorem* from Section 3.12, $N(G'(u))^\perp = R(G'(u)^\top)$, and hence

$$f'(u) \in R(G'(u)^\top).$$

Thus, there exists a functional $\lambda \in Y^*$ such that $f'(u) = G'(u)^\top \lambda$, namely,

$$\langle f'(u), h \rangle = \langle G'(u)^\top \lambda, h \rangle = \langle \lambda, G'(u)h \rangle \quad \text{for all } h \in X.$$

Hence $(f'(u) + \lambda G'(u))h = 0$ for all $h \in X$. This is (98). \square

Problems

In the following, let X , X_j , Y , and Z denote Banach spaces over \mathbb{K} . Furthermore, let $U(u)$ denote an open neighborhood of the point u . Recall that

$$A^{(n)}(u)h_1 \cdots h_n \equiv d^n A(u)h_1 \cdots h_n.$$

Let $-\infty < a < b < \infty$.

4.1. Linear operators. Let $A: X \rightarrow Y$ be a linear continuous operator. Show that

$$A' = A \quad \text{and} \quad A^{(n)} = 0 \quad \text{for all } n \geq 2.$$

4.2. Composition. We are given the operator $A: U(u) \subseteq X \rightarrow Y$ and the two linear continuous operators $L: Y \rightarrow Z$ and $M: Z \rightarrow X$. Let $n \geq 1$.

Show that if the F -derivative $A^{(n)}(u)$ exists, then the two F -derivatives $(L \circ A)^{(n)}(u)$ and $(A \circ M)^{(n)}(u)$ exist, where

$$(L \circ A)^{(n)}(u) = L \circ A^{(n)}(u)$$

and

$$(A \circ M)^{(n)}(u)(h_1, \dots, h_n) = A^{(n)}(Mu)(Mh_1, \dots, Mh_n),$$

for all $h_j \in X$, $j = 1, \dots, n$.

4.3. The superposition operator. Set $X := C[a, b]$. Define

$$(Au)(x) := f(u(x)) \quad \text{for all } x \in [a, b].$$

Suppose that $f: \mathbb{R} \rightarrow \mathbb{R}$ is C^n , $n \geq 1$. Differentiating $A(u + th)(x) = f(u(x) + th(x))$, $t \in \mathbb{R}$, at the point $t = 0$, we formally obtain

$$[A'(u)h](x) = f'(u(x))h(x) \quad \text{for all } x \in [a, b]. \quad (101)$$

Show that the operator $A: X \rightarrow X$ is C^n and $A'(u)$ is given by (101) for all $u, h \in X$. Moreover, show that

$$[A^{(n)}(u)h_1 h_2 \cdots h_n](x) = f^{(n)}(u(x))h_1(x)h_2(x) \cdots h_n(x)$$

for all $u, h_1, \dots, h_n \in X$ and all $x \in [a, b]$.

Hint: Use the classic Taylor theorem and an induction argument.

4.4. Nonlinear integral operator. Set $X := C[a, b]$. Define

$$(Au)(x) := \int_a^b \mathcal{A}(x, y)f(u(y))dy \quad \text{for all } x \in [a, b].$$

Suppose that $\mathcal{A}: [a, b] \times [a, b] \rightarrow \mathbb{R}$ is continuous and that $f: \mathbb{R} \rightarrow \mathbb{R}$ is C^n , $n \geq 1$. Differentiating

$$A(u + th)(x) = \int_a^b \mathcal{A}(x, y) f(u(y) + th(y)) dy, \quad t \in \mathbb{R},$$

at the point $t = 0$, we formally obtain

$$[A'(u)h](x) = \int_a^b \mathcal{A}(x, y) f'(u(y)) h(y) dy \quad \text{for all } x \in [a, b]. \quad (102)$$

Show that the operator $A: X \rightarrow X$ is C^n and $A'(u)$ is given by (102) for all $u, h \in X$. Moreover, show that

$$[A^{(n)}(u)h_1 h_2 \cdots h_n](x) = \int_a^b \mathcal{A}(x, y) f^{(n)}(u(y)) h_1(y) h_2(y) \cdots h_n(y) dy$$

for all $u, h_1, \dots, h_n \in X$ and all $x \in [a, b]$.

4.5. More general superposition operators. Set $X := C[a, b]$. Define

$$(Au)(x) := g(x, u_1(x), \dots, u_m(x)) \quad \text{for all } x \in [a, b],$$

where $u = (u_1, \dots, u_m)$, $m \geq 1$. Suppose that $g: [a, b] \times \mathbb{R}^m \rightarrow \mathbb{R}$ is C^n , $n \geq 1$.

Show that the operator $A: X \times \cdots \times X \rightarrow X$ is C^n and

$$[A'(u)h](x) = \sum_{j=1}^m g_{u_j}(x, u(x)) h_j(x) \quad \text{for all } x \in [a, b],$$

and for all $u, h \in X \times \cdots \times X$, where $h = (h_1, \dots, h_m)$.

4.6. Nonlinear differential operator. Set $X := C^2[a, b]$ and $Y := C[a, b]$. Define

$$(Au)(x) := u''(x) + g(x, u(x), u'(x)) \quad \text{for all } x \in [a, b].$$

Suppose that $g: [a, b] \times \mathbb{R}^2 \rightarrow \mathbb{R}$ is C^n , $n \geq 1$. Differentiating $A(u + th)(x)$ at the point $t = 0$, we formally obtain

$$\begin{aligned} [A'(u)h](x) &= h''(x) + g_u(x, u(x), u'(x)) h(x) \\ &\quad + g_{u'}(x, u(x), u'(x)) h'(x) \quad \text{for all } x \in [a, b]. \end{aligned} \quad (103)$$

Show that the operator $A: X \rightarrow Y$ is C^n and that $A'(u)$ is given by (103) for all $u, h \in X$. Compute $A^{(n)}(u)h_1 \cdots h_n$ for $n = 2, 3$.

4.7. Generalizations. Formulate and prove analogous results for the operators related to

$$\int_a^b g(x, y, u(y)) dy \quad \text{and} \quad g(x, u'(x), \dots, u^{(m)}(x)).$$

4.8. Nonlinear systems of real equations. Let $X := \mathbb{R}^m$ and $Y := \mathbb{R}^k$, where $k, m \geq 1$. Define the operator $A: X \rightarrow Y$ by $v = Au$ and

$$v_j = f_j(u_1, \dots, u_m), \quad j = 1, \dots, k. \quad (104)$$

Suppose that $f_j: \mathbb{R}^m \rightarrow \mathbb{R}$, $j = 1, \dots, k$, is C^n with $n \geq 1$.

Show that $A: X \rightarrow Y$ is C^n and

$$A^{(n)}(u)h_1 \cdots h_n = (d^n f_1(u)h_1 \cdots h_n, \dots, d^n f_m(u)h_1 \cdots h_n),$$

where $d^n f_j$ has been computed in Example 4 of Section 4.2.

In particular, the equation $v = A'(u)w$ corresponds to the linearization of (104), namely,

$$v_j = \sum_{s=1}^m \partial_s f_j(u_1, \dots, u_m) w_s, \quad j = 1, \dots, k,$$

where $\partial_s := \partial/\partial u_s$.

Formulate the implicit function theorem, the inverse mapping theorem, and the rank theorem (Theorem 4.G) in terms of nonlinear systems of real equations.

Hint: Cf. Zeidler (1986), Vol. 1, Section 4.8 and Problem 4.4b.

4.9. The Gâteaux derivative (G -derivative). Let $A: U(u_0) \subseteq X \rightarrow Y$. By definition, the operator A is G -differentiable at u_0 iff there exists a linear bounded operator $L: X \rightarrow Y$ such that

$$A(u_0 + th) - A(u_0) = tLh + o(t), \quad t \rightarrow 0,$$

for all $h \in X$ with $\|h\| \leq 1$ and all real numbers t in some neighborhood of zero. We call the operator $A'_G(u_0) := L$ the G -derivative of A at u_0 .

Show that if $A'_G(u)$ exists on some open neighborhood \mathcal{U} of u_0 and the map $u \mapsto A'_G(u)$ from \mathcal{U} to $L(X, Y)$ is continuous at u_0 , then the F -derivative $A'(u_0)$ exists and $A'(u_0) = A'_G(u_0)$.

Hint: Cf. Zeidler (1986), Vol. 1, Section 4.2.

4.10. Symmetry of higher derivatives. Suppose that $A: U \subseteq X \rightarrow Y$ is C^n , $n \geq 2$, on the open set U . Then, for each $u \in U$, the map

$$(h_1, \dots, h_n) \mapsto A^{(n)}(u)h_1 \cdots h_n$$

is symmetric on $X \times \cdots \times X$, that is, $A^{(n)}(u)h_1 \cdots h_n$ remains invariant under any permutation of h_1, \dots, h_n .

Solution: For example, let $n = 2$. For fixed $h, k \in X$ and fixed functional $y^* \in Y^*$, introduce the *real function*

$$\phi(t, s) := \langle y^*, A(u + th + sk) \rangle, \quad t, s \in \mathbb{R},$$

where $|t|$ and $|s|$ are sufficiently small. Then

$$\begin{aligned}\phi_t(t, s) &= \langle y^*, A'(u + th + sk)h \rangle, \\ \phi_{st}(0, 0) &= \langle y^*, A''(u)kh \rangle, \quad \phi_{ts}(0, 0) = \langle y^*, A''(u)hk \rangle.\end{aligned}$$

Since ϕ is C^2 in some neighborhood of $(0, 0)$, we obtain $\phi_{ts}(0, 0) = \phi_{st}(0, 0)$ by a well-known classical result. Hence

$$A''(u)hk = A''(u)kh \quad \text{for all } h, k \in X, u \in U,$$

since $y^* \in Y^*$ is arbitrary.

4.11. Partial F -derivatives. Let the operator

$$A: U \subseteq X_1 \times \cdots \times X_m \rightarrow Y$$

be given, where U is open and $m \geq 1$. Let $n \geq 1$.

Show that A is C^n on U iff the partial F -derivative $D_j A$ is C^{n-1} on U for all $j = 1, \dots, m$. In addition, we have

$$A'(u)h = \sum_{j=1}^m D_j A(u)h_j \quad \text{for all } u \in U, h \in X_1 \times \cdots \times X_m, \quad (105)$$

where $h = (h_1, \dots, h_m)$.

Solution: Let $m = 2$. The general case proceeds analogously.

Suppose first that A is C^n on U . Then recall that (105) has been proved in Section 4.2. Since

$$D_1 A(u)h = A'(u)(h, 0) \quad \text{and} \quad D_2 A(u)h = A'(u)(0, h),$$

$D_1 A$ and $D_2 A$ are C^{n-1} on U .

Conversely, suppose that $D_1 A$ and $D_2 A$ are C^{n-1} on U . We first prove (105). Set $u := (v, w)$ and $h := (\alpha, \beta)$, where $v, \alpha \in X_1$ and $w, \beta \in X_2$. Also set $A_v := D_1 A$ and $A_w := D_2 A$. By the triangle inequality,

$$\begin{aligned}&\|A(v + \alpha, w + \beta) - A(v, w) - A_v(v, w)\alpha - A_w(v, w)\beta\| \\ &\leq \|A(v + \alpha, w + \beta) - A(v, w + \beta) - A_v(v, w + \beta)\alpha\| \\ &\quad + \|A_v(v, w + \beta) - A_v(v, w)\alpha\| \\ &\quad + \|A(v, w + \beta) - A(v, w) - A_w(v, w)\beta\|.\end{aligned}$$

It follows from the *Taylor theorem* in Section 4.5 that the right side can be bounded by

$$\begin{aligned}&\sup_{0 \leq \tau \leq 1} \|A_v(v + \tau\alpha, w + \beta) - A_v(v, w + \beta)\| \|\alpha\| \\ &\quad + \|A_v(v, w + \beta) - A_v(v, w)\| \|\alpha\| \\ &\quad + \sup_{0 \leq \tau \leq 1} \|A_v(v, w + \tau\beta) - A_w(v, w)\| \|\beta\|.\end{aligned}$$

By the *continuity* of A_v and A_w at the point (v, w) , this last expression can be bounded by $r(\alpha, \beta)(\|\alpha\| + \|\beta\|)$ with $r(\alpha, \beta) \rightarrow 0$ as $(\alpha, \beta) \rightarrow 0$. This proves (105).

Set $P_j(h_1, h_2) := h_j$, $j = 1, 2$. Then (105) reads as

$$A'(u) = D_1 A(u) \circ P_1 + D_2 A(u) \circ P_2.$$

The operator $P_j: X_1 \times X_2 \rightarrow X_j$ is linear and continuous. Since $D_j A$ is C^{n-1} on U , the operator A is C^n on U , by Problem 4.2.

4.12.* The global inverse mapping theorem. Let $f: X \rightarrow Y$ be a C^1 -map, where X and Y are Banach spaces over \mathbb{K} . Then the following two conditions are equivalent:

- (i) $f: X \rightarrow Y$ is a C^1 -diffeomorphism.
- (ii) $f'(u): X \rightarrow Y$ is bijective for all $u \in X$, and f is *proper* (i.e., the compactness of C in Y implies the compactness of $f^{-1}(C)$).

Study the proof in Berger (1977), p. 221.

4.13. A simple proof for a variant of the global inverse mapping theorem via the continuation method. Let $f: X \rightarrow Y$ be a proper C^r -map, $1 \leq r \leq \infty$, where X and Y are Banach spaces over \mathbb{K} . Suppose that

$$f'(u): X \rightarrow Y \text{ is bijective for all } u \in X.$$

Show that

- (a) $f: X \rightarrow Y$ is surjective.
- (b) If $f: X \rightarrow Y$ is injective, then f is a C^r -diffeomorphism.

Hint: For $t \in \mathbb{R}$, use the equation

$$f(u(t)) = tv. \quad (106)$$

Solution: Ad (a): Let $v \in Y$. Without loss of generality, we may assume that $f(0) = 0$. By the local inverse mapping theorem (Theorem 4.F), there exists a number $t_0 > 0$ such that equation (106) has a solution $u(t)$ for each $t \in [0, t_0]$. Let T be the supremum of all the numbers t_0 . If $T = \infty$, then equation (106) has a solution for $t = 1$ and we are done.

We show that $T = \infty$. On the contrary, suppose that $T < \infty$. Let (t_n) be a sequence such that $t_n \rightarrow T$ as $n \rightarrow \infty$ and $t_n < T$ for all n . By (106),

$$f(u(t_n)) = t_n v \quad \text{for all } n.$$

Since f is *proper*, there exists a subsequence, again denoted by $(u(t_n))$, such that

$$u(t_n) \rightarrow w \quad \text{as } n \rightarrow \infty$$

for some w . Hence $f(w) = Tv$. By the local inverse mapping theorem (Theorem 4.F), f is a local C^1 -diffeomorphism at the point w . Hence equation (106) has a unique solution for all t in an open neighborhood of T in \mathbb{R} . This contradicts the maximality of T .

Ad (b). Since $f: X \rightarrow Y$ is *bijective*, the inverse map $f^{-1}: Y \rightarrow X$ exists. The local inverse mapping theorem (Theorem 4.F) tells us that f^{-1} is C^r . Hence $f: X \rightarrow Y$ is a C^r -diffeomorphism. \square

4.14. *Algebraic operations for multilinear forms.* Let X be a linear space over \mathbb{K} . For $n \geq 1$, denote the set of all n -linear forms

$$A: X \times \cdots \times X \rightarrow \mathbb{K}$$

by $\mathcal{M}^n(X)$. Set $\mathcal{M}^0(X) := \mathbb{K}$.

4.14a. *The tensor algebra of multilinear forms.* Let $m, n \geq 1$, and let $A \in \mathcal{M}^m(X)$ and $B \in \mathcal{M}^n(X)$. Define the tensor product $A \otimes B$ through

$$(A \otimes B)(u_1, \dots, u_{m+n}) := A(u_1, \dots, u_m)B(u_{m+1}, \dots, u_{m+n})$$

for all $u_1, \dots, u_{m+n} \in X$. If $\alpha, \beta \in \mathbb{K}$ and $A \in \mathcal{M}^n(X)$, $n \geq 1$, then define

$$\alpha \otimes A = A \otimes \alpha := \alpha A \quad \text{and} \quad \alpha \otimes \beta := \alpha\beta.$$

Let $m, n, r \geq 0$. Show that for all $A \in \mathcal{M}^m(X)$, $B \in \mathcal{M}^n(X)$, and $C \in \mathcal{M}^r(X)$, the following conditions are met:

- (i) $A \otimes B \in \mathcal{M}^{m+n}(X)$;
- (ii) $A \otimes (B \otimes C) = (A \otimes B) \otimes C$;
- (iii) $A \otimes (B + C) = (A \otimes B) + (A \otimes C)$ and $(B + C) \otimes A = (B \otimes A) + (C \otimes A)$.

Naturally enough, suppose that $n = r$ in (iii).

Definition. Let $\otimes \mathcal{M}(X)$ denote the set of all finite sums of multilinear forms over X , that is,

$$A_0 + A_1 + \cdots + A_k,$$

where $A_m \in \mathcal{M}^m(X)$ for all m . Then, $\otimes \mathcal{M}(X)$ is a linear space over \mathbb{K} . Moreover, $\otimes \mathcal{M}(X)$ becomes an algebra with respect to the operations “+” and “ \otimes ”.

4.14b. *The Grassmann algebra of antisymmetric multilinear forms.* Let $m \geq 1$. The set of all antisymmetric m -linear forms $A \in \mathcal{M}^m(X)$ is denoted by $\mathcal{A}^m(X)$. Thus, $A \in \mathcal{A}^m(X)$ means that

$$A(u_{\pi(1)}, \dots, u_{\pi(m+n)}) = (\operatorname{sgn} \pi)A(u_1, \dots, u_m),$$

for all u_1, \dots, u_m , where π is a permutation of $1, \dots, m$, and $\operatorname{sgn} \pi$ denotes the sign of π .

Let $m, n \geq 1$. For $A \in \mathcal{A}^m(X)$ and $B \in \mathcal{A}^n(X)$, define the *exterior product* $A \wedge B$ through

$$(A \wedge B)(u_1, \dots, u_{m+n}) := \sum_{\pi} (\text{sgn } \pi) A(u_{\pi(1)}, \dots, u_{\pi(m)}) B(u_{\pi(m+1)}, \dots, u_{\pi(m+n)})$$

for all $u_1, \dots, u_{m+n} \in X$. Here, we sum over all permutations π of $1, \dots, m+n$ that have the following additional property:

$$\pi(1) < \dots < \pi(m) \quad \text{and} \quad \pi(m+1) < \dots < \pi(m+n).$$

For $\alpha, \beta \in \mathbb{K}$ and $A \in \mathcal{A}^n(X)$, $n \geq 1$, define

$$\alpha \wedge A = A \wedge \alpha := \alpha A \quad \text{and} \quad \alpha \wedge \beta := \alpha \beta.$$

Let $m, n, r \geq 0$. Show that for all $A \in \mathcal{A}^m(X)$, $B \in \mathcal{A}^n(X)$, and $C \in \mathcal{A}^r(X)$, the following conditions are met:

- (i) $A \wedge B \in \mathcal{A}^{m+n}(X)$;
- (ii) $A \wedge (B \wedge C) = (A \wedge B) \wedge C$;
- (iii) $A \wedge B = (-1)^{mn} B \wedge A$ (supercommutativity);
- (iv) $A \wedge (B+C) = (A \wedge B) + (A \wedge C)$ and $(B+C) \wedge A = (B \wedge A) + (C \wedge A)$.

Naturally enough, suppose that $n = r$ in (iv).

Definition. Let $\wedge \mathcal{A}(X)$ denote the set of all finite sums of antisymmetric multilinear forms over X , that is,

$$A_0 + A_1 + \dots + A_k,$$

where $A \in \mathcal{A}^m(X)$ for all m . Then, $\wedge \mathcal{A}(X)$ is a linear space over \mathbb{K} . Moreover, $\wedge \mathcal{A}(X)$ becomes an algebra with respect to the operations “+” and “ \wedge ”.

4.14c. The relation to determinants. Show that if $a, b, c \in X^T$, then

$$(a \wedge b)(u, v) = \begin{vmatrix} a(u) & a(v) \\ b(u) & b(v) \end{vmatrix}$$

and

$$((a \wedge b) \wedge c)(u, v, w) = \begin{vmatrix} a(u) & a(v) & a(w) \\ b(u) & b(v) & b(w) \\ c(u) & c(v) & c(w) \end{vmatrix}$$

for all $u, v, w \in X$. Generally, if $a_1, \dots, a_n \in X$, then

$$(a_1 \wedge \dots \wedge a_n)(u_1, \dots, u_n) = \det(a_j(u_k))$$

for all $u_1, \dots, u_n \in X$.

4.14d. *The Grassmann algebra $\wedge(X)$ of a linear space X .* Let X be a linear space over \mathbb{K} . We are given $u \in X$. Define

$$u(u^*) := u^*(u) \quad \text{for all } u^* \in X^T.$$

This way, we regard u as an element of $\mathcal{A}^1(X^T)$. By definition, the Grassmann algebra $\wedge(X)$ of the linear space X consists of all finite sums of all possible finite \wedge -products, that is,

$$\alpha + u_1 + (u_2 \wedge u_3) + (u_4 \wedge u_5 \wedge u_6) + \cdots$$

for all $u_j \in X$. Obviously, $\wedge(X)$ is a linear space over \mathbb{K} , and $\wedge(X)$ is an algebra with respect to the operations “+” and “ \wedge ”.

5

Fredholm Operators

It came as a complete surprise, when, in a short note published in 1900, the Swedish mathematician Ivar Fredholm (1866–1927) showed that the general theory of all integral equations considered prior to him was, in fact, extremely simple.

Jean Dieudonné (1981)

The purpose of this note is to introduce a nonlinear version of Fredholm operators, and to prove that in this context Sard's theorem holds if zero measure is replaced by first category.

Steve Smale (1965)

Before you generalize, formalize, and axiomatize there must be mathematical substance.

Hermann Weyl (1885–1955)

Another characteristic of mathematical thought is that it can have no success where it cannot generalize.

Charles Sanders Peirce (1839–1914)

Let us first consider the *linear* operator equation

$$Au = b, \quad u \in X. \quad (1)$$

It is quite natural to look for a class of linear operators that have the following properties:

- (i) Equation (1) has a solution u iff a *finite* number of solvability conditions is satisfied for b .
- (ii) The general solution of (1) depends on a *finite* number of parameters.

The class of linear Fredholm operators satisfies conditions (i) and (ii). The *index* of a linear Fredholm operator A is defined through

$$\text{ind } A := \dim N(A) - \text{codim } R(A).$$

Large classes of linear differential and integral operators represent Fredholm operators in appropriate function spaces. In particular, if the operator A is Fredholm of *index zero*, then the following fundamental principle holds for equation (1):

Uniqueness implies existence.

For example, the Riesz–Schauder theory tells us that the operator

$$I + C: X \rightarrow X \quad (2)$$

is Fredholm of index zero on the Banach space X provided the linear operator $C: X \rightarrow X$ is *compact*. This generalizes the classic Fredholm theory for integral equations (cf. Section 5.3).

Now suppose that the operator A from equation (1) is *nonlinear*. Then it is quite natural to look for a class of nonlinear operators that possess the following property:

For “most” right-hand sides b , the equation $Au = b$, $u \in X$, has at most a finite number of solutions.

This leads us to the class of *nonlinear Fredholm operators* introduced by Smale in 1965. For example, each operator

$$B + C: X \rightarrow X$$

is a nonlinear Fredholm operator of index zero on the Banach space X provided

- (a) the operator $B: X \rightarrow X$ is linear, continuous, and *bijective*, and
- (b) the operator $C: X \rightarrow X$ is *compact* and C^1 .

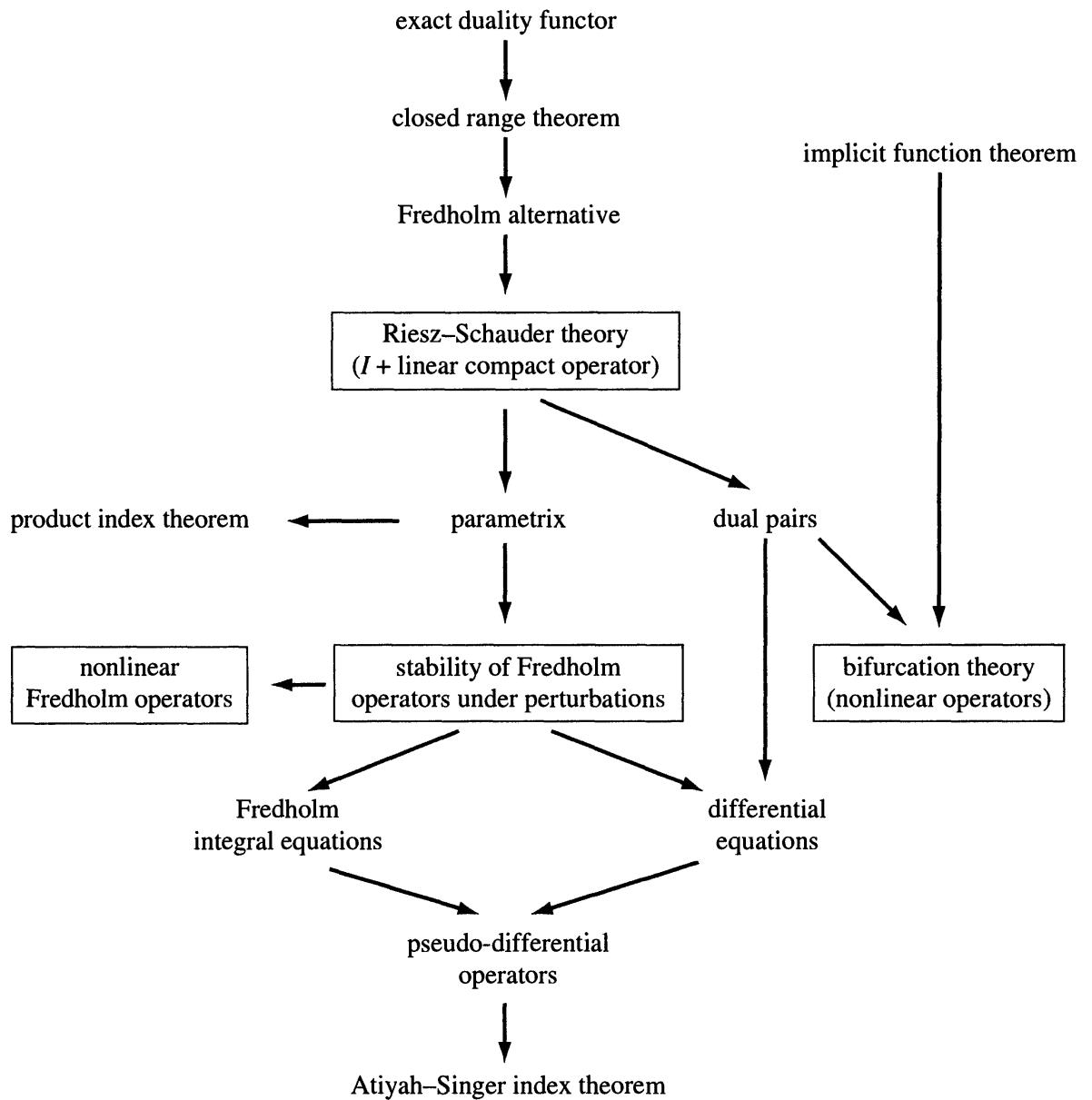


FIGURE 5.1.

The *index* of a linear Fredholm operator A plays a fundamental role, since it is *invariant* under appropriate perturbations of A (cf. Section 5.8). One of the most important achievements of twentieth-century mathematics is represented by the famous *Atiyah–Singer index theorem*.¹

Roughly speaking, this index theorem tells us that the index of elliptic-type differential operators and certain classes of pseudo-differential operators on compact manifolds depends only on the *topology* of the manifold. This way we obtain a deep relation between *analysis* and *topology* which has its roots in the ingenious work of Riemann in the middle of the nine-

¹As an introduction to the Atiyah–Singer index theorem we recommend Zeidler (1995), Gilkey (1984), Booss and Bleecker (1985), and Cycon et al. (1986) (supersymmetric approach).

teenth century. Pseudo-differential operators² generalize both differential and integral operators.

Figure 5.1 displays important interrelations. For the convenience of the reader, we start with the Riesz–Schauder theory on Hilbert spaces, which is based on a simple variant of the closed graph theorem.

5.1 Duality for Linear Compact Operators

Theorem 5.A. *Let X and Y be Banach spaces over \mathbb{K} .*

If the linear operator $A: X \rightarrow Y$ is compact, then so is the dual operator $A^T: Y^ \rightarrow X^*$.*

Schauder proved this theorem in 1930.

Proof. Let $f, f_j \in Y^*$. Then, for all $u, u_i \in X$, we get the following *key inequality*:

$$\begin{aligned} |f(Au) - f_j(Au)| &\leq |f(Au) - f(Au_i)| + |f(Au_i) - f_j(Au_i)| + |f_j(Au_i) - f_j(Au)| \\ &\leq \|f\| \|Au - Au_i\| + |f(Au_i) - f_j(Au_i)| + \|f_j\| \|Au - Au_i\|. \end{aligned} \quad (3)$$

In the following we will critically use the fact that a subset of a Banach space is relatively compact iff it has a *finite ε -net* for each $\varepsilon > 0$ (cf. Proposition 10 in Section 1.11 of AMS Vol. 108).

Let B^* be a bounded set in Y^* . We have to show that the set $A^T(B^*)$ is relatively compact.

To this end, fix $\varepsilon > 0$. Let B denote the closed unit ball in X . Since the operator A is compact, the set $A(B)$ is relatively compact, and hence $A(B)$ has a finite ε -net. That is, there are points $u_1, \dots, u_N \in B$ such that

$$\min_{1 \leq i \leq N} \|Au - Au_i\| \leq \varepsilon \quad \text{for all } u \in B.$$

Since the set B^* is bounded, we obtain from

$$|f(Au_i)| \leq \|f\| \|Au_i\|$$

that the set

$$F := \{(f(Au_1), \dots, f(Au_N)): f^* \in B^*\}$$

is bounded in the finite-dimensional Banach space \mathbb{K}^N , and hence F is relatively compact, that is, F has a finite ε -net. Thus, there exist points $f_1, \dots, f_M \in B^*$ such that

$$\min_{1 \leq j \leq M} |f(Au_i) - f_j(Au_i)| \leq \varepsilon \quad \text{for all } i.$$

²The modern theory of pseudo-differential operators can be found in Hörmander (1983).

It follows now from (3) that, for all $u \in B$ and $f \in B^*$,

$$|f(Au) - f_j(Au)| \leq (\|f\| + \|f_j\|) \|Au - Au_i\| + |f(Au_i) - f_j(Au_i)|,$$

and hence

$$\min_{1 \leq j \leq M} |f(Au) - f_j(Au)| \leq \text{const} \cdot \varepsilon + \varepsilon.$$

Finally, observe that $A^T(f - f_j) \in X^*$, and hence

$$\|A^T f - A^T f_j\| = \sup_{u \in B} |\langle A^T(f - f_j), u \rangle| = \sup_{u \in B} |f(Au) - f_j(Au)|,$$

by the definition of the dual operator A^T . This implies

$$\min_{1 \leq j \leq N} \|A^T f - A^T f_j\| \leq \text{const} \cdot \varepsilon + \varepsilon \quad \text{for all } f \in B^*. \quad (4)$$

Varying the number $\varepsilon > 0$, relation (4) tells us that, for each $\eta > 0$, the set $A^T(B^*)$ has a finite η -net, that is, $A^T(B^*)$ is relatively compact. \square

Corollary 1. *Let X be a Hilbert space over \mathbb{K} . If the linear operator $A: X \rightarrow X$ is compact, then so is the adjoint operator $A^*: X \rightarrow X$.*

Proof. The duality map $\mathcal{J}: X \rightarrow X^*$ is a homeomorphism with $\|\mathcal{J}u\| = \|u\|$ for all $u \in X$, and

$$A^* = \mathcal{J}^{-1} A^T \mathcal{J}.$$

Since A^T is compact, the compactness of A^* follows from Proposition 4 ahead. \square

Proposition 2 (Sums). *Let the operators*

$$A, B: X \rightarrow Y$$

be compact, where X and Y are normed spaces over \mathbb{K} .

Then the sum $A + B: X \rightarrow Y$ is also compact.

Proof. Let (u_n) be a bounded sequence in X . Since A is compact, there exists a subsequence $(u_{n'})$ such that $(Au_{n'})$ is convergent. Furthermore, since B is compact, there exists a subsequence $(u_{n''})$ of $(u_{n'})$ such that $(Bu_{n''})$ is convergent. Hence $(Au_{n''} + Bu_{n''})$ is convergent. \square

Definition 3. Let X and Y be normed spaces over \mathbb{K} . Then the operator $A: D(A) \subseteq X \rightarrow Y$ is called *bounded* if it maps bounded sets onto bounded sets.

For example, each linear continuous operator $A: X \rightarrow Y$ is bounded, since $\|Au\| \leq \|A\| \|u\|$ for all $u \in X$.

Proposition 4 (Products). *The operators*

$$BC \quad \text{and} \quad CE$$

are compact provided the following hold:

- (i) X, Y, V, W are normed spaces over \mathbb{K} .
- (ii) $C: X \rightarrow Y$ is compact.
- (iii) $E: V \rightarrow X$ is continuous and bounded, and
- (iv) $B: Y \rightarrow W$ is continuous.

Proof. Ad $CD: V \rightarrow Y$. If M is a bounded set in V , then so is $E(M)$, and hence the set $C(E(M))$ is relatively compact.

Ad $BC: X \rightarrow W$. If N is a bounded set in X , then $C(N)$ is relatively compact. Since $u_n \rightarrow u$ as $n \rightarrow \infty$ implies $Bu_n \rightarrow Bu$, the set $B(C(N))$ is also relatively compact. \square

Proposition 5 (Finite rank). *Let $A: X \rightarrow Y$ be a bounded continuous operator with*

$$\dim R(A) < \infty,$$

where X and Y are normed spaces over \mathbb{K} . Then A is compact.

Proof. If M is a bounded subset of X , then $A(M)$ is a bounded set in the finite-dimensional normed space $R(A)$. Hence $A(M)$ is relatively compact. \square

5.2 The Riesz–Schauder Theory on Hilbert Spaces

We consider the operator equation

$$Bu + Cu = b, \quad u \in X, \tag{5}$$

along with the dual equation

$$B^*v + C^*v = b^*, \quad v \in X. \tag{5*}$$

By definition, the homogeneous original equation and the homogeneous dual equation correspond to (5) and (5*) with $b = 0$ and $b^* = 0$, respectively.

Theorem 5.B. *Suppose that*

- (i) *The operator $B: X \rightarrow X$ is linear, continuous, and bijective on the Hilbert space X over \mathbb{K} (e.g., $B = I$).*
- (ii) *The operator $C: X \rightarrow X$ is linear and compact.*

Then the following properties are met:

- (a) Original problem. For given $b \in X$, equation (5) has a solution $u \in X$ iff b satisfies the solvability condition

$$(b | v) = 0$$

for all solutions v of the homogeneous dual equation (5*).

- (b) Finiteness. The homogeneous dual equation (5*) and the homogeneous original equation (5) have the same finite number of linearly independent solutions.

- (c) Well-posedness. If $Bu + Cu = 0$ implies $u = 0$, then the original equation (5) has a unique solution u for each given $b \in X$.

Moreover, the solution u depends continuously on b , that is, the inverse operator

$$(B + C)^{-1}: X \rightarrow X$$

is continuous.

- (d) Dual equation. For given $b^* \in X$, equation (5*) has a solution v iff b^* satisfies the solvability condition

$$(b^* | u) = 0$$

for all solutions u of the homogeneous original equation (5).

In terms of Section 5.4 ahead, this theorem tells us that the operator $B + C$ is Fredholm of index zero. Statement (c) says that

Uniqueness implies existence.

Let u_1, \dots, u_n and v_1, \dots, v_n be a maximal number of linearly independent solutions of the homogeneous original equation (5) and the homogeneous dual equation (5*), respectively. Then, for given $b \in X$, the original problem (5) has a solution u iff

$$(b | v_j) = 0 \quad \text{for } j = 1, \dots, n,$$

and the general solution of (5) is given by

$$u = u_0 + \sum_{j=1}^n \alpha_j u_j,$$

where u_0 is a special solution of (5), and $\alpha_1, \dots, \alpha_n$ are arbitrary numbers from \mathbb{K} .

Moreover, we will show in Section 5.5 that this theorem remains true for Banach spaces X provided we *replace* the adjoint operators

$$B^*, C^*: X \rightarrow X \quad \text{and} \quad \text{the inner product } (\cdot | \cdot)$$

with the dual operators

$$B^T, C^T: X^* \rightarrow X^* \quad \text{and} \quad \text{the symbol } \langle \cdot, \cdot \rangle,$$

respectively. Recall that $\langle b^*, u \rangle = b^*(u)$ for all $b^* \in X^*$ and $u \in X$.

Theorem 5.B represents a *Fredholm alternative*. Such an alternative for *integral equations* was first proved by Fredholm in 1900. The generalization to Banach spaces dates back to papers by Riesz in 1918 and Schauder in 1930.

Proof. We set

$$S := B + C, \quad N := N(S), \quad R := R(S).$$

Obviously, the null space N of the operator S is closed, since S is continuous. We shall show ahead that the range R of S is closed. Therefore, by Section 2.9 in AMS Vol. 108, we have the orthogonal direct sums

$$X = N \oplus N^\perp = R \oplus R^\perp.$$

We also introduce the index of S by

$$\text{ind } S := \dim N - \text{codim } R.$$

Step 1: We show that $\dim N < \infty$. Let (u_n) be a bounded sequence in N (i.e., $Bu_n + Cu_n = 0$ for all n). Since C is compact, there is a subsequence, again denoted by (u_n) , such that $Cu_n \rightarrow w$ as $n \rightarrow \infty$. Hence $u_n \rightarrow v$ as $n \rightarrow \infty$, where $v := -B^{-1}w$. This implies $Bv + Cv = 0$ (i.e., $v \in N$).

Thus, the closed unit ball in the *closed* subspace N of X is compact. By Section 2.3, this implies $\dim N < \infty$.

Step 2: We show that $\dim N(S^*) < \infty$. Note that $S^* = B^* + C^*$. Since C is compact, so is C^* by Corollary 1 in Section 5.1. Furthermore, $B^{*-1} = (B^{-1})^*$, by the proof of Proposition 13 in Section 5.2 in AMS Vol. 108. The same argument as in Step 1 yields $\dim N(S^*) < \infty$.

Step 3: We show that the range $R(S)$ is closed. By Theorem 3.E(iii) in Section 3.12, it is sufficient to prove that

$$c \cdot \text{dist}(u, N) \leq \|Su\| \quad \text{for all } u \in X \text{ and fixed } c > 0.$$

If this is not true, then there exists a sequence (u_n) such that

$$Su_n = Bu_n + Cu_n \rightarrow 0 \quad \text{as } n \rightarrow \infty \tag{6}$$

and $\text{dist}(u_n, N) = 1$ for all n . As in Step 1, there exists a subsequence, again denoted by (u_n) , such that $u_n \rightarrow v$ as $n \rightarrow \infty$ and $v \in N$, contradicting $\text{dist}(u_n, N) = 1$ for all n .

Step 4: Special case of the *closed graph theorem*. We show that

$$R = N(S^*)^\perp, \quad (7)$$

where \perp denotes the orthogonal complement. In fact, it follows from

$$(Su \mid v) = (u \mid S^*v) \quad \text{for all } u, v \in X$$

that $N(S^*) \subseteq R^\perp$ and $R^\perp \subseteq N(S^*)$. Hence $R^\perp = N(S^*)$, that is,

$$X = R \oplus N(S^*),$$

since R is closed. This yields (7).

It follows from (7) that

$$\text{codim } R = \dim N(B^*) < \infty.$$

Step 5: The fundamental *stability* of the index against small perturbations. We show that if the operator $T: X \rightarrow X$ is linear and continuous, and if $\|T - S\|$ is sufficiently small, then

$$\text{ind } S = \text{ind } T.$$

The *decisive trick* consists in constructing the operator

$$\tilde{T}(u, v) := Tu + v \quad \text{for all } u \in N^\perp \text{ and } v \in R^\perp.$$

Then the operator $\tilde{T}: N^\perp \times R^\perp \rightarrow X$ is linear and continuous.

Moreover, if $T = S$, then the operator \tilde{S} is *bijective*. This is the *key observation*. In fact, the operator \tilde{S} is surjective, by construction.³ Furthermore, \tilde{S} is injective. To see this, assume $\tilde{S}(u, v) = 0$. Then,

$$Su + v = 0, \quad u \in N^\perp, \quad v \in R^\perp,$$

and hence $v = 0$, $Su = 0$. From $Su = 0$ and $u \in N^\perp$ we get $u = 0$.

If $\|\tilde{T} - \tilde{S}\|$ is sufficiently small, then the operator \tilde{T} is also bijective, by Proposition 7 in Section 1.23 of AMS Vol. 108. Thus, if $\|T - S\|$ is sufficiently small, then \tilde{T} is *bijective*, that is, we obtain the direct sum⁴

$$X = T(N^\perp) \oplus R^\perp, \quad (8)$$

³Observe that $S(N^\perp) = R(S) = R$.

⁴Note that

$$Tu_1 + v_1 = Tu_2 + v_2, \quad u_1, u_2 \in N^\perp, \quad v_1, v_2 \in R^\perp,$$

implies $u_1 = u_2$ and $v_1 = v_2$.

by definition of \tilde{T} . Moreover, we get

$$N(T) \subseteq N. \quad (9)$$

In fact, $Tu = 0$ along with $u \in N^\perp$ implies $\tilde{T}(u, 0) = 0$, and hence $u = 0$. Therefore, $N(T) \subseteq N$. By (9),

$$\dim N(T) \leq \dim N < \infty.$$

The proof follows now from the two relations (8) and (9) in a simple manner. By (8),

$$\text{codim } T(N^\perp) = \dim R^\perp = \text{codim } R. \quad (10)$$

Let us choose a linear subspace M of N such that we get the direct sum

$$N = N(T) \oplus M. \quad (11)$$

Then, $X = (N(T) \oplus M) \oplus N^\perp = N(T) \oplus (M \oplus N^\perp)$. Hence T is *injective* on $M \oplus N^\perp$. This implies

$$R(T) = T(M) \oplus T(N^\perp) \quad \text{and} \quad \dim T(M) = \dim M,$$

that is,

$$X = (T(M) \oplus T(N^\perp)) \oplus R(T)^\perp. \quad (12)$$

From this we immediately obtain the following:

$$\begin{aligned} \text{codim } R &= \text{codim } T(N^\perp) = \text{codim } R(T) + \dim M \quad (\text{by (10) and (12)}), \\ \dim N &= \dim N(T) + \dim M \quad (\text{by (11)}). \end{aligned}$$

Hence $\text{ind } S = \dim N - \text{codim } R = \dim N(T) - \text{codim } R(T) = \text{ind } T$.

Step 6: We show that

$$\text{ind } (B + C) = 0.$$

In fact, the function $t \mapsto B + tC$ is continuous from $[0, 1]$ to $L(X, X)$. By Step 5, $\text{ind}(B + tC) = \text{const}$ for all $t \in [0, 1]$. Since the operator B is bijective, we get $\dim N(B) = 0$ and $\text{codim } R(B) = 0$, and hence

$$\text{ind } B = \dim N(B) - \text{codim } R(B) = 0.$$

Ad (a). This follows from $R = N(S^*)^\perp$ in Step 4.

Ad (b). By Step 4, $\dim N(S^*) = \text{codim } R$. Hence

$$\dim N(B^*) = \dim N - \text{ind}(B + C) = \dim N < \infty.$$

Ad (c). If $Bu + Cu = 0$ implies $u = 0$, then $B + C$ is injective. Moreover, $\dim N = 0$ implies $\dim N(B^*) = 0$, and hence $\text{codim } R = 0$ (i.e., $B + C$ is

surjective). By the *continuous inverse theorem* from Section 3.5, the inverse operator $(B + C)^{-1}: X \rightarrow X$ is continuous.

Ad (d). Observe that $(S^*)^* = S$. Therefore, the dual equation to (5*) is given by (5). Consequently, statement (d) follows immediately from (a). \square

This proof has been chosen in such a way that it can be generalized immediately to more general situations as appear ahead (the Riesz–Schauder theory on Banach spaces and the perturbation theory for Fredholm operators.)

5.3 Applications to Integral Equations

Parallel to Section 4.4 in AMS Vol. 108, let us consider the integral equation

$$\int_a^b \mathcal{A}(x, y)u(y)dy - \lambda u(x) = h(x), \quad a \leq x \leq b, \quad (13)$$

along with the dual integral equation

$$\int_a^b \mathcal{A}(y, x)v(y)dy - \lambda v(x) = 0, \quad a \leq x \leq b, \quad (13^*)$$

where $-\infty < a < b < \infty$. In contrast to Proposition 4 in Section 4.4 in AMS Vol. 108 we do *not* assume that the kernel \mathcal{A} is symmetric. The real number λ is called an *eigenvalue* of (13) iff the homogeneous equation (13) with $h \equiv 0$ has a nontrivial solution $u \not\equiv 0$ on $[a, b]$.

Proposition 1. *Assume that the function $\mathcal{A}: [a, b] \times [a, b] \rightarrow \mathbb{R}$ is continuous. Let the function $h \in L_2(a, b)$ and the real number $\lambda \neq 0$ be given. Then the following statements hold true:*

- (i) *If λ is not an eigenvalue of (13), then (13) has a unique solution $u \in L_2(a, b)$.*
- (ii) *If λ is an eigenvalue of (13), then λ is also an eigenvalue of (13*) with the same multiplicity, and (13) has a solution $u \in L_2(a, b)$ iff*

$$\int_a^b h(x)v(x)dx = 0$$

for all eigensolutions v of (13).*

Corollary 2. *The eigensolutions of (13) and (13*) are continuous. If h is continuous on $[a, b]$, then so is each solution u of (13).*

Proof. Let $X = L_2(a, b)$, and define the operator C through

$$(Cu)(x) := \int_a^b \mathcal{A}(x, y)u(y)dy \quad \text{for all } x \in [a, b].$$

By Lemma 3 in Section 4.4 of AMS Vol. 108, the operator $C: X \rightarrow X$ is linear and compact.

The adjoint operator $C^*: X \rightarrow X$ is given through

$$(C^*v)(x) = \int_a^b \mathcal{A}(y, x)v(y)dy \quad \text{for all } x \in [a, b].$$

In fact, it follows from the *Tonnelli* theorem (cf. “Iterated Integration” in the appendix to AMS Vol. 108) that, for all $u, v \in X$,

$$\begin{aligned} (Cu | v) &= \int_a^b \left(\int_a^b \mathcal{A}(x, y)u(y)dy \right) v(x)dx \\ &= \int_a^b \left(\int_a^b \mathcal{A}(x, y)v(x)dx \right) u(y)dy = (u | C^*v). \end{aligned}$$

Now use Theorem 5.B with $B := -\lambda I$. □

Corollary 2 follows from the continuity of parameter integrals (cf. the appendix to AMS Vol. 108). In fact, if $u \in L_2(a, b)$, then the function g defined by

$$g(x) := \int_a^b \mathcal{A}(x, y)u(y)dy$$

is continuous on $[a, b]$.

5.4 Linear Fredholm Operators

Definition 1. Let X and Y be normed spaces over \mathbb{K} . By a *linear Fredholm operator*

$$A: X \rightarrow Y$$

we understand a linear continuous operator with

$$\dim N(A) < \infty \quad \text{and} \quad \operatorname{codim} R(A) < \infty.$$

The *index* of A is defined to be the integer

$$\operatorname{ind} A := \dim A - \operatorname{codim} R(A).$$

Example 2 (Finite-dimensional operators). Let X and Y be finite-dimensional normed spaces over \mathbb{K} . Then each linear operator $A: X \rightarrow Y$ is Fredholm and

$$\operatorname{ind} A = \dim X - \dim Y.$$

Proof. By (45*) in Chapter 3,

$$\operatorname{codim} N(A) = \dim R(A).$$

Hence

$$\begin{aligned}\operatorname{ind} A &= \dim N(A) - \operatorname{codim} R(A) \\ &= (\dim X - \operatorname{codim} N(A)) - (\dim Y - \dim R(A)) \\ &= \dim X - \dim Y.\end{aligned}\quad \square$$

Example 3 (Differential operator). Let $X = C^1[a, b]$ and $Y = C[a, b]$, where $-\infty < a < b < \infty$. Set

$$(Au)(x) := u'(x) \quad \text{for all } x \in [a, b].$$

Then the linear operator $A: X \rightarrow Y$ is Fredholm with $\operatorname{ind} A = 1$.

Proof. For each given $f \in C[a, b]$, the equation

$$u' = f \quad \text{on } [a, b] \tag{14}$$

has a solution $u \in C^1[a, b]$ given through

$$u(x) = \int_a^x f(t)dt.$$

Hence $R(A) = Y$, showing that $\operatorname{codim} R(A) = 0$.

Furthermore, it follows from (3) with $f \equiv 0$ that $u = \text{const}$, and hence $\dim N(A) = 1$. \square

Example 4 (Integral operator). Let $\lambda \in \mathbb{R}$ with $\lambda \neq 0$. Set

$$(Au)(x) := \int_a^b \mathcal{A}(x, y)u(y)dy - \lambda u(x),$$

for all $x \in [a, b]$, where the function $\mathcal{A}: [a, b] \times [a, b] \rightarrow \mathbb{R}$ is continuous, $-\infty < a < b < \infty$.

Then the operator $A: X \rightarrow X$ is Fredholm of index zero⁵ provided we set $X := L_2(a, b)$.

⁵We shall show in Section 5.11 that this remains true for $X = C[a, b]$.

This follows from Section 5.3. □

Proposition 5 (The role of the index). *Let $A: X \rightarrow Y$ be a linear Fredholm operator, where X and Y are normed spaces over \mathbb{K} . Then*

- (i) *A is surjective iff $\text{ind } A = \dim N(A)$.*
- (ii) *A is injective iff $\dim N(A) = 0$.*
- (iii) *A is bijective iff $\text{ind } A = \dim N(A) = 0$.*
- (iv) *If X and Y are Banach spaces, then the equation*

$$Au = b, \quad u \in X,$$

is well posed iff $\text{ind } A = \dim N(A) = 0$.

Proof. Ad (i). A is surjective iff $\text{codim } R(A) = 0$.

Ad (ii), (iii). This is obvious.

Ad (iv). This follows from Corollary 2 in Section 3.5. □

By Proposition 5(iv), Fredholm operators of index zero play a special role. We now make the following assumption.

(H) Let $A: X \rightarrow Y$ be a linear Fredholm operator, where X and Y are Banach spaces over \mathbb{K} .

Proposition 6. *Assume (H). Then the following properties are met:*

- (i) *The range $R(A)$ is closed.*
- (ii) *$R(A) = {}^\perp N(A^T)$ and $R(A^T) = N(A)^\perp$.*
- (iii) *$\text{codim } R(A^T) = \dim N(A)$.*
- (iv) *$\text{codim } R(A) = \dim N(A^T) = \dim N(A) - \text{ind } A$.*
- (v) *The dual operator $A^T: Y^* \rightarrow X^*$ is also Fredholm, and*

$$\text{ind } A^T = -\text{ind } A.$$

Proof. Ad (i). Cf. Standard Example 2 in Section 3.11.

Ad (ii). Cf. the *closed graph theorem* (Theorem 3.E).

Ad (iii), (iv). Use (ii) and Proposition 21 in Section 3.9.

Ad (v). Observe that

$$\begin{aligned} \text{ind } A^T &= \dim N(A^T) - \text{codim } R(A^T) \\ &= \text{codim } R(A) - \dim N(A) = -\text{ind } A. \end{aligned} \quad \square$$

In terms of the operator equation

$$Au = b, \quad u \in X, \tag{E}$$

and the dual equation

$$A^T u^* = b^*, \quad u^* \in Y^*, \tag{E^*}$$

Proposition 6 tells us the following. Assume (H). Then the following properties are met:

- (i) Original problem. *For given $b \in Y$, equation (E) has a solution $u \in X$ iff b satisfies the solvability condition*

$$\langle u^*, b \rangle = 0$$

for all solutions u^ of the homogeneous dual equation (E*)*.⁶

- (ii) Finiteness. *The homogeneous original equation (E) and the homogeneous dual equation (E*) have only a finite number of linearly independent solutions, and*

$$\text{ind } A = \dim N(A) - \dim N(A^T).$$

- (iii) Dual equation. *For given $b^* \in X^*$, equation (E*) has a solution $u^* \in Y^*$ iff*

$$\langle b^*, u \rangle = 0$$

for all solutions u of the homogeneous original equation (E).

- (iv) Well-posedness. *Let $\text{ind } A = 0$ and suppose that $Au = 0$ implies $u = 0$.*

Then, for each given $b \in Y$, the original equation (E) has a unique solution u , which depends continuously on b .

Moreover, for each given $b^ \in X^*$, the dual equation (E*) has a unique solution u^* , which depends continuously on b^* .*

5.5 The Riesz–Schauder Theory on Banach Spaces

Theorem 5.C. *Let X and Y be Banach spaces over \mathbb{K} . Then the operator*

$$B + C: X \rightarrow Y$$

is Fredholm of index zero provided the following hold:

⁶Recall that the homogeneous equations (E) and (E*) correspond to $b = 0$ and $b^* = 0$, respectively.

- (i) *The linear operator $B: X \rightarrow Y$ is continuous and bijective.*
- (ii) *The linear operator $C: X \rightarrow Y$ is compact.*

Proof. Use the proof of Theorem 5.B with the following obvious modifications:

- (a) Replace the adjoint operators B^* and C^* with the *dual* operators B^T and C^T , respectively.
- (b) Replace the special closed graph theorem with the general *closed graph theorem* (Theorem 3.E) in order to prove that

$$\text{codim}(B + C) < \infty.$$

- (c) Replace orthogonal direct sums with *direct sums*.
- (d) Replace orthogonal complements such as N^\perp , R^\perp , and so forth, with *topological complements*. □

5.6 Applications to the Spectrum of Linear Compact Operators

Theorem 5.D. *Let $A: X \rightarrow X$ be a linear compact operator on the complex Banach space $X \neq \{0\}$. Then the following statements hold true:*

- (i) *All nonzero points λ in the spectrum $\sigma(A)$ of A are eigenvalues of A with finite multiplicity.⁷*
- (ii) *The spectrum $\sigma(A)$ is either a finite set or a countable subset of \mathbb{C} with the only limit point $\lambda = 0$, which belongs to $\sigma(A)$.*
- (iii) *The spectrum $\sigma(A)$ is not empty.*

Proof. Ad (i). Let $\lambda \in \sigma(A)$ with $\lambda \neq 0$. By the definition of $\sigma(A)$, the operator $A - \lambda I$ is not bijective. Since this operator is Fredholm of index zero, we get $0 < \dim N(A - \lambda I) < \infty$ (i.e., λ is an eigenvalue of A with finite multiplicity).

Ad (ii). By Section 1.25 in AMS Vol. 108, $\sigma(A)$ is compact.

Suppose that $\lambda_n \in \sigma(A)$ for all n and let $\lambda_n \rightarrow \lambda$ as $n \rightarrow \infty$. Assume that $\lambda_n \neq \lambda_m$ if $n \neq m$. We have to show that $\lambda = 0$.

⁷The spectrum $\sigma(A)$ of an operator A was defined in Section 1.25 of AMS Vol. 108.

There exists a sequence (u_n) with $Au_n - \lambda_n u_n = 0$ for all n and $u_n \neq u_m$ if $n \neq m$. In addition, $u_n \neq 0$ for all n . Define

$$X_n := \text{span}\{u_1, \dots, u_n\}.$$

We show that the eigenvectors u_1, \dots, u_n are linearly independent, by an induction argument. Suppose that u_1, \dots, u_{n-1} are linearly independent and

$$u_n = \sum_{j=1}^{n-1} \alpha_j u_j \quad \text{for some } \alpha_1, \dots, \alpha_{n-1} \in \mathbb{C}.$$

It follows from

$$0 = Au_n - \lambda_n u_n = \sum_{j=1}^{n-1} \alpha_j (\lambda_j - \lambda_n) u_j$$

along with $\lambda_j \neq \lambda_n$ for all $j = 1, \dots, n-1$ that $\alpha_j = 0$, and hence $u_n = 0$. This contradicts $u_n \neq 0$.

Since X_{n-1} is a proper subspace of X_n , there exists a point $v_n \in X_n$ such that $\|v_n\| = 1$ and

$$\text{dist}(v_n, X_{n-1}) \geq \frac{1}{2} \quad \text{for all } n \geq 2, \quad (15)$$

by Step 1 of the proof of Theorem 2.B (almost orthogonal elements). Then $v_n = \alpha_n u_n + w_{n-1}$ for some $w_{n-1} \in X_{n-1}$ and some $\alpha_n \in \mathbb{C}$. Hence

$$Av_n - \lambda_n v_n = Aw_{n-1} - \lambda_n w_{n-1} \in X_{n-1},$$

since the space X_{n-1} is invariant under A . For $m < n$, it follows from (15) that

$$\left\| A\left(\frac{v_n}{\lambda_n}\right) - A\left(\frac{v_m}{\lambda_m}\right) \right\| = \left\| v_n + \frac{1}{\lambda_n}(Av_n - \lambda_n v_n) - \frac{1}{\lambda_m}Av_m \right\| \geq \frac{1}{2}, \quad (16)$$

since $Av_m \in X_{n-1}$. This implies

$$\frac{1}{|\lambda_n|} = \left\| \frac{v_n}{\lambda_n} \right\| \rightarrow \infty \quad \text{as } n \rightarrow \infty.$$

Otherwise, there would exist a bounded subsequence $\left(\frac{v_{n'}}{\lambda_{n'}}\right)$. Since A is compact, we may assume that $\left(A\left(\frac{v_{n'}}{\lambda_{n'}}\right)\right)$ is convergent, contradicting (16).

Ad (iii). We will use standard arguments from classic complex function theory.

Suppose that $\sigma(A)$ is empty. Then the operator $(A - \lambda I)^{-1}: X \rightarrow X$ is linear and continuous for each $\lambda \in \mathbb{C}$. Set $\mu := \frac{1}{\lambda}$. Then

$$R_\lambda := (A - \lambda I)^{-1} = \mu(\mu A - I)^{-1}. \quad (17)$$

By Example 6 in Section 1.23 of AMS Vol. 108, the series

$$\mu(\mu A - I)^{-1} = -\mu(I + \mu A + \mu^2 A^2 + \dots)$$

converges in $L(X, X)$ for all $\mu \in \mathbb{C}$ with $|\mu| < r$ if r is sufficiently small. Hence the series

$$(A - \lambda I)^{-1} = -\lambda^{-1}I[-\lambda^{-2}A + \dots] \quad (18)$$

converges in $L(X, X)$ for all $\lambda \in \mathbb{C}$ with $|\lambda| > r$. For each fixed $\lambda_0 \in \mathbb{C}$, we get

$$\begin{aligned} R_\lambda &= (A - \lambda_0 I - (\lambda - \lambda_0)I)^{-1} = R_{\lambda_0}(I - (\lambda - \lambda_0)R_{\lambda_0})^{-1} \\ &= R_{\lambda_0}(I + (\lambda - \lambda_0)R_{\lambda_0} + (\lambda - \lambda_0)^2R_{\lambda_0}^2 + \dots). \end{aligned} \quad (19)$$

This series converges in $L(X, X)$ for all $\lambda \in \mathbb{C}$ with $|\lambda - \lambda_0| < \rho$, where ρ is sufficiently small.

After these preparations, we define the function $\phi: \mathbb{C} \rightarrow \mathbb{C}$ through

$$\phi(\lambda) := f(R_\lambda u) \quad \text{for all } \lambda \in \mathbb{C},$$

where $f \in X^*$ and $u \in X$ are fixed such that $f(u) \neq 0$. By (19), the function ϕ allows the convergent series expansion

$$\phi(\lambda) = \phi(\lambda_0) + a_1(\lambda - \lambda_0) + \dots$$

for sufficiently small $|\lambda - \lambda_0|$. Note that the point λ_0 can be chosen arbitrarily. Thus, ϕ is holomorphic on \mathbb{C} . Hence

$$\oint_C \phi(\lambda) d\lambda = 0 \quad (20)$$

for each circle C around the origin. If C is sufficiently large, then it follows from (18) that

$$\begin{aligned} -\oint_C \phi(\lambda) d\lambda &= \oint_C (\lambda^{-1}f(u) + \lambda^{-2}f(Au) + \dots) d\lambda \\ &= \oint_C \lambda^{-1}f(u) d\lambda = 2\pi i f(u). \end{aligned}$$

Since $f(u) \neq 0$, this contradicts (20). \square

5.7 The Parametrix

Proposition 1. *Let $A: X \rightarrow Y$ be a linear continuous operator, where X and Y are Banach spaces over \mathbb{K} . Then the following two statements are equivalent:*

- (i) *The operator A is Fredholm.*
- (ii) *There exist linear continuous operators $P_l, P_r: Y \rightarrow X$ and linear compact operators $C_l: X \rightarrow X$ and $C_r: Y \rightarrow Y$ such that*

$$P_l A = I + C_l, \quad (21a)$$

$$A P_r = I + C_r. \quad (21b)$$

The operators P_l and P_r are called a left and right *parametrix* for A , respectively. The theory of pseudo-differential operators provides a systematic method for constructing the parametrices corresponding to large classes of differential and integral operators. This can be found in Hörmander (1983).

Proof. (i) \Rightarrow (ii). Choose linear subspaces V and W of X and Y , respectively, such that

$$X = N(A) \oplus V \quad \text{and} \quad Y = R(A) \oplus W.$$

This is possible by Proposition 6(i) in Section 5.4 and by Standard Example 17 from Section 3.9. Let

$$P: X \rightarrow N(A) \quad \text{and} \quad Q: Y \rightarrow W$$

be the corresponding linear continuous projection operators onto $N(A)$ and W , respectively. Define the linear continuous operator

$$B: R(A) \oplus W \rightarrow X$$

through

$$B(u + w) := A_0^{-1} u \quad \text{for all } u \in R(A), w \in W,$$

where $A_0: V \rightarrow R(A)$ denotes the restriction of $A: X \rightarrow Y$ to V . By Proposition 13 in Section 3.9, the operator A_0 is a linear *homeomorphism*. Finally, observe that

$$BA = I - P \quad \text{and} \quad AB = I - Q.$$

Since $P(X)$ and $Q(Y)$ are *finite-dimensional* linear spaces, the operators P and Q are *compact*. Now set $P_r = P_l := B$.

(ii) \Rightarrow (i). It follows from Theorem 5.C that the operators $I + C_l$ and $I + C_r$ are Fredholm of index zero.

By (21a), $N(A) \subseteq N(I + C_l)$. Hence

$$\dim N(A) \leq \dim N(I + C_l) < \infty.$$

Furthermore, it follows from (21b) that

$$R(I + C_r) \subseteq R(A).$$

Since $\text{codim } R(I + C_r) < \infty$ and $R(I + C_r)$ is closed, we get $\text{codim } R(A) < \infty$, by Corollary 11 in Section 3.9. \square

5.8 Applications to the Perturbation of Fredholm Operators

Let $F(X, Y)$ denote the set of all linear *Fredholm operators* $A: X \rightarrow Y$, where X and Y are Banach spaces over \mathbb{K} . Recall that $L(X, Y)$ denotes the Banach space of all linear continuous operators $B: X \rightarrow Y$ equipped with the operator norm $\|B\|$.

Proposition 1. *Let $S \in F(X, Y)$. Then there exists a number $\varepsilon > 0$ such that*

$$T \in F(X, Y) \quad \text{and} \quad \operatorname{ind} T = \operatorname{ind} S$$

for all operators $T \in L(X, Y)$ with $\|T - S\| < \varepsilon$.

Proof. Use Step 5 of the proof of Theorem 5.B along with the following modifications:

- (a) Replace orthogonal direct sums with *direct sums*.
- (b) Replace orthogonal complements such as N^\perp , R^\perp , and so on, with *topological complements*. \square

Since the index is an integer, Proposition 1 can be formulated in the following equivalent way:

The set $F(X, Y)$ is open in $L(X, Y)$, and the function $S \mapsto \operatorname{ind} S$ is continuous on $L(X, Y)$.

Theorem 5.E (Compact perturbations of Fredholm operators). *Let $S \in F(X, Y)$, and let $A \in L(X, Y)$ be a compact operator.*

Then the linear operator $S + A$ is Fredholm and

$$\operatorname{ind}(S + A) = \operatorname{ind} S.$$

Proof. We will use the method of the parametrix. Since $S \in F(X, Y)$, there exist operators $P_l, P_r \in L(X, Y)$ and compact operators $C_l \in L(X, X)$, $C_r \in L(Y, Y)$ such that

$$P_l S = I + C_l \quad \text{and} \quad S P_r = I + C_r,$$

by Section 5.7. Hence

$$P_l(S + A) = I + C_l + P_l A$$

and

$$(S + A)P_r = I + C_r + A P_r.$$

Since the operator A is compact, so are $P_l A$ and $A P_r$, by Proposition 4 in Section 5.1. Thus, P_l and P_r are a left and right *parametrix* for $S + A$, respectively. Hence $S + A$ is *Fredholm*.

Now consider the continuous function

$$t \mapsto S + tA$$

from $[0, 1]$ to $L(X, Y)$. Since tA is compact, the operator $S + tA$ is Fredholm for all $t \in [0, 1]$. By Proposition 1, $\text{ind}(S + tA) = \text{const}$ for all $t \in [0, 1]$. \square

5.9 Applications to the Product Index Theorem

Theorem 5.F. *Let*

$$X \xrightarrow{A} Y \xrightarrow{B} Z$$

be a sequence of linear Fredholm operators A and B , where X , Y , and Z are Banach spaces over \mathbb{K} .

Then the linear operator

$$X \xrightarrow{BA} Z$$

is also Fredholm and

$$\text{ind}(BA) = \text{ind } B + \text{ind } A. \quad (22)$$

Proof. We will use the method of the parametrix. Note that the products of *parametrices* for A and B produce parametrices for BA . In fact, it follows from Section 5.7 that, for $j = 1, 2$, there exist linear continuous operators $P_l^{(j)}$, $P_r^{(j)}$ and linear compact operators $C_l^{(j)}$, $C_r^{(j)}$ such that

$$P_l^{(1)} A = I + C_l^{(1)}, \quad AP_r^{(1)} = I + C_r^{(1)},$$

$$P_l^{(2)} B = I + C_l^{(2)}, \quad BP_r^{(2)} = I + C_r^{(2)}.$$

By Proposition 4 in Section 5.1, we obtain

$$P_l^{(1)} P_l^{(2)} BA = I + \text{linear compact operator};$$

$$B A P_r^{(1)} P_r^{(2)} = I + \text{linear compact operator}.$$

Thus, the product BA possesses right and left parametrices, that is, BA is Fredholm, by Section 5.7.

Let us compute the *index* of BA . The following sequences of finite-dimensional linear spaces are *exact*:

$$0 \rightarrow N(A) \rightarrow N(BA) \xrightarrow{A} R(A) \cap N(B) \rightarrow 0$$

$$0 \rightarrow R(B)/R(BA) \rightarrow Z/R(BA) \longrightarrow Z/R(B) \rightarrow 0$$

$$0 \rightarrow (R(A) + N(B))/R(A) \rightarrow Y/R(A) \xrightarrow{[B]} R(B)/R(BA) \rightarrow 0 \quad (22^*)$$

$$0 \rightarrow N(B) \cap R(A) \rightarrow N(B) \xrightarrow{\pi} (N(B) + R(A))/R(A) \rightarrow 0.$$

Here, the arrows without A , $[B]$, and π correspond to trivial inclusion maps, and π denotes the canonical map $\pi: Y \rightarrow Y/R(A)$. Recall that $\pi(u) := u + R(A)$. Furthermore,

$$[B](u + R(A)) := Bu + BR(A) = Bu + R(BA).$$

Observe that the factor space W/V is the collection of all the different subsets $w + V$, where $w \in W$.

The exactness of all the preceding sequences follows simply from

$$N(A) \subseteq N(BA) \quad \text{and} \quad R(BA) \subseteq R(B).$$

For example, $N(A) \subseteq N(BA)$ implies that the inclusion map $N(A) \rightarrow N(BA)$ is *injective*, that is,

$$0 \rightarrow N(A) \rightarrow N(BA)$$

is exact. Moreover, the map $A: N(BA) \rightarrow R(A) \cap N(B)$ is *surjective*, that is,

$$N(BA) \xrightarrow{A} R(A) \cap N(B) \rightarrow 0$$

is exact, and hence

$$0 \rightarrow N(A) \rightarrow N(BA) \xrightarrow{A} R(A) \cap N(B) \rightarrow 0$$

is exact, and so forth.

By Proposition 3 in Section 3.11, the exactness of $0 \rightarrow \mathcal{A} \rightarrow \mathcal{B} \rightarrow \mathcal{C} \rightarrow 0$ implies

$$\pm(\dim \mathcal{A} - \dim \mathcal{B} + \dim \mathcal{C}) = 0.$$

Applying this to the exact sequences in (22^*) , and summing the corresponding relations for the dimensions (with $+, -, +, -$), we obtain

$$\begin{aligned} \dim N(A) - \dim N(BA) + \dim (Z/R(BA)) - \dim (Z/R(B)) \\ - \dim (Y/R(A)) + \dim N(B) = 0. \end{aligned}$$

Observe that $\text{codim } R(BA) = \dim(Z/R(BA))$, and so on. Hence

$$\text{ind } A - \text{ind } BA + \text{ind } B = 0.$$

□

5.10 Fredholm Alternatives via Dual Pairs

We want to reformulate the Fredholm alternative in terms of dual pairs, which is convenient with a view to differential and integral equations (cf. Section 5.11).

Definition 1. Let X and Y be normed spaces over \mathbb{K} . We call $\{Y, X\}$ a *dual pair* iff there exists a bounded bilinear map $\langle \cdot, \cdot \rangle_D : Y \times X \rightarrow \mathbb{K}$ such that

- (i) $\langle v, u \rangle_D = 0$ for all $u \in X$ implies $v = 0$;
- (ii) $\langle v, u \rangle_D = 0$ for all $v \in Y$ implies $u = 0$.

Example 2. Let X be a normed space over \mathbb{K} . Then $\{X^*, X\}$ forms a *dual pair* with⁸

$$\langle v, u \rangle_D := \langle v, u \rangle \quad \text{for all } v \in X^*, u \in X.$$

Proof. For all $u \in X$ and $v \in X^*$,

$$|\langle v, u \rangle| \leq \|v\| \|u\|.$$

If $\langle v, u \rangle = 0$ for all $u \in X$ and fixed $v \in X^*$, then $v = 0$.

Conversely, if

$$\langle v, u \rangle = 0 \quad \text{for all } v \in X^* \text{ and fixed } u \in X,$$

then $u = 0$, by the *Hahn–Banach theorem* (cf. Standard Example 1 in Section 1.1).

Example 3. Let $X := C[a, b]$, where $-\infty < a < b < \infty$. Then $\{X, X\}$ forms a *dual pair* with respect to

$$\langle v, u \rangle_D := \int_a^b v(x)u(x)dx \quad \text{for all } u, v \in X.$$

Proof. Recall that $\|u\| = \max_{a \leq x \leq b} |u(x)|$. Hence

$$|\langle v, u \rangle_D| \leq (b - a)\|v\| \|u\| \quad \text{for all } u, v \in X.$$

If $\langle v, u \rangle_D = 0$ for all $u \in X$ and fixed $v \in X$, then $v(x) = 0$ on $[a, b]$, by Variational Lemma 10 in Section 2.2 of AMS Vol. 108. \square

⁸Recall that $\langle v, u \rangle = v(u)$.

Note that the dual space $C[a, b]^*$ corresponds to functions of bounded variation, by Section 1.3. Therefore, the dual pair $\{X, X\}$ from Example 3 possesses a *simpler structure* than the usual dual pair $\{X^*, X\}$.

Let us now consider the operator equation

$$Au = b, \quad u \in X, \tag{23}$$

along with the “dual” equation

$$A^D v = 0, \quad v \in X, \tag{23*}$$

where

$$\langle Au, v \rangle_D = \langle A^D v, u \rangle_D \quad \text{for all } u, v \in X. \tag{24}$$

Theorem 5.G (The Fredholm alternative). *Suppose that*

- (i) $\{Y, X\}$ is a dual pair, where X and Y are Banach spaces over \mathbb{K} .
- (ii) The linear continuous operators $A, A^D: X \rightarrow Y$ are Fredholm with

$$\operatorname{ind} A = -\operatorname{ind} A^D.$$

- (iii) The operator A^D is “dual” to A , meaning that relation (24) is satisfied.

Then, for each given $b \in Y$, the original equation (23) has a solution u iff

$$\langle b, v \rangle_D = 0$$

for all solutions v of the “dual” equation (23*).

The advantage of this theorem over the usual formulation is that the operators A and A^D live in the same space, in contrast to A and A^T . In applications, relation (24) corresponds frequently to integration by parts, as we shall see ahead.

Proof. We want to reduce this new situation to the usual Fredholm alternative.

Step 1: We first show that

$$\dim N(A^D) \leq \dim N(A^T). \tag{25}$$

To this end, let $\{v_1, \dots, v_n\}$ be a basis of $N(A^D)$. Define

$$f_j(u) := \langle u, v_j \rangle \quad \text{for all } u \in Y.$$

Then $f_j \in Y^*$, since $|f_j(u)| \leq \operatorname{const} \|u\| \|v_j\|$ for all $u \in Y$. By (24),

$$f_j(Au) = \langle Au, v_j \rangle_D = \langle A^D v_j, u \rangle_D = 0.$$

Hence

$$\langle A^T f_j, u \rangle = \langle f_j, Au \rangle = 0 \quad \text{for all } u \in Y,$$

that is, $f_1, \dots, f_n \in N(A^T)$. This implies (25).

The functionals $f_1, \dots, f_n \in Y^*$ are linearly independent. In fact it follows from

$$\alpha_1 f_1 + \cdots + \alpha_n f_n = 0 \quad \text{with } \alpha_1, \dots, \alpha_n \in \mathbb{K}$$

that $\langle u, \alpha_1 v_1 + \cdots + \alpha_n v_n \rangle_D = 0$ for all $u \in Y$. This implies $\alpha_1 v_1 + \cdots + \alpha_n v_n = 0$, and hence $\alpha_1 = \cdots = \alpha_n = 0$.

Step 2: Similarly, we get

$$\dim N(A) \leq \dim N((A^D)^T). \quad (26)$$

Step 3: Since the operators A and A^D are Fredholm, we obtain

$$\text{codim } R(A) = \dim N(A^T) \quad \text{and} \quad \text{codim } R(A^D) = \dim N((A^D)^T).$$

Thus, it follows from

$$\begin{aligned} \text{ind } A &= \dim N(A) - \text{codim } R(A) \\ &\leq \text{codim } R(A^D) - \dim N(A^D) = -\text{ind } A^D \end{aligned}$$

along with $\text{ind } A = -\text{ind } A^D$ that we can replace \leq with $=$ in equations (25) and (26). Consequently, $\{f_1, \dots, f_n\}$ forms a basis of $N(A^T)$.

Step 4: Since the operator A is Fredholm, the original equation $Au = b$, $u \in X$, has a solution iff

$$\langle f_j, b \rangle = 0 \quad \text{for all } j = 1, \dots, n.$$

By the definition of f_j , this is equivalent to

$$\langle b, v_j \rangle_D = 0 \quad \text{for all } j = 1, \dots, n. \quad \square$$

Dual pairs play an important role in the modern theory of nonlinear partial differential equations. This can be found in Zeidler (1986), Vol. 2B, Theorems 27.B and 30.B, as well as Vol. 5, Theorems 83.Uff.

5.11 Applications to Integral Equations and Boundary-Value Problems

Let us first consider the *integral equation*

$$u(x) - \int_a^b \mathcal{A}(x, y)u(y)dy = h(x), \quad a \leq x \leq b, \quad (27)$$

along with the dual integral equation

$$v(x) - \int_a^b \mathcal{A}(y, x)v(y)dy = 0, \quad a \leq x \leq b. \quad (27^*)$$

Standard Example 1. Let the function $\mathcal{A}: [a, b] \times [a, b] \rightarrow \mathbb{R}$ be continuous, where $-\infty < a < b < \infty$.

Then, for given $h \in C[a, b]$, the original problem (27) has a solution $u \in C[a, b]$ iff

$$\int_a^b h(x)v(x)dx = 0,$$

for all solutions $v \in C[a, b]$ of the dual equation (27*).

The same result has been obtained in Section 5.3 by means of Hilbert space methods combined with a regularization argument.

Proof. Set $X := C[a, b]$. We will use the dual pair $\{X, X\}$ with

$$\langle v, u \rangle_D := \int_a^b v(x)u(x)dx \quad \text{for all } u, v \in X$$

(cf. Example 3 in Section 5.10). Define

$$(Au)(x) := u(x) - \int_a^b \mathcal{A}(x, y)u(y)dy$$

and

$$(A^D v)(x) := v(x) - \int_a^b \mathcal{A}(y, x)v(y)dy$$

for all $x \in [a, b]$. By Standard Example 12 in Section 1.11 of AMS Vol. 108, the linear operator $A: X \rightarrow X$ is a compact perturbation of the identity. Hence Theorem 5.E tells us that $A: X \rightarrow X$ is Fredholm of index zero. The same argument shows that $A^D: X \rightarrow X$ is Fredholm of index zero.

Finally, for all $u, v \in X$,

$$\int_a^b \left(\int_a^b \mathcal{A}(x, y)u(y)dy \right) v(x)dx = \int_a^b \left(\int_a^b \mathcal{A}(x, y)v(x)dx \right) u(y)dy,$$

and hence we get the *duality relation*

$$\langle Au, v \rangle_D = \langle A^D v, u \rangle_D \quad \text{for all } u, v \in X.$$

Thus, the assertion follows from Theorem 5.G. □

Next we want to study the following *boundary-value problem*:

$$\alpha u'' + \beta u' + \gamma u = h \quad \text{on } [a, b], \quad u(a) = u(b) = 0, \quad (28)$$

along with the dual problem

$$(\alpha v)'' - (\beta v)' + \gamma v = 0 \quad \text{on } [a, b], \quad v(a) = v(b) = 0. \quad (28^*)$$

Standard Example 2. Let

$$\alpha \in C^2[a, b], \quad \beta \in C^1[a, b], \quad \gamma \in C[a, b],$$

where $-\infty < a < b < \infty$, and suppose that $\alpha(x) > 0$ on $[a, b]$.

Then, for given $h \in C[a, b]$, the original problem (28) has a solution $u \in C^2[a, b]$ iff

$$\int_a^b h(x)v(x)dx = 0,$$

for all solutions $v \in C^2[a, b]$ of the dual problem (28*).

Proof. Set

$$X := \{u \in C^2[a, b] : u(a) = u(b) = 0\} \quad \text{and} \quad Y := C[a, b],$$

along with the norm

$$\|u\|_X := \max_{a \leq x \leq b} |u(x)| + \max_{a \leq x \leq b} |u'(x)| + \max_{a \leq x \leq b} |u''(x)|,$$

and $\|u\|_Y := \max_{a \leq x \leq b} |u(x)|$. Then X and Y are real *Banach spaces*.

We want to use the *dual pair* $\{Y, X\}$ with

$$\langle v, u \rangle_D := \int_a^b v(x)u(x)dx \quad \text{for all } v \in Y, u \in X.$$

Define the linear operators $A, A^D : X \rightarrow Y$ through

$$Au := \alpha u'' + \beta u' + \gamma u \quad \text{for all } u \in X,$$

and

$$A^Dv := (\alpha v)'' - (\beta v)' + \gamma v \quad \text{for all } v \in X.$$

For all $u, v \in X$, *integration by parts* yields the *duality relation*

$$\begin{aligned} \langle Au, v \rangle_D &= \int_a^b (\alpha u'' + \beta u' + \gamma u)v dx \\ &= \int_a^b (-(\alpha v)'u' - (\beta v)'u + \gamma uv)dx \\ &= \int_a^b ((\alpha v)'' - (\beta v)' + \gamma v)u dx = \langle A^Dv, u \rangle_D. \end{aligned}$$

Finally, let us prove that the operators $A, A^D: X \rightarrow Y$ are *Fredholm of index zero*. Then, the assertion follows from Theorem 5.G. To this end, define the linear operators $B, C: X \rightarrow Y$ through

$$Bu := \alpha u''$$

and

$$Cu := \beta u' + \gamma u \quad \text{for all } u \in X.$$

We will show the following:

- (a) $B: X \rightarrow Y$ is linear, continuous, and bijective.
- (b) $C: X \rightarrow Y$ is linear and compact.

This implies that the operator $A := B + C$ is a compact perturbation of the Fredholm operator B of index zero. By Theorem 5.E, the operator A is Fredholm of index zero.

Ad (a). Obviously,

$$\|Bu\|_Y \leq \max_{a \leq x \leq b} |\alpha(x)| \max_{a \leq x \leq b} |u''(x)| \leq \text{const} \|u\|_X \quad \text{for all } u \in X. \quad (29)$$

Hence B is continuous. For given $h \in Y$ and $\rho \in \mathbb{R}$, the initial-value problem

$$\alpha u'' = h \quad \text{on } [a, b], \quad u(a) = 0, \quad u'(a) = \rho,$$

has the unique solution

$$u = u_0 + \rho(x - a),$$

where u_0 corresponds to $\rho = 0$. Choosing ρ in an appropriate way, we find that, for each given $h \in Y$, the boundary-value problem

$$\dot{\alpha} u'' = h \quad \text{on } [a, b], \quad u(a) = u(b) = 0,$$

has a unique solution $u \in X$. Thus, $B: X \rightarrow Y$ is bijective.

Ad (b). As in (29), we obtain

$$\|Cu\|_Y \leq \text{const} \|u\|_X \quad \text{for all } u \in X.$$

Thus, C is continuous. Let M be a bounded subset of X . Then, for all $u \in X$ and all $x, y \in [a, b]$, we get

$$\begin{aligned} |u'(x) - u'(y)| &\leq \left(\max_{a \leq z \leq b} |u''(z)| \right) |x - y| \\ &\leq \left(\sup_{u \in M} \|u\|_X \right) |x - y| \end{aligned}$$

and

$$|u(x) - u(y)| \leq \left(\sup_{u \in M} \|u\|_X \right) |x - y|.$$

By the *Arzelà–Ascoli theorem* (Standard Example 7 in Section 1.11 of AMS Vol. 108), the set $C(M)$ is relatively compact in Y , that is, the operator $C: X \rightarrow Y$ is compact.

Using the same argument, we see that the operator $A^D: X \rightarrow Y$ is also Fredholm of index zero. \square

5.12 Bifurcation Theory

To explain the basic idea of bifurcation theory, let us consider the following two real equations:

$$u - u_0 - (p - p_0) = 0 \quad (30)$$

and

$$(u - u_0)[(u - u_0) - (p - p_0)] = 0. \quad (31)$$

The solutions are pictured in Figures 5.2(a) and 5.2(b), respectively. Obviously, in Figure 5.2(a), the solution curve through the point (u_0, p_0) is unique in a neighborhood of (u_0, p_0) , in contrast to Figure 5.2(b). We say that (u_0, p_0) is a bifurcation point in Figure 5.2(b).

In the natural sciences, bifurcation points correspond to substantial changes in the behavior of systems. For example, Figure 5.3 displays a beam buckling under the influence of an outer force p . If the force becomes critical (i.e., $p = p_0$), then the rest state passes over to a buckled state.⁹ One frequently observes the following principle in nature:

Loss of stability leads to bifurcation.

We begin with the operator equation

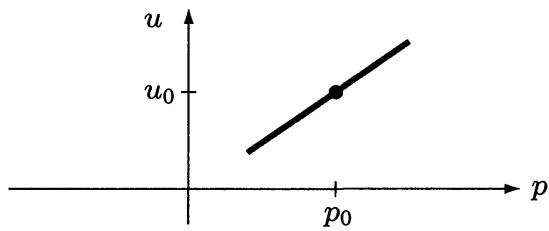
$$F(u, p) = 0, \quad u \in X, \quad p \in \Pi, \quad (32)$$

where X and Π are Banach spaces over \mathbb{K} . Here p is regarded as a parameter living in the parameter space Π (e.g., $\Pi = \mathbb{K}$).

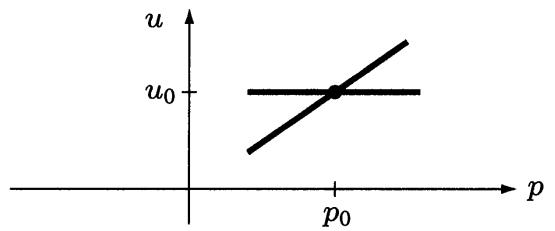
Definition 1. The point (u_0, p_0) is called a *bifurcation point* of (32) iff the following conditions are met:

- (i) $F(u_0, p_0) = 0$;
- (ii) For $n = 1, 2, \dots$, there are two sequences $\{(u_n, p_n)\}$ and $\{(v_n, p_n)\}$ of solutions to equation (32) that converge to (u_0, p_0) as $n \rightarrow \infty$.

⁹ A detailed mathematical study of this problem can be found in Zeidler (1986), Vol. 2B, Section 29.13.

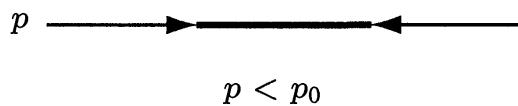


(a) no bifurcation point

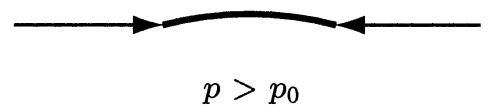


(b) bifurcation point

FIGURE 5.2.



(a)



(b)

FIGURE 5.3.

These are *distinct* sequences, that is, $u_n \neq v_n$ for all $n = 1, 2, \dots$.

Proposition 2 (Necessary bifurcation condition). *Let*

$$F: U(u_0, p_0) \subseteq X \times \Pi \rightarrow Y$$

be a C^1 -map on an open neighborhood of the point (u_0, p_0) , where X, Y , and Π are Banach spaces over \mathbb{K} .

If (u_0, p_0) is a bifurcation point of (32), then the linearization

$$F_u(u_0, p_0): X \rightarrow Y$$

is not bijective.

Proof. This follows immediately from the implicit function theorem in Section 4.8. \square

We now want to formulate an important, *sufficient* bifurcation condition in the case where $\Pi := \mathbb{K}^n$, that is, $p = (p_1, \dots, p_n)$. We assume the following:

- (H1) Let $F: U(u_0, p_0) \subseteq X \times \mathbb{K}^n \rightarrow Y$ be a C^2 -map on an open neighborhood of the point (u_0, p_0) , where X and Y are Banach spaces over \mathbb{K} , and $n \geq 1$.
- (H2) Trivial solution. For all $p \in \mathbb{K}^n$ in an open neighborhood of $p_0 \in \mathbb{K}^n$,

$$F(u_0, p) = 0.$$

- (H3) Linearization. The linearized operator $F_u(u_0, p_0): X \rightarrow Y$ is *Fredholm*.

Suppose that there exists a $b \in N(F_u(u_0, p_0))$ with $b \neq 0$, that is,

$$F_u(u_0, p_0)b = 0.$$

Moreover, suppose that $v_1^*, \dots, v_n^* \in Y^*$ form a basis of $R(F_u(u_0, p_0))^\perp$. This is equivalent to the fact that v_1^*, \dots, v_n^* are linearly independent and that the linearized equation

$$F_u(u_0, p_0)u = v, \quad u \in X, \quad (33)$$

has a solution for given $v \in Y$ iff $\langle v_j^*, v \rangle = 0$ for all $j = 1, \dots, n$.

(H4) Bifurcation condition. Let

$$\det(a_{jk}) \neq 0,$$

where $a_{jk} := \langle v_j^*, F_{p_k u}(u_0, p_0)b \rangle$, $j, k = 1, \dots, n$.

Theorem 5.H. Assume (H1) through (H4). Then (u_0, p_0) is a bifurcation point of the equation

$$F(u, p) = 0, \quad u \in X, \quad p \in \mathbb{K}^n. \quad (34)$$

Proof. Without loss of generality, we may assume that $u_0 = 0$ and $p_0 = 0$.

Since $\dim N(F_u(0, 0)) < \infty$, there exists a topological direct sum

$$X = N(F_u(0, 0)) \oplus W.$$

Choose elements $v_1, \dots, v_n \in Y$ with $\langle v_i^*, v_j \rangle = \delta_{ij}$ for $i, j = 1, \dots, n$, and set

$$Qv := \sum_{i=1}^n \langle v_i^*, v \rangle v_i.$$

Then the linearized equation (33) has a solution iff $Qv = 0$, by (H3). Thus,

$$I - Q: Y \rightarrow R(F_u(0, 0))$$

represents a continuous projection onto the range $R(F_u(0, 0))$. Define

$$H(w, p, s) := \begin{cases} s^{-1}F(sb + sw, p) & \text{if } s \neq 0, \\ F_u(0, p)(b + w) & \text{if } s = 0. \end{cases} \quad (35)$$

We will show the following:

- (a) The operator $H: U(0, 0, 0) \subseteq W \times \mathbb{K}^n \times \mathbb{K} \rightarrow \mathbb{K}$ is C^1 on an open neighborhood of the point $(0, 0, 0)$.

(b) The linearization

$$(w, p) \mapsto H_w(0, 0, 0)w + H_p(0, 0, 0)p$$

is *bijective* from $W \times \mathbb{K}^n$ onto Y .

It then follows from the *implicit function theorem* in Section 4.8 that, for each $s \in \mathbb{K}$ in an open neighborhood of $s = 0$, the equation

$$H(w, p, s) = 0 \quad (36a)$$

has a unique C^1 -solution

$$w = w(s), \quad p = p(s), \quad s = s \quad (36b)$$

in an open neighborhood of the point $(w, p, s) = (0, 0, 0)$ in $W \times \mathbb{K}^n \times \mathbb{K}$. Since $H(0, 0, 0) = 0$, we get $w(0) = 0$ and $p(0) = 0$. Consequently, for sufficiently small $|s|$ with $s \neq 0$, the original equation (34) has the *nontrivial* solution

$$u = sb + sw(s), \quad p = p(s)$$

(i.e., $u \neq 0$). Since equation (34) also possesses the trivial solution $(u, p) = (0, p)$, the point $(0, 0)$ is a *bifurcation point* of (34).

Ad (a). Assume first that F is *analytic* on a neighborhood of the point $(0, 0)$ in $X \times \Pi$. Hence

$$F(u, p) = \sum_{j,k=0}^{\infty} M_{jk} u^j p^k \quad (37)$$

with

$$\sum_{j,k=0}^{\infty} \|M_{jk}\| \|u\|^j \|p\|^k < \infty,$$

for all (u, p) in some open neighborhood of $(0, 0)$ in $X \times \Pi$. Since $F(0, p) \equiv 0$,

$$M_{0k} = 0 \quad \text{for all } k = 0, 1, 2, \dots$$

Furthermore, differentiating relation (37) at $(0, p)$, we get

$$F_u(0, p)h = \sum_{k=0}^{\infty} M_{1k} h p^k \quad \text{for all } h \in X.$$

Thus,

$$H(w, p, s) = \sum_{j \geq 1, k \geq 0} s^{j-1} M_{jk} (b + w)^j p^k, \quad (38)$$

along with

$$\sum \|M_{jk}\| |s|^{j-1} \|b + w\|^j \|p\|^k < \infty,$$

meaning that H is analytic on some open neighborhood of $(0, 0, 0)$. Hence H is C^∞ on that neighborhood. Furthermore,

$$H_w(0, 0, 0) = M_{10} = F_u(0, 0), \quad H_p(0, 0, 0)p = M_{11}bp = F_{pu}(0, 0)pb. \quad (39)$$

For the general case, the proof of (a) will be given in Problem 5.1, by means of the Taylor theorem.

Ad (b). Set $B := F_u(0, 0)$ and $C := F_{pu}(0, 0)$. Then the linearized equation $H_w(0, 0, 0)w + H_p(0, 0, 0)p = h$ is identical to

$$Bw + Cpb = h, \quad (w, p) \in W \times \mathbb{K}^n. \quad (40)$$

If we note that $QB = 0$, then (40), after applying Q and $(I - Q)$, is equivalent to

$$QCpb = Qh, \quad (41a)$$

$$Bw = (I - Q)(h - Cpb), \quad (w, p) \in W \times \mathbb{K}^n. \quad (41b)$$

Equation (41a) is equivalent to

$$\sum_{j=1}^n p_j QF_{p_j u}(0, 0)b = Qh,$$

that is,

$$\sum_{j=1}^n p_j \langle v_i^*, F_{p_j u}(0, 0)b \rangle = \langle v_i^*, Qh \rangle, \quad i = 1, \dots, n.$$

By (H4), this equation has a unique solution $p \in \mathbb{K}^n$, for each given $h \in Y$. Thus, equation (41a) can be solved uniquely for p ; since $w \in W$, we can then solve the second equation (41b) uniquely for w , by (H3). This proves (b). \square

5.13 Applications to Nonlinear Integral Equations

Let us consider the integral equation

$$u(t) = p \sum_{k=1}^{\infty} \int_a^b \mathcal{A}_k(t, s)u(s)^k ds, \quad a \leq t \leq b, \quad p \in \mathbb{R}, \quad u \in X, \quad (42)$$

with the real Banach space $X := C[a, b]$, $-\infty < a < b < \infty$. The linearized problem is

$$u(t) = p_0 \int_a^b \mathcal{A}_1(t, s)u(s)ds, \quad a \leq t \leq b, \quad p_0 \in \mathbb{R}, \quad u \in X, \quad (43)$$

with the corresponding dual equation

$$v(t) = p_0 \int_a^b \mathcal{A}_1(s, t)v(s)ds, \quad a \leq t \leq b, \quad p_0 \in \mathbb{R}, \quad v \in X. \quad (43^*)$$

We set $(u | v) := \int_a^b u(t)v(t)dt$ for all $u, v \in X$. Assume that every function $\mathcal{A}_k: [a, b] \times [a, b] \rightarrow \mathbb{R}$ is continuous and that the majorant series

$$\sum_{k=1}^{\infty} \max_{a \leq t, s \leq b} |\mathcal{A}_k(t, s)| \xi^k$$

is convergent for all real numbers ξ in some open neighborhood of zero.

Recall that p_0 is called a *characteristic number* of (43) iff equation (43) has a nontrivial solution u . Moreover, p_0 is called a *simple characteristic number* iff (43) has precisely one linearly independent solution.

Proposition 1. (i) The regular case. Suppose that p_0 is not a characteristic number of (43). Then there exist numbers $\rho > 0$ and $r > 0$ such that, for each given $p \in \mathbb{R}$ with $|p - p_0| < \rho$, the original problem (42) has a unique solution $u \in X$ with¹⁰ $\|u\| < r$.

(ii) The bifurcation case. Let p_0 be a simple characteristic number of (43). Then $(0, p_0) \in X \times \mathbb{R}$ is a bifurcation point of the original problem (43) provided that

$$(u | v) \neq 0, \quad (44)$$

where u and v are nontrivial solutions to (43) and (43*), respectively.

Proof. Let us write the original equation (42) in the form

$$F(u, p) = 0, \quad u \in X, \quad p \in \mathbb{R}.$$

Obviously, the operator $F: U(0, p_0) \subseteq X \times \mathbb{R} \rightarrow X$ is analytic on some open neighborhood of the point $(0, p_0)$ in $X \times \mathbb{R}$. The linearized equation

$$F_u(0, p_0)h = w, \quad h \in X, \quad (45)$$

corresponds to

$$h(t) - p_0 \int_a^b \mathcal{A}_1(t, s)h(s)ds = w(t), \quad a \leq t \leq b, \quad h \in X.$$

Note that $F_u(0, p_0): X \rightarrow X$ is a *Fredholm operator of index zero*.

Moreover, the partial derivative $(h, p) \mapsto F_{pu}(0, p_0)ph$ corresponds to the integral operator

$$p \int_a^b \mathcal{A}_1(t, s)h(s)ds.$$

¹⁰Recall that $\|u\| = \max_{a \leq x \leq b} |u(x)|$.

Ad (i). By hypothesis, equation (45) with $w = 0$ has only the trivial solution $h = 0$. Hence $F_u(0, p_0): X \rightarrow X$ is bijective. The assertion now follows from the implicit function theorem in Section 4.8.

Ad (ii). Let $v \in X$ be a nontrivial solution of (43*). Set

$$\langle v^*, g \rangle := (v | g) \quad \text{for all } g \in X.$$

Then, $v^* \in X^*$. By Example 1 in Section 5.11, for given $w \in X$, equation (45) has a solution h iff

$$\langle v^*, w \rangle = 0.$$

The bifurcation condition (H4) from Section 5.12(i) reads as follows:

$$\begin{aligned} \langle v^*, F_{pu}(0, p_0)u \rangle &= \int_a^b v(t) \left(\int_a^b \mathcal{A}_1(t, s)u(s)ds \right) dt \\ &= p_0^{-1}(v | u) \neq 0. \end{aligned}$$

The assertion now follows from Theorem 5.H. \square

Remark 2. Suppose that $\mathcal{A}_1(t, s) > 0$ for all $t, s \in [a, b] \times [a, b]$. Then, by the classical theorem of Jentzsch,¹¹ the integral operator $L: X \rightarrow X$ belonging to the kernel \mathcal{A}_1 , namely,

$$(Lh)(t) := \int_a^b \mathcal{A}_1(t, s)h(s)ds \quad \text{for all } t \in [a, b]$$

has a positive spectral radius r , and $p_0 := r^{-1}$ is a simple characteristic number of both (43) and (43*), where the corresponding eigenfunctions u and v are positive on $[a, b]$. Thus, condition (44) is satisfied automatically, and hence $(0, p_0) \in X \times \mathbb{R}$ is a bifurcation point of (42).

5.14 Applications to Nonlinear Boundary-Value Problems

Let us study the boundary-value problem

$$-u''(t) + q(t)u(t) = p \left(u(t) + \sum_{k=2}^{\infty} a_k u(t)^k \right) \quad \text{on } [a, b], \quad (46)$$

$$u(a) = u(b) = 0, \quad p \in \mathbb{R},$$

¹¹This is a special case of the functional analytic Krein–Rutman theorem (cf. Zeidler (1986), Vol. 1, Example 7.30).

where $-\infty < a < b < \infty$, along with the linearized problem

$$\begin{aligned} -u''(t) + q(t)u(t) &= p_0u(t) && \text{on } [a, b], \\ u(a) = u(b) &= 0, && p_0 \in \mathbb{R}. \end{aligned} \quad (47)$$

Let a_k be fixed real numbers for which the series $\sum_k a_k \xi^k$ converges in an open neighborhood of $\xi = 0$ in \mathbb{R} . Let $q: [a, b] \rightarrow \mathbb{R}$ be a continuous function. Set

$$X := \{u \in C^2[a, b]: u(a) = u(b) = 0\}, \quad Y := C[a, b].$$

Proposition 1. *The point $(0, p_0) \in X \times \mathbb{R}$ is a bifurcation point of (46) iff p_0 is an eigenvalue of (47).*

Proof. Let us write equation (46) in the form

$$F(u, p) = 0, \quad u \in X, \quad p \in \mathbb{R}.$$

Then $F: U(0, p_0) \subseteq X \times \mathbb{R} \rightarrow X$ is analytic on some open neighborhood of the point $(0, p_0)$. The linearized equation

$$F_u(0, p_0)h = w, \quad h \in X, \quad (48)$$

corresponds to the boundary-value problem

$$\begin{aligned} -h''(t) + q(t)h(t) &= p_0h(t) + w(t) && \text{on } [a, b], \\ h(a) = h(b) &= 0. \end{aligned}$$

Let p_0 be an eigenvalue of (47) with the eigenfunctions u and v . Since $u(a) = u'(a) = 0$ along with (47) imply $u \equiv 0$, we get $u'(a) \neq 0$ and $v'(a) \neq 0$. Hence $u'(a) = \lambda v'(a)$ for some $\lambda \in \mathbb{R}$, and the uniqueness of the solution to the initial-value problem in (47) yields $u \equiv \lambda v$ on $[a, b]$. Thus, there exists precisely one linearly independent eigenfunction u to p_0 .

By Example 2 in Section 5.11, for given $w \in Y$, problem (49) has a solution $h \in X$ iff

$$(u \mid w) := \int_a^b u(t)w(t)dt = 0 \quad (50)$$

and the operator $F_u(0, p_0): X \rightarrow Y$ is Fredholm of index zero.

The partial derivative $(h, p) \mapsto F_{pu}(0, p_0)ph$ corresponds to the operator

$$(h, p) \mapsto -ph.$$

By (50), the decisive *bifurcation condition* (H4) of Theorem 5.H reads as follows:

$$(u \mid F_{pu}(0, p_0)u) = - \int_a^b u(t)^2 dt \neq 0.$$

The assertion now follows from Proposition 2 and Theorem 5.H in Section 5.12. \square

5.15 Nonlinear Fredholm Operators

In this section, X and Y denote real Banach spaces.

Definition 1. Let $A: U \subseteq X \rightarrow Y$ be a C^1 -map, where U is open. Then A is called a *Fredholm map* iff the linearization

$$A'(u): X \rightarrow Y$$

is Fredholm for all $u \in U$.

Since $u \mapsto A'(u)$ is continuous, the index $\text{ind } A'(u)$ is locally constant, by Section 5.8. Thus, if $U = X$, then

$$\text{ind } A := \text{ind } A'(u)$$

is independent of $u \in X$. This number is called the *index* of A .

Definition 2. The operator $A: U \subseteq X \rightarrow Y$ is called *proper* iff the preimage $A^{-1}(\mathcal{C})$ of every compact set \mathcal{C} in Y is also compact.

Standard Example 3. The operator $A: X \rightarrow Y$ is proper provided

$$A = B + C,$$

where $B: X \rightarrow Y$ is a homeomorphism, $C: X \rightarrow Y$ is compact, and

$$\|Au\| \rightarrow \infty \quad \text{as } \|u\| \rightarrow \infty. \quad (51)$$

Proof. Let \mathcal{C} be a compact subset of Y , and let $Au_n \in \mathcal{C}$ for all $n \in \mathbb{N}$. It suffices to show that (u_n) contains a convergent subsequence $(u_{n'})$. Then $u_{n'} \rightarrow u$ as $n \rightarrow \infty$, and hence $Au \in \mathcal{C}$, that is, $u \in A^{-1}(\mathcal{C})$. In the following, we do not distinguish between sequences and subsequences.

Since \mathcal{C} is bounded, it follows from (51) that (u_n) is bounded. Consequently,

$$Cu_n \rightarrow w \quad \text{as } n \rightarrow \infty$$

for some w . All Au_n live in the compact set \mathcal{C} , and so

$$Au_n \rightarrow v \quad \text{as } n \rightarrow \infty$$

for some v . Thus, $Bu_n \rightarrow v - w$ as $n \rightarrow \infty$. Since B is a homeomorphism, $u_n \rightarrow B^{-1}(v - w)$ as $n \rightarrow \infty$. \square

Let us now consider the equation

$$Au = w, \quad u \in X. \quad (52)$$

Theorem 5.I (The Smale principle). *Suppose that the C^1 -operator $A: X \rightarrow Y$ is Fredholm and proper with $\text{ind } A = 0$.*

Then, for each $w_0 \in Y$ and $\varepsilon > 0$, there exists a point $w \in Y$ with $\|w - w_0\| < \varepsilon$ such that the original equation (52) has at most a finite number of solutions.

Smale proved this theorem in 1965. Roughly speaking, Theorem 5.I tells us that in “most cases” equation (52) has at most a finite number of solutions. One also says that (52) has generically at most a finite number of solutions. The proof will be given later after some preparations.

Definition 4. Let the map $A: X \rightarrow Y$ be C^1 .

- (i) The point $u \in X$ is called a *regular point* of A iff $A'(u): X \rightarrow Y$ is surjective. Otherwise, u is called a singular point of A .
- (ii) The point $v \in Y$ is called a *regular value* of A iff the preimage $A^{-1}(v)$ is empty or consists solely of regular points. Otherwise, v is called a singular value of A (i.e., $A^{-1}(v)$ contains at least one singular point).

Proposition 5 (Sard’s theorem). *If $A: \mathbb{R}^m \rightarrow \mathbb{R}^n$ is a C^k -mapping with $k > \max(0, m - n)$ and $m, n \in \mathbb{N}$, then the set of singular values of A has n -dimensional Lebesgue measure zero in \mathbb{R}^n .*

Consequently, the set of regular values of A is dense in \mathbb{R}^n .

Sard proved this famous classic result in 1942. A proof can be found in Abraham and Robbin (1967).

Example 6. Let $f: \mathbb{R} \rightarrow \mathbb{R}$ be a C^1 -function. Then u is a regular point of f iff

$$f'(u) \neq 0.$$

Moreover, v is a singular value of f iff there exists a point u such that

$$f(u) = v \quad \text{and} \quad f'(u) = 0.$$

The Sard theorem tells us that

Singular values are rare (cf. Figure 5.4).

Example 7. Suppose that the C^1 -operator $A: X \rightarrow Y$ is Fredholm and proper with $\text{ind } A = 0$. Let w be a regular value of A . Then

- (i) If $Au = w$, A is a local C^1 -diffeomorphism at u .
- (ii) The equation $Au = w$ has at most a finite number of solutions.

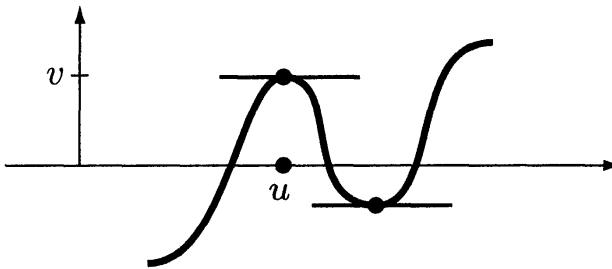


FIGURE 5.4.

Proof. Ad (i). Since the Fredholm operator $A'(u): X \rightarrow Y$ is surjective and $\text{ind } A'(u) = 0$, $A'(u): X \rightarrow Y$ is bijective. By the local inverse mapping theorem from Section 4.10, A is a local C^1 -diffeomorphism at u .

Ad (ii). Since A is proper, the set $A^{-1}(w)$ is compact. Suppose that there exists an infinite sequence (u_n) with

$$Au_n = w \quad \text{for all } n.$$

Since $A^{-1}(w)$ is compact, there exists a subsequence, again denoted by (u_n) , such that $u_n \rightarrow u$ as $n \rightarrow \infty$. Hence $Au = w$. By (i), u is an *isolated* solution of $Au = w$. This is a contradiction. \square

Let us assume the following.

(H) $A: U(u_0) \subseteq X \rightarrow Y$ is a C^k -Fredholm map with $k > \max(\text{ind } A'(u_0), 0)$, where $U(u_0)$ is an open neighborhood of u_0 .

Lemma 8. *If (H) holds, then there exists an open neighborhood $V(u_0)$ of u_0 in X such that the regular values of the restriction $A|_{V(u_0)}$ are dense in Y .*

Proof. We make essential use of the local *normal form* (81) from Section 4.12. Let $N := N(A'(u_0))$ and $R := R(A'(u_0))$. We choose topological direct sums

$$X = N \oplus N_c \quad \text{and} \quad Y = R \oplus R_c. \quad (53)$$

By Proposition 1 in Section 4.12, there exists a C^k -diffeomorphism

$$\phi: \mathcal{U}(0, 0) \subseteq N \times R \rightarrow V(u_0)$$

such that the relation

$$h(n, r) = Au_0 + r + g(n, r) \quad \text{on } \mathcal{U}(0, 0) \quad (54)$$

holds for $h(n, r) := A(\phi(n, r))$, with $n \in N$, $r \in R$, and $g(n, r) \in R_c$ on $\mathcal{U}(0, 0)$.

The dimensions of N and R_c are finite, because $A'(u_0)$ is Fredholm. The product of a linear surjective operator with a linear bijective operator is

surjective. Thus, by the local inverse mapping theorem from Section 4.10, regular values are invariant under diffeomorphisms. Consequently, it suffices to show that the regular values of h are dense in Y .

Let $v \in Y$. We decompose

$$v = A(u_0) + v_1 + v_2 \quad \text{where } v_1 \in R \text{ and } v_2 \in R_c. \quad (54^*)$$

Let $\psi(n) := g(n, v_1)$. Then $\psi: U(0) \subseteq N \rightarrow R_c$ is C^k , by (54). Letting

$$m := \dim N - \dim R_c,$$

we obtain $m = \text{ind } A'(u_0)$ and $k > \max(m, 0)$ from (H).

(a) According to *Sard's theorem* (Proposition 5), the regular values of ψ are dense in R_c .

(b) We show that if v_2 is a regular value of ψ , then v is a regular value of h . Indeed, by (54) and (54*), from $h(n, r) = v$ it follows that $r = v_1$ and $\psi(n) = v_2$. Moreover, by (54), we have

$$h'(n, v_1)(\tilde{n}, \tilde{r}) = \tilde{r} + \psi'(n)\tilde{n} + g_r(n, v_1)(0, \tilde{r})$$

for all $\tilde{n} \in N$, $\tilde{r} \in R$. Note that $Y = R \oplus R_c$ as well as $g_r(n, v_1)(0, \tilde{r}) \in R_c$ for all $\tilde{r} \in R$. Therefore, the surjectivity of $\psi'(n): N \rightarrow R_c$ implies the surjectivity of

$$h'(n, v_1): N \times R \rightarrow Y.$$

It follows from (a) and (b) that the regular values of h are dense in Y . \square

Lemma 9. *If (H) holds, then A is locally closed, that is, A maps closed sets contained in a sufficiently small open neighborhood of the point u_0 onto closed sets.*

Proof. Since $h = A \circ \phi$, it suffices to show that h maps closed sets onto closed sets.¹² Let

$$h(n_k, r_k) \rightarrow w \quad \text{as } k \rightarrow \infty,$$

and let

$$w = Au_0 + w_1 + w_2, \quad \text{where } w_1 \in R \text{ and } w_2 \in R_c.$$

Assume that (n_k, r_k) lives in the bounded closed set M for all k . It follows from (53) and (54) that $r_k \rightarrow w_1$ as $k \rightarrow \infty$. Since $\dim N < \infty$, we find that $n_k \rightarrow n$ as $k \rightarrow \infty$ for some $n \in N$, after passing to a subsequence, if necessary. Hence $h(n, w_1) = w$ and $(n, w_1) \in M$. \square

Lemma 10. *If (H) holds and u_0 is a regular point of A , then there exists an open neighborhood of Au_0 that contains only regular points of A .*

¹²Observe that ϕ is a homeomorphism and use Lemma 12 from Section 3.9.

Proof. This follows from normal form (54) with $R = Y$ and $g \equiv 0$. Note that

$$h'(n, r)(\tilde{n}, \tilde{r}) = \tilde{r} \quad \text{for all } \tilde{r} \in Y$$

and all $(n, r) \in U(0, 0)$, that is, $h'(n, r)$ is surjective for these points. \square

Corollary 11. *If (H) holds, then there exists an open neighborhood $W(u_0)$ of u_0 in X such that the set of singular values of the restriction $A|_{W(u_0)}$ is closed in Y , and the set of regular values of $A|_{W(u_0)}$ is open and dense in Y .*

Thus, the set of singular values of $A|_{W(u_0)}$ is nowhere dense in Y .

Proof. Let \mathcal{R} be the set of regular points of A in a sufficiently small neighborhood $W(u_0)$ of the point u_0 . By Lemma 10, the set \mathcal{R} is open, and hence the set $S := W(u_0) - \mathcal{R}$ of singular points of A in $W(u_0)$ is closed. Lemma 9 tells us that $A(S)$ is closed in Y . Thus, the set $Y - A(S)$ of regular values of $A|_{W(u_0)}$ is open and dense in Y , by Lemma 8. \square

Lemma 12. *If $A: X \rightarrow Y$ is continuous and proper, then A transforms closed sets onto closed sets.*

Proof. Let C be a closed set in X , and let $Au_n = v_n$, where $u_n \in C$ for all $n \in \mathbb{N}$ and $v_n \rightarrow v$ as $n \rightarrow \infty$. The set of all v_n together with v is compact. Therefore, (v_n) contains a convergent subsequence with $v_{n'} \rightarrow u$ as $n \rightarrow \infty$. Since C is closed and A is continuous, we have $u \in C$ and $Au = v$. This means that $A(C)$ is closed. \square

Proposition 13 (Sard–Smale theorem). *Let $A: X \rightarrow Y$ be a proper C^k -Fredholm map with $k > \max(\text{ind } A, 0)$.*

Then the set of regular values of A is open and dense in Y .

Proof of Theorem 5.I. Proposition 13 and Example 7 immediately imply Theorem 5.I. \square

Proof of Proposition 13. *Step 1:* By Lemma 10, the set of the regular points of A is open in X . Thus, the set of the singular points of A is closed in X , and hence Lemma 12 tells us that the set of singular values of A is closed in Y . Consequently, the set $\text{reg}(A)$ of the regular values of A is open in Y .

Step 2: Choose a fixed point $v \in Y$. Let U be an open set in X such that $A^{-1}(v) \subseteq U$. Then there exists an open neighborhood $V(v)$ of the point v with

$$A^{-1}(V(v)) \subseteq U.$$

Otherwise, there exists a convergent sequence $Au_n \rightarrow v$ as $n \rightarrow \infty$ with $u_n \notin U$ for all n . Since A is proper, we may assume that, $u_n \rightarrow u$ as

$n \rightarrow \infty$ after passing to a subsequence, if necessary. This yields the desired contradiction $u \notin U$ and $Au = v$.

Step 3: By Corollary 11, there exists an open neighborhood $W(u)$ for each point $u \in A^{-1}(v)$ such that the *singular values* of the restriction

$$A|_{W(u)}$$

form a nowhere dense set $S(u)$ in Y . Since A is *proper*, the set $A^{-1}(v)$ is compact, and hence finitely many sets $W(u_1), \dots, W(u_m)$ already cover the set $A^{-1}(v)$ (cf. Problem 1.14). Set

$$U := \bigcup_{j=1}^m W(u_j).$$

Then the set

$$S_U := \bigcup_{j=1}^m S(u_j)$$

of singular values of the restriction $A|_U$ is of the first Baire category.

Step 3: According to Step 2 there exists an open neighborhood $V(v)$ of v such that $A^{-1}(V(v)) \subseteq U$. Hence the set of singular values S of $A: X \rightarrow Y$ in $V(v)$ is equal to S_U (i.e., S is of the *first Baire category*). Hence the set $V(v) - S$ of regular values of A in $V(v)$ is *dense* in $V(v)$ (cf. Problem 3.1).

Since the point v can be chosen arbitrarily, the set of regular values of operator A is *dense* in Y . \square

5.16 Interpolation Inequalities

Let $0 < \alpha < 1$. Inequalities of the type

$$\|u\|_Y \leq \text{const} \|u\|_X^\alpha \|u\|_Z^{1-\alpha}, \quad \text{for all } u \in X, \quad (55)$$

are called *interpolation inequalities*. Such inequalities play a fundamental role in modern analysis. They allow us to give efficient existence proofs for nonlinear partial differential equations. For example, the interpolation inequality (60) implies the compactness of the embedding $\overset{\circ}{W}_2^1(G) \subseteq L_4(G)$. This will be used in the next section in order to give an existence proof for the famous stationary Navier–Stokes equations.¹³ Interpolation inequalities for Sobolev spaces follow from integral inequalities based on the Hölder inequality.

¹³A detailed study of interpolation inequalities can be found in Zeidler (1986), Vol. 2A, Section 21.17ff.

Recall the following. Let X and Z be Banach spaces over \mathbb{K} such that $X \subseteq Z$. Define an operator $E: X \rightarrow Z$ that assigns to each element u of X the same u , but now regarded as an element of Z . The operator E is called an *embedding operator*. The embedding $X \subseteq Z$ is called *continuous* iff the operator E is continuous, that is,

$$\|u\|_Z \leq \text{const} \|u\|_X \quad \text{for all } u \in X.$$

Furthermore, the embedding $X \subseteq Z$ is called compact iff the operator E is compact, that is, the embedding $X \subseteq Z$ is continuous and each bounded sequence in X has a subsequence that converges in Z .

The following simple result is crucial.

Proposition 1 (Compact embedding). *Let X , Y , and Z be Banach spaces over \mathbb{K} such that we have the inclusions*

$$X \subseteq Y \quad \text{and} \quad X \subseteq Z$$

along with the interpolation inequality (55) for the norms. Then the following conditions are met:

- (i) *If the embedding $X \subseteq Z$ is continuous, then so is $X \subseteq Y$.*
- (ii) *If the embedding $X \subseteq Z$ is compact, then so is $X \subseteq Y$.*

Proof. Ad (i). Since the embedding $X \subseteq Z$ is continuous,

$$\|u\|_Z \leq \text{const} \|u\|_X \quad \text{for all } u \in X.$$

By (55), $\|u\|_Y \leq \text{const} \|u\|_X$ for all $u \in X$.

Ad (ii). Let (u_n) be a bounded sequence in X . Since the embedding $X \subseteq Z$ is compact, there is a subsequence $(u_{n'})$ that converges in Z . Let us denote $(u_{n'})$ by (u_n) . It follows from (55) that

$$\begin{aligned} \|u_n - u_m\|_Y &\leq \text{const} \|u_n - u_m\|_X^\alpha \|u_n - u_m\|_Z^{1-\alpha} \\ &\leq \text{const} \|u_n - u_m\|_Z^{1-\alpha} \quad \text{for all } n, m. \end{aligned} \tag{56}$$

The sequence (u_n) is Cauchy in Z . By (56), (u_n) is also Cauchy in Y . Thus, the sequence (u_n) converges in Y . \square

Let G be a nonempty bounded open set in \mathbb{R}^3 . Define¹⁴

$$\begin{aligned} (u | v)_2 &:= \int_G uv dx, \quad (u | v) := \int_G \partial_j u \partial_j v dx, \\ (u | v)_{1,2} &:= (u | v)_2 + (u | v), \end{aligned}$$

¹⁴We sum over two equal indices from 1 to 3.

and

$$\|u\|_2 := (u \mid u)_2^{\frac{1}{2}}, \quad \|u\| := (u \mid u)^{\frac{1}{2}}.$$

Furthermore, let

$$\|u\|_4 := \left(\int_G u^4 dx \right)^{\frac{1}{4}}.$$

Lemma 2. *The norm $\|\cdot\|$ is equivalent to the original norm $\|\cdot\|_{1,2}$ on the Sobolev space $\overset{\circ}{W}_2^1(G)$.*

Proof. By the Poincaré–Friedrichs inequality, there is a constant $c > 0$ such that

$$c\|u\|_2^2 \leq \|u\|^2 \quad \text{for all } u \in \overset{\circ}{W}_2^1(G)$$

(cf. Section 2.5.6 in AMS Vol. 108). Hence $c\|u\|_2^2 + c\|u\|^2 \leq (1 + c)\|u\|^2$. Thus, there is a constant $d > 0$ such that

$$d\|u\|_{1,2} \leq \|u\| \leq \|u\|_{1,2} \quad \text{for all } u \in \overset{\circ}{W}_2^1(G). \quad (57)$$

□

Let (u_n) be a sequence in $\overset{\circ}{W}_2^1(G)$. Relation (57) tells us that (u_n) is Cauchy with respect to the norm $\|\cdot\|_{1,2}$ iff it is Cauchy with respect to the equivalent norm $\|\cdot\|$. Consequently, $\overset{\circ}{W}_2^1(G)$ is a Hilbert space equipped with the new inner product $(\cdot \mid \cdot)$. In what follows we will always refer to this new inner product.

Definition 3. Let $L_4(G)$ denote the set of all measurable functions $u: G \rightarrow \mathbb{R}$ such that $\|u\|_4 < \infty$.

Then, $L_4(G)$ becomes a real Banach space¹⁵ with respect to the norm $\|\cdot\|_4$. The following result will be critically used in the next section.

Proposition 4. *The embedding $\overset{\circ}{W}_2^1(G) \subseteq L_4(G)$ is compact.*

Proof. Set

$$X := \overset{\circ}{W}_2^1(G), \quad Y := L_4(G), \quad \text{and} \quad Z := L_2(G).$$

We will show in Lemma 7 ahead that in this case the interpolation inequality (55) holds true.

¹⁵See Problem 5.9 for a more general result. By definition, two functions u and v represent the same element of $L_4(G)$ iff their values differ only on a set of measure zero.

Rellich's compactness theorem from Section 5.7 in AMS Vol. 108 tells us that the embedding $X \subseteq Z$ is compact. Thus, it follows from Proposition 1 that the embedding $X \subseteq Y$ is also compact. \square

It remains to prove Lemma 7. To accomplish this, we need some preparations.

Lemma 5. *For all $u \in C_0^\infty(\mathbb{R}^2)$,*

$$\int_{\mathbb{R}^2} u^4 dx \leq 4 \int_{\mathbb{R}^2} u^2 dx \int_{\mathbb{R}^2} (u_\xi^2 + u_\eta^2) dx, \quad (58)$$

where $x = (\xi, \eta)$.

Proof. In what follows we write briefly the integral \int instead of $\int_{-\infty}^\infty$. From

$$u(\xi, \eta)^2 = \int_{-\infty}^\xi (u^2)_\xi d\xi = 2 \int_{-\infty}^\xi uu_\xi dx$$

we obtain the *key inequality*:

$$\max_{\xi \in \mathbb{R}} u(\xi, \eta)^2 \leq 2 \int |uu_\xi| d\xi. \quad (59)$$

Hence,

$$\begin{aligned} \iint u^4 d\xi d\eta &\leq \iint \max_{\xi \in \mathbb{R}} u(\xi, \eta)^2 \max_{\eta \in \mathbb{R}} u(\xi, \eta)^2 d\xi d\eta \\ &= \int \max_{\xi \in \mathbb{R}} u(\xi, \eta)^2 d\eta \int \max_{\eta \in \mathbb{R}} u(\xi, \eta)^2 d\xi \\ &\leq 4 \iint |uu_\xi| d\xi d\eta \iint |uu_\eta| d\xi d\eta. \end{aligned}$$

Therefore, by the Schwarz inequality,

$$\iint u^4 dx \leq 4 \left(\iint u^2 dx \right)^{\frac{1}{2}} \left(\iint u_\xi^2 dx \right)^{\frac{1}{2}} \left(\iint u^2 dx \right)^{\frac{1}{2}} \left(\iint u_\eta^2 dx \right)^{\frac{1}{2}}.$$

This immediately implies (58). \square

Lemma 6. *For all $u \in C_0^\infty(\mathbb{R}^3)$,*

$$\int_{\mathbb{R}^3} u^4 dx \leq 8 \left(\int_{\mathbb{R}^3} u^2 dx \right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^3} (u_\xi^2 + u_\eta^2 + u_\zeta^2) dx \right)^{\frac{3}{2}},$$

where $x = (\xi, \eta, \zeta)$.

Proof. Set $J := \int_{\mathbb{R}^3} u^4 dx$. By Lemma 5,

$$\begin{aligned} J &= \int d\zeta \iint u(\xi, \eta, \zeta)^4 d\xi d\eta \\ &\leq 4 \int d\zeta \iint u^2 d\xi d\eta \iint (u_\xi^2 + u_\eta^2) d\xi d\eta \\ &\leq 4 \iint \max_{\zeta \in \mathbb{R}} u(\xi, \eta, \zeta)^2 d\xi d\eta \iint \int (u_\xi^2 + u_\eta^2) d\xi d\eta d\zeta. \end{aligned}$$

Applying the key inequality (59), we obtain

$$J \leq 8 \iint \int |uu_\zeta| dx \iint \int (u_\xi^2 + u_\eta^2) dx.$$

By the Schwarz inequality,

$$J \leq 8 \left(\iint \int u^2 dx \right)^{\frac{1}{2}} \left(\iint \int u_\zeta^2 dx \right)^{\frac{1}{2}} \left(\iint \int (u_\xi^2 + u_\eta^2) dx \right). \quad \square$$

Lemma 7. For all $u \in \overset{\circ}{W}_2^1(G)$, we have the crucial Ladyzhenskaya inequality

$$\|u\|_4 \leq 8 \|u\|_2^{\frac{1}{4}} \|u\|_2^{\frac{3}{4}}. \quad (60)$$

Recall that $\|\cdot\|$ denotes the norm on $\overset{\circ}{W}_2^1(G)$ corresponding to the inner product $(\cdot | \cdot)$.

Proof. Let $u \in \overset{\circ}{W}_2^1(G)$. Since the set $C_0^\infty(G)$ is dense in $\overset{\circ}{W}_2^1(G)$, there exists a sequence (u_n) in $\overset{\circ}{W}_2^1(G)$ such that $u_n \rightarrow u$ in $\overset{\circ}{W}_2^1(G)$ as $n \rightarrow \infty$. Hence,

$$u_n \rightarrow u \text{ in } L_2(G) \text{ as } n \rightarrow \infty, \quad (61)$$

by Lemma 2. Furthermore, Lemma 6 tells us that

$$\|u\|_4 \leq 8 \|u_n\|_2^{\frac{1}{4}} \|u_n\|_2^{\frac{3}{4}} \quad \text{for all } n, \quad (62)$$

and

$$\|u_n - u_m\|_4 \leq 8 \|u_n - u_m\|_2^{\frac{1}{4}} \|u_n - u_m\|_2^{\frac{3}{4}} \quad \text{for all } n, m. \quad (63)$$

Thus, the sequence (u_n) is Cauchy in the Banach space $L_4(G)$. This implies the existence of a function $v \in L_4(G)$ such that

$$u_n \rightarrow v \text{ in } L_4(G) \text{ as } n \rightarrow \infty. \quad (64)$$

We want to show that

$$u(x) = v(x) \quad \text{for almost all } x \in G. \quad (65)$$

Then, letting $n \rightarrow \infty$, we obtain the desired inequality (60) from (62).

To prove (65) we will use the following standard trick. By Problem 5.9a(iv), it follows from (61) that there is a subsequence $(u_{n'})$ of (u_n) such that

$$u_{n'}(x) \rightarrow u(x) \quad \text{as } n' \rightarrow \infty \text{ for almost all } x \in G.$$

Similarly, it follows from (64) that there is a subsequence $(u_{n''})$ of $(u_{n'})$ such that

$$u_{n''}(x) \rightarrow v(x) \quad \text{as } n'' \rightarrow \infty \text{ for almost all } x \in G.$$

This implies (65). \square

Proposition 8. *Let \mathcal{H} and Z be real Hilbert spaces such that the embedding*

$$\mathcal{H} \subseteq Z$$

is continuous, and \mathcal{H} is dense in Z . Then the embedding

$$Z \subseteq \mathcal{H}^* \tag{66}$$

is continuous. In addition, Z is dense in the dual space \mathcal{H}^ .*

The precise interpretation of relation (66) will be given in Step 3 of the following proof.

Proof. *Step 1:* The injective map $\psi: Z \rightarrow \mathcal{H}^*$. Let $v \in Z$. Define v^* through

$$v^*(u) := (v | u)_Z \quad \text{for all } u \in \mathcal{H}. \tag{67}$$

Then

$$|v^*(u)| \leq \|v\|_Z \|u\|_Z \quad \text{for all } u \in \mathcal{H}.$$

Since the embedding $\mathcal{H} \subseteq Z$ is continuous, $\|u\|_Z \leq \text{const} \|u\|_{\mathcal{H}}$ for all $u \in \mathcal{H}$. Hence,

$$|v^*(u)| \leq \text{const} \|v\|_Z \|u\|_{\mathcal{H}} \quad \text{for all } u \in \mathcal{H}. \tag{68}$$

This shows that $v^*: \mathcal{H} \rightarrow \mathbb{R}$ is a linear continuous functional on \mathcal{H} (i.e., $v^* \in \mathcal{H}^*$). Furthermore, it follows from (68) that

$$\|v^*\|_{\mathcal{H}^*} \leq \text{const} \|v\|_Z \quad \text{for all } v \in Z. \tag{69}$$

Define now the map $\psi: Z \rightarrow \mathcal{H}^*$ through

$$\psi(v) := v^*.$$

Then, ψ is linear and continuous, by (69). Moreover, ψ is injective. In fact, if $\psi(v) = 0$ for fixed $v \in Z$, then $(v | u)_Z = 0$ for all $u \in \mathcal{H}$. Since \mathcal{H} is dense in Z , $v = 0$.

Step 2: We show that the set $\psi(Z)$ is *dense* in \mathcal{H}^* . Suppose that this is not true. Then the closure $\overline{\psi(Z)}$ of $\psi(Z)$ in \mathcal{H}^* is a proper closed linear subspace of \mathcal{H}^* . Choose a point $u^* \in \mathcal{H}^*$ such that $u^* \notin \overline{\psi(Z)}$. By Section 1.2, there exists a functional $f \in (\mathcal{H}^*)^*$ such that $f(u^*) \neq 0$ and

$$f(v^*) = 0 \quad \text{for all } v^* \in \overline{\psi(Z)}. \quad (70)$$

Since the Hilbert space \mathcal{H} is *reflexive*, there exists a point $w \in \mathcal{H}$ such that

$$f(v^*) = v^*(w) \quad \text{for all } v^* \in \mathcal{H}^*.$$

By (70), $v^*(w) = 0$ for all $v^* \in \psi(Z)$. Thus, relation (67) tells us that

$$(v | w)_Z = 0 \quad \text{for all } v \in Z.$$

Hence $w = 0$. Therefore, we obtain $f = 0$, contradicting $f(u^*) \neq 0$.

Step 3: Interpretation of relation (66). Since the map $\psi: Z \rightarrow \mathcal{H}^*$ is injective, we can identify the point v in Z with the point $\psi(v)$ in \mathcal{H}^* . In this sense, we write $Z \subseteq \mathcal{H}^*$. \square

Proposition 9. *Let X and Y be normed spaces over \mathbb{K} such that the embedding $X \subseteq Y$ is continuous. In addition, let M be a set in the dual space X^* . Then the following hold true:*

- (i) *The embedding $Y^* \subseteq X^*$ is continuous.*
- (ii) *If M is open in X^* , then the intersection $Y^* \cap M$ is open in Y^* .*
- (iii) *If M is dense in X^* , then $Y^* \cap M$ is dense in Y^* .*

Proof. Ad (i). Let $f \in Y^*$. Then the functional $f: Y \rightarrow \mathbb{K}$ is linear and continuous. Hence the restriction $f: X \rightarrow \mathbb{K}$ is also linear and continuous. Moreover, it follows from

$$\sup\{|f(u)|: u \in X, \|u\|_X \leq 1\} \leq \sup\{|f(u)|: u \in Y, \|u\|_Y \leq 1\}$$

that

$$\|f\|_{X^*} \leq \|f\|_{Y^*}.$$

Ad (ii). The embedding operator $E: Y^* \rightarrow X^*$ is continuous. Thus, the preimage of open sets is again open. In particular, $E^{-1}(M)$ is open in Y^* . Observe that $E^{-1}(M) = Y^* \cap M$.

Ad (iii). Let $f \in Y^*$. Choose any $\varepsilon > 0$. Since M is dense in X^* , there exists a linear continuous functional $g: X \rightarrow \mathbb{K}$ such that $g \in M$ and

$$\|f - g\|_{X^*} < \varepsilon.$$

Since X is a linear subspace of Y , there exists an extension $g: Y \rightarrow \mathbb{K}$ such that $g \in Y^*$ and

$$\|f - g\|_{Y^*} < \varepsilon,$$

by the Hahn–Banach theorem (which allows norm-preserving extensions).

Finally, observe that $g \in Y^* \cap M$. \square

5.17 Applications to the Navier–Stokes Equations

Undoubtedly, the Navier–Stokes equations are of basic importance within the context of modern theory of partial differential equations. Although the range of their applicability to concrete problems has now been clearly recognized to be limited, as my dear friend and bright colleague K.R. Rajagopal has showed me by several examples during the last six years, the mathematical questions that remain open are of such a fascinating and challenging nature that analysts and applied mathematicians cannot help being attracted by them and trying to contribute to their resolution. Thus, it is not a coincidence that over the past ten years more than seventy significant research papers have appeared concerning the well-posedness of boundary and initial-boundary value problems.¹⁶

Giovanni Paolo Galdi (1994)

In this section, we want to combine important tools from functional analysis in order to solve the famous stationary Navier–Stokes equations. In particular, we will use the Riesz theorem, the closed range theorem, the Leray–Schauder principle, and the Smale principle for nonlinear proper Fredholm operators of index zero. Furthermore, we will use the theory of distributions and the theory of Sobolev spaces introduced in AMS Vol. 108.

We recommend that the reader study this section carefully. This way, the reader gets an impression of the modern approach to nonlinear partial differential equations arising in mathematical physics. We also like to show that the language of functional analysis is the right language for modern mathematical physics.

Let G be a nonempty, bounded, open, and connected set in \mathbb{R}^3 . The stationary motion of an incompressible viscous fluid in G is governed by the following Navier–Stokes equations:

$$\begin{aligned} -\eta \Delta \mathbf{v} + \rho(\mathbf{v} \nabla) \mathbf{v} &= \mathbf{K} - \nabla p && \text{(equation of motion),} \\ \nabla \cdot \mathbf{v} &= 0 && \text{(incompressibility conditions),} \\ \mathbf{v} &= 0 && \text{on } \partial G \quad \text{(boundary condition).} \end{aligned} \tag{71}$$

Here, we use the following notation:¹⁷

- $\mathbf{v}(x)$ = velocity vector of the fluid at the point x ,
- $p(x)$ = pressure at the point x ,
- ρ = constant density of the fluid,

¹⁶The four-volume monograph by Galdi (1994) contains a detailed, up-to-date study of the Navier–Stokes equations.

¹⁷The total force acting on the fluid equals $\int_G \mathbf{K}(x) dx$.

$\mathbf{K}(x)$ = outer force density at the point x ,
 η = viscosity (a positive constant).

The full equation of motion

$$\rho \mathbf{v}_t - \eta \Delta \mathbf{v} + \rho(\mathbf{v} \nabla) \mathbf{v} = \mathbf{K} - \nabla p$$

corresponds to Newton's law: mass \times acceleration = force. In the stationary case, all the quantities are independent of time t . In particular, the time derivative \mathbf{v}_t of the velocity field vanishes. This yields the first equation in (71). The boundary condition in (71) reflects the experimental fact that a viscous fluid sticks to the boundary. To simplify notation, set $\rho = 1$.

Turbulence. Physical experiments show that turbulence occurs if the outer force densities \mathbf{K} are sufficiently large. This physical effect strongly complicates the mathematics of the Navier–Stokes equations.

A detailed physical motivation of the Navier–Stokes equations can be found in Zeidler (1986), Vol. 4, Section 70.3.

5.17.1 Reformulation of the Classical Problem

The original problem (71) can be written in the following form:

$$\begin{aligned} -\eta \Delta \mathbf{v} + \nabla(\mathbf{v} \otimes \mathbf{v}) &= \mathbf{K} - \nabla p && \text{(equation of motion),} \\ \nabla \mathbf{v} &= 0 && \text{(incompressibility condition),} \\ \mathbf{v} &= 0 && \text{(boundary condition).} \end{aligned} \quad (72)$$

To show this, let us choose a Cartesian coordinate system with the orthonormal basis \mathbf{e}_1 , \mathbf{e}_2 , and \mathbf{e}_3 . Then, $\mathbf{x} = x_j \mathbf{e}_j$ along with¹⁸

$$\mathbf{v} = v_j \mathbf{e}_j \quad \text{and} \quad \mathbf{K} = K_j \mathbf{e}_j.$$

Set $\partial_j := \partial/\partial_j$. Observe that

$$\begin{aligned} (\mathbf{v} \nabla) \mathbf{v} &= v_j \partial_j v_m \mathbf{e}_m = \partial_j(v_j v_m) \mathbf{e}_m - v_m(\partial_j v_j) \mathbf{e}_m \\ &= \partial_j(v_j v_m) \mathbf{e}_m = \nabla(\mathbf{v} \otimes \mathbf{v}) \end{aligned}$$

because $\nabla \mathbf{v} = \partial_j v_j = 0$. Thus, for $m = 1, 2, 3$, problem (71) reads as follows:

$$\begin{aligned} -\eta \partial_j \partial_j v_m + \partial_j(v_j v_m) &= K_m - \partial_m p \quad \text{(equation of motion),} \\ \partial_j v_j &= 0 \quad \text{on } G \quad \text{(incompressibility condition),} \\ v_m &= 0 \quad \text{on } \partial G \quad \text{(boundary condition).} \end{aligned} \quad (73)$$

This is identical to the invariant formulation in (72).

¹⁸We sum over two equal indices from 1 to 3.

5.17.2 The Classical Basic Idea

To find smooth solutions to the original problem (72), let us write this in the following form

$$\mathbf{P} = -\eta \Delta \mathbf{v} + \nabla(\mathbf{v} \otimes \mathbf{v}) - \mathbf{K} \quad \text{on } G, \quad \mathbf{v} \in \mathcal{X}, \quad (74)$$

$$-\nabla p = \mathbf{P} \quad \text{on } G, \quad (75)$$

$$\int_G \mathbf{P} \mathbf{w} dx = 0 \quad \text{for all } \mathbf{w} \in \mathcal{X}. \quad (76)$$

Here, \mathcal{X} denotes the set of all smooth velocity fluids \mathbf{v} on G that satisfy both the incompressibility condition

$$\nabla \mathbf{v} = 0 \quad \text{on } G$$

and the boundary condition $\mathbf{v} = 0$ on ∂G . From the physical point of view, \mathbf{P} is the force density generated by the pressure p in the fluid.

Let us discuss this.

Step 1: Suppose that (\mathbf{v}, p) is a smooth solution of the original problem (72). We want to show that (\mathbf{v}, p) satisfies equations (74) through (76).

In fact, $\mathbf{v} \in \mathcal{X}$. Moreover, it follows immediately from (72) that the velocity field \mathbf{v} and the pressure p satisfy equations (74) and (75). Equation (75) implies

$$\int_G \mathbf{P} \mathbf{w} dx = \int_G -(\nabla p) \mathbf{w} dx \quad \text{for all } \mathbf{w} \in \mathcal{X}.$$

Integration by parts yields

$$\int_G \mathbf{P} \mathbf{w} dx = \int_G p (\nabla \mathbf{w}) dx = 0 \quad \text{for all } \mathbf{w} \in \mathcal{X},$$

since $\nabla \mathbf{w} = 0$ on G . This shows that

Relation (76) represents a necessary solvability condition for the pressure equation (75).

In summary, (\mathbf{v}, p) is a solution to (74) through (76).

Step 2: Conversely, assume that there is a solution (\mathbf{v}, \mathbf{P}) of equation (74) such that relation (76) is additionally satisfied. We want to show that there exists a pressure function p that satisfies equation (75).

Let G be a simply connected region with smooth boundary. To determine the pressure p , we use the basic fact from classical vector calculus that

Relation (76) is also a sufficient solvability condition for the pressure equation (75).

Thus, equation (75) indeed has a solution p . To summarize, $(\mathbf{v}, p, \mathbf{P})$ is a solution of (74) through (76). This implies that (\mathbf{v}, p) is a solution of the original problem (72).

The preceding discussion shows the crucial fact that it suffices to find a velocity field $\mathbf{v} \in \mathcal{X}$ such that the following *orthogonality relation* holds true:

$$\int_G (-\eta \Delta \mathbf{v} + \nabla(\mathbf{v} \otimes \mathbf{v}) - \mathbf{K}) \mathbf{w} dx = 0 \quad \text{for all } \mathbf{w} \in \mathcal{X}. \quad (77)$$

This is the *key* to our approach.

Remark 1 (Generalized solutions). Our aim is to prove the existence of generalized (nonsmooth) solutions. Since equation (77) is related to orthogonality, we will use a *Hilbert space* approach. To reduce the order of highest derivatives that appear in our generalized problem, consider first a sufficiently smooth situation. Then, integrating relation (77) by parts yields¹⁹

$$\int_G (\eta \nabla \mathbf{v} \nabla \mathbf{w} - \mathbf{v}(\mathbf{v} \nabla) \mathbf{w} - \mathbf{K} \mathbf{w}) dx = 0 \quad \text{for all } \mathbf{w} \in \mathcal{X} \quad (78)$$

and fixed $\mathbf{v} \in \mathcal{X}$. Observe that $\mathbf{w} = 0$ on the boundary ∂G .

Using regularity theory, it can be shown that generalized solutions are also classical smooth solutions provided the boundary and the outer force densities are smooth (Remark 8).

To simplify notation, let

$$\mathbf{v} := (v_1, v_2, v_3), \quad \mathbf{K} := (K_1, K_2, K_3), \quad \phi := (\phi_1, \phi_2, \phi_3).$$

5.17.3 The Generalized Problem

Motivated by (78), define

$$(v | w) := \int_G \nabla \mathbf{v} \nabla \mathbf{w} dx = \int_G \partial_j v_m \partial_j w_m dx,$$

$$a(u, v, w) := - \int_G \mathbf{u}(\mathbf{v} \nabla) \mathbf{w} dx = - \int_G u_j v_m \partial_j w_m dx,$$

¹⁹Using components, equation (77) reads as follows:

$$\int_G (-\eta \partial_j \partial_j v_m + \partial_j(v_j v_m) - K_m) w_m dx = 0.$$

Since $w_m = 0$ on ∂G , integration by parts yields

$$\int_G (\eta \partial_j v_m \partial_j w_m - (v_j v_m) \partial_j w_m - K_m w_m) dx = 0.$$

This coincides with (78).

and

$$K(w) := \int_G \mathbf{K} \mathbf{w} dx = \int_G K_j w_j dx.$$

Then, the key equation (78) reads as follows:

$$\eta(v | w) + a(v, v, w) = K(w) \quad \text{for all } w \in \mathcal{X} \quad (79)$$

and fixed $v \in \mathcal{X}$.

Let us now introduce the relevant function spaces. Consider first the product space

$$\mathcal{H} := \overset{\circ}{W}_2^1(G) \times \overset{\circ}{W}_2^1(G) \times \overset{\circ}{W}_2^1(G).$$

By Lemma 2 in Section 5.16, \mathcal{H} is a real Hilbert space with respect to the inner product $(\cdot | \cdot)$. Moreover, define

$$X := \text{closure of the set } D \text{ in the Hilbert space } \mathcal{H}, \quad (80)$$

where

$$D := \{\phi: \phi_j \in C_0^\infty(G) \text{ for all } j \text{ and } \nabla \phi = 0\}.$$

Here, $\nabla \phi = \partial_j \phi_j$. Note that

$$D \subseteq \mathcal{X},$$

where \mathcal{X} was introduced in Section 5.17.2. By (80), X is a closed subspace of the Hilbert space \mathcal{H} . Thus, X is a Hilbert space equipped with the inner product $(\cdot | \cdot)$. In addition, the set D is dense in X . Observe that²⁰

$$(v | w) := \sum_{j=1}^3 (v | w)_{1,2}.$$

Furthermore, define

$$Z := L_2(G) \times L_2(G) \times L_2(G),$$

along with

$$(K | L)_Z := \sum_{j=1}^3 (K_j | L_j)_2 = \sum_{j=1}^3 \int_G K_j L_j dx.$$

Then Z is a real Hilbert space with respect to the inner product $(\cdot | \cdot)_Z$.

Definition 2. The *generalized problem* to the original problem (72) reads as follows. Let $\eta > 0$. We are given the outer force density $K \in \mathcal{H}^*$. We are

²⁰We use the notation introduced in Section 5.16.

looking for a velocity field $v \in X$ such that the following equation holds true:

$$\eta(v | \phi) + a(v, v, \phi) = K(\phi) \quad \text{for all } \phi \in D. \quad (81)$$

Remark 3 (Motivation). Let v be a smooth solution to the classical problem (72). Our discussion in Section 5.17.2 shows that v is also a generalized solution.

Conversely, we will show in Corollary 6 that each generalized solution is a solution to the original problem (72), in the sense of distribution theory and of generalized boundary values.

Remark 4 (The outer forces). (i) *Classical outer force densities.* Let the outer force density $K \in Z$ be given. Then the components K_1, K_2 , and K_3 of K live in $L_2(G)$. Define the functional

$$K(\phi) := \int_G K_j \phi_j dx \quad \text{for all } \phi \in \mathcal{H}. \quad (82)$$

It follows from²¹

$$|K(\phi)| \leq \|K_j\|_2 \|\phi_j\|_2 \leq \text{const} \|K\|_Z \|\phi\| \quad \text{for all } \phi \in \mathcal{H}$$

that $K \in \mathcal{H}^*$.

Since $C_0^\infty(G)$ is dense in $L_2(G)$, the set \mathcal{H} is dense in Z . This implies the continuous embedding

$$Z \subseteq \mathcal{H}^*.$$

Moreover, Z is dense in \mathcal{H}^* , by Proposition 8 in Section 5.16.

The outer force densities $K \in Z$ are called *classical outer force densities*.

(ii) *Generalized outer force densities.* Denote by $W_2^{-1}(G)$ the dual space to $\overset{\circ}{W}_2^1(G)$. Let $K_j \in W_2^{-1}(G)$ for all j . Define

$$K(\phi) := K_j(\phi_j) \quad \text{for all } \phi \in \mathcal{H}.$$

Then, $K \in \mathcal{H}^*$. Moreover, all elements from \mathcal{H}^* are obtained this way.

All the outer force densities $K \in \mathcal{H}^*$ are called *generalized outer force densities*. It follows from the continuous embedding $Z \subseteq \mathcal{H}^*$ along with

$$\|K\|_{\mathcal{H}^*} \leq \text{const} \|K\|_Z$$

that each classical outer force density is also a generalized outer force density, and the norm $\|K\|_{\mathcal{H}^*}$ is arbitrarily small provided the classical norm $\|K\|_Z$ is sufficiently small.

In modern mathematical physics, forces correspond quite often to functionals.

²¹Cf. Lemma 2 in Section 5.16.

The following proposition justifies the choice of the space X for velocity fields.

Proposition 5 (Velocity fields). *Let $v \in X$. Then, for $m = 1, 2, 3$,*

$$\begin{aligned}\partial_j v_j &= 0 \quad \text{on } G \quad (\text{incompressibility condition}), \\ v_m &= 0 \quad \text{on } \partial G \quad (\text{boundary condition}),\end{aligned}$$

where the incompressibility condition is to be understood in the sense of distribution theory, and where the boundary condition is satisfied in the sense of generalized boundary values.

Proof. If $v \in X$, then $v_m \in \overset{\circ}{W}_2^1(G)$ for all m . By Section 2.5.5 of AMS Vol. 108, the function v_m vanishes on the boundary ∂G (in the generalized sense).

To prove the incompressibility condition, let $v \in D$. Since $\partial_m v_m = 0$, we obtain

$$\int_G \phi \partial_m v_m dx = 0 \quad \text{for all } \phi \in C_0^\infty(G).$$

Integration by parts yields

$$\int_G v_m \partial_m \phi dx = 0 \quad \text{for all } \phi \in C_0^\infty(G). \quad (84)$$

Since D is dense in X , and since X -convergence implies $L_2(G)$ -convergence of v_m , a passage to the limit shows that relation (84) remains true for all $v \in X$.

Define

$$v_m(\phi) := \int_G v_m \phi dx \quad \text{for all } \phi \in C_0^\infty(G). \quad (85)$$

Since the function v_m lives in $L_2(G)$, the functional from (85) represents a distribution (cf. Section 2.8.3 in AMS Vol. 108). Furthermore,

$$(\partial_m v_m)(\phi) = -v_m(\partial_m \phi) = 0 \quad \text{for all } \phi \in C_0^\infty(G).$$

Consequently,

$$\partial_m v_m = 0 \quad \text{on } G,$$

in the sense of distribution theory. This is precisely the incompressibility condition. \square

5.17.4 The Fundamental Existence Theorem

Theorem 5.J. *Let G be a nonempty, bounded, open, connected set in \mathbb{R}^3 . Consider a viscous fluid in G with density²² $\rho > 0$ and viscosity $\eta > 0$.*

²²Here we do not use the convention $\rho = 1$ from Section 5.17.1 in order to clarify the physical statement. Therefore, in the proof of Theorem 5.J, we have to replace η and K with $\rho^{-1}\eta$ and $\rho^{-1}K$, respectively.

Then, for given outer force density $K \in \mathcal{H}^*$, the following conditions are met:

- (i) Existence. The generalized problem (81) has a solution $v \in X$.
- (ii) Uniqueness. If the outer force density is sufficiently small, then the velocity field $v \in X$ is unique.

More precisely, we have to assume that the dimensionless quantity

$$\frac{\rho}{\eta^2} \|K\|_{\mathcal{H}^*}$$

is sufficiently small.

- (iii) Generic finiteness. There is an open, dense subset \mathcal{H}_0^* of \mathcal{H}^* such that, for each outer force density $K \in \mathcal{H}_0^*$, the generalized Problem (81) has only a finite number of solutions $v \in X$.

There is an open, dense subset²³ Z_0 of Z such that for each (classical) outer force density $K \in Z_0$, the generalized problem (81) has only a finite number of solutions $v \in X$.

Corollary 6 (The pressure p). Suppose that the boundary ∂G of the region G is sufficiently smooth.²⁴ Let $v \in X$ be a solution of the generalized problem (81). Then there exists a pressure function $p \in L_2(G)$ such that

$$\begin{aligned} -\eta \partial_j \partial_j v_m + \partial_j(v_j v_m) &= K_m - \partial_m p && \text{on } G \text{ (equation of motion),} \\ \partial_j v_j &= 0 && \text{on } G \text{ (incompressibility condition),} \\ v_m &= 0 && \text{on } G \text{ (boundary condition),} \end{aligned} \tag{86}$$

where both the equation of motion and the incompressibility condition are to be understood in the sense of distribution theory, and where the boundary condition is satisfied in the sense of generalized boundary values.²⁵

Naturally enough, the pressure function p is unique up to a constant. If we use the normalization condition

$$\int_G p dx = 0,$$

then p is unique (as an element of the space $L_2(G)$).

²³Choose $Z_0 := \mathcal{H}_0^* \cap Z$.

²⁴For example, suppose that ∂G is a two-dimensional C^1 -manifold that lies locally on one side of ∂G .

²⁵Cf. Sections 2.5.5 and 2.8.3 of AMS Vol. 108.

Let us first discuss this.

Remark 7 (Turbulence). The lack of uniqueness²⁶ corresponds to the experimental fact that turbulence appears for sufficiently large outer forces. The perfect mathematical description of turbulence is a famous open problem in mathematical physics.

As an introduction to the modern theory of turbulence, we recommend the books by Chorin (1975, 1994), Sirovich (1991), and Foias et al. (1993).

Remark 8 (The stability property of reasonable classical forces). Recall that

$$Z = L_2(G) \times L_2(G) \times L_2(G),$$

along with

$$\|K\|_Z = \left(\int_G \mathbf{K}^2 dx \right)^{\frac{1}{2}}.$$

Furthermore, recall that the forces $K \in Z$ are called classical forces. Let us designate the forces $K_0 \in Z_0$ as *reasonable* classical forces because they are classical functions and they generate only a finite number of velocity fields $v \in X$. Since the set Z_0 is dense and open in Z , we obtain the following:

- (i) If K is a classical force, then for each $\varepsilon > 0$, there is a reasonable classical force K_* with $\|K - K_*\|_Z < \varepsilon$.
- (ii) If K is a reasonable classical force, then there is a number $\delta > 0$ such that each classical force K_* with $\|K - K_*\|_Z < \delta$ is reasonable.

This can be expressed briefly by saying that

Most classical forces are reasonable, and they remain reasonable under small perturbations.

That is, reasonable classical forces are generic.

Let us summarize our results for classical forces:

- (i) *For all classical outer force densities $K \in Z$, there exists a velocity field $v \in X$.*
- (ii) *If the dimensionless quantity*

$$\frac{\rho^2}{\eta^4} \int_G \mathbf{K}^2 dx$$

is sufficiently small, then the velocity field is unique.

²⁶In fact, there exists a counterexample, where nonuniqueness appears (cf. Galdi (1994), Vol. 2, p. 11).

- (iii) *Most classical outer force densities generate only a finite number of velocity fields. This property remains unchanged under small perturbations of the outer force densities.*
- (iv) *To each velocity field $v \in X$, there corresponds a pressure function p that is unique up to a constant (as an element of $L_2(G)$).²⁷*

Remark 9 (The efficiency of modern calculus). Formally, the equations from (86) look like the classical equations from (73). It is a decisive advantage of the modern theory of distributions that, on the one hand, this theory is powerful enough to force the existence of solutions to important problems in mathematical physics. On the other hand, modern calculus resembles classical calculus.

Historical Remark 10. The Navier–Stokes equations were formulated by Navier in 1822 and studied by Stokes in 1845. Existence and uniqueness theorems for the stationary Navier–Stokes equations were first proved by Odquist in 1930 and then by Leray in 1933. This time gap between 1822 and 1930 characterizes a *retardation* between physics and mathematics that happens quite often. For example, the Laplace equation was introduced by Laplace near 1800, but a deeper mathematical understanding of this equation via interpolation theory was gained only in the 1960s.

Our proof of statements (i) and (ii) from Theorem 5.J follows the elegant approach discovered by Ladyshenskaya in 1959. Foias and Temam proved a statement of type (iii) in 1977, by using the Sard–Smale theorem.

The proofs of Theorem 5.J and Corollary 6 given in Sections 5.17.6 and 5.17.8, respectively, will be a simple consequence of an abstract theorem based on the Riesz theorem, the Leray–Schauder principle, and the Smale principle. In addition, we will use the Hölder inequality and the results from Section 5.16 about interpolation inequalities and compact embeddings. Furthermore, we will use the generalization of a theorem from classical vector calculus to distribution theory.

A different proof via the Galerkin method for pseudomonotone operators can be found in Zeidler (1986), Vol. 4, Section 72.4.

5.17.5 A Functional Analytic Theorem

Let us make the following assumptions:

²⁷In addition, the sophisticated *regularity theory* for the stationary Navier–Stokes equations shows that if both the boundary and the outer forces are *smooth*, then so are the velocity field and the pressure.

More precisely, if the boundary ∂G is C^∞ , and if $K_j \in C^\infty(\overline{G})$ for all j , then $v_j, p \in C^\infty(\overline{G})$ for all j , and the equations in (86) are satisfied in the classical sense (cf. Galdi (1994), Vol. 2, Section 8.5).

- (H1) \mathcal{H} and Z are real Hilbert spaces, where the embedding $\mathcal{H} \subseteq Z$ is continuous, and \mathcal{H} is dense in Z .
- (H2) X is a closed linear subspace of \mathcal{H} , and D is a dense subset of X . Denote the inner product on X by $(\cdot | \cdot)$.
- (H3) Y is a real Banach space such that the embedding $X \subseteq Y$ is compact.
- (H4) $a: X \times X \times X \rightarrow \mathbb{R}$ is trilinear (i.e., $a(u, v, w)$ is linear with respect to each argument). In addition, for all $u, v, w \in X$,

$$|a(u, v, w)| \leq \text{const} \|u\|_Y \|v\|_Y \|w\|_X.$$

- (H5) $a(v, v, v) = 0$ for all $v \in D$.

Condition (H5) is crucial for obtaining a priori estimates.

Proposition 11. *Let $\eta > 0$. For a given functional $K \in \mathcal{H}^*$, the following hold true:*

- (i) Existence. *There exists a solution $v \in X$ to the equation*

$$\eta(v | \phi) + a(v, v, \phi) = K(\phi) \quad \text{for all } \phi \in D. \quad (87)$$

- (ii) Uniqueness. *If the norm of the functional K is sufficiently small (i.e., $\|K\|_{\mathcal{H}^*} < \delta$), then the solution v is unique.*
- (iii) Generic finiteness. *There exists an open, dense subset \mathcal{H}_0^* of \mathcal{H}^* such that, for each $K \in \mathcal{H}_0^*$, equation (87) has only a finite number of solutions $v \in X$.*

The intersection $\mathcal{H}_0^ \cap Z$ is open and dense in Z .*

Remark 12 (Elementary approach). If one wants to present an elementary approach to the Navier–Stokes equations in a lecture, it is convenient only to prove statements (i) and (ii) of Theorem 5.J on existence and uniqueness. In this connection, it suffices to use statements (i) and (ii) of Proposition 11. Thus, for the convenience of the lecturer, we divide the proof of Proposition 11 into two parts.

The proof of Proposition 11(iii), and hence the proof of Theorem 5.J(iii), is based on the sophisticated Smale principle.

By Proposition 8 in Section 5.16, we have the continuous embedding

$$Z \subseteq \mathcal{H}^*,$$

and Z is dense in \mathcal{H} . Moreover, it follows from the continuous embedding $X \subseteq \mathcal{H}$ that the embedding

$$\mathcal{H} \subseteq X^*$$

is also continuous.

The following existence and uniqueness proof will work not only for given $K \in \mathcal{H}^*$, but also for $K \in X^*$. Observe that $\|K\|_{X^*} \leq \text{const}\|K\|_{\mathcal{H}^*}$.

Proof of Proposition 11(i), (ii). The idea of the proof is to reduce the original problem (87) to an equivalent operator equation,

$$\eta v + Av = K_*, \quad v \in X, \quad (88)$$

by using the Riesz theorem. Next we will apply the Leray–Schauder principle to (88) in order to obtain statements (i) and (ii).

Step 1: The functional $K \in X^$.* By the Riesz theorem, there is a unique element K_* in the Hilbert space X such that

$$K(w) = (K_* | w) \quad \text{for all } w \in X$$

(cf. Section 2.10 of AMS Vol. 108). The duality map $J: X \rightarrow X^*$ defined through $J(K_*) := K$ is linear, bijective, and norm preserving (i.e., $\|J(K_*)\| = \|K\|$). Therefore, if X_0 is an open, dense subset of X , then the image $X_0^* := J(X_0)$ is an open, dense subset of X^* .

Step 2: The operator B . Let $u, v \in X$ be fixed. By (H4),

$$|a(u, v, w)| \leq \text{const}\|u\|_Y\|v\|_Y\|w\|_X \quad \text{for all } w \in X. \quad (89)$$

It follows from the Riesz theorem that there exists a unique element in X denoted by $B(u, v)$ such that

$$a(u, v, w) = (B(u, v) | w) \quad \text{for all } w \in X.$$

In addition,

$$\|B(u, v)\|_X = \sup |a(u, v, w)|, \quad (90)$$

where the supremum is taken over all $w \in X$, with $\|w\|_X \leq 1$.

Varying u and v , we obtain an operator $B: X \times X \rightarrow X$ that has the following properties:

(a) B is bilinear and bounded.

(b) For all $u, v \in X$,

$$\|B(u, v)\|_X \leq \text{const}\|u\|_Y\|v\|_Y. \quad (91)$$

In fact, statement (b) follows from (89) and (90). This implies

$$\|B(u, v)\|_X \leq \text{const}\|u\|_X\|v\|_X$$

because the embedding $X \subseteq Y$ is continuous. The bilinearity of B follows from the bilinearity of the map $(u, v) \mapsto a(u, v, w)$ for fixed w .

Step 3: The operator A . Define

$$Av := B(v, v) \quad \text{for all } v \in X.$$

The operator $A: X \rightarrow X$ has the following properties:

(a) A is locally Lipschitz continuous. This means that

$$\|Au - Av\|_X \leq \text{const} \cdot r \|u - v\|_X$$

for all $u, v \in X$ with $\|u\|_X, \|v\|_X \leq r$. Here, r denotes a fixed, but otherwise arbitrary, positive number.

- (b) A is compact.
- (c) $(Av | v) = 0$ for all $v \in X$.

Let us prove this. Recall that

$$\|v\|_Y \leq \text{const} \|v\|_X \quad \text{for all } v \in X \quad (92)$$

because the embedding $X \subseteq Y$ is compact.

Ad (a). Let $u, v \in X$ with $\|u\|_X, \|v\|_X \leq r$. Note that

$$Au - Av = B(u, u) - B(v, v) = B(u - v, u) + B(v, u - v).$$

By (91) and (92),

$$\begin{aligned} \|Au - Av\|_X &\leq \text{const} (\|u - v\|_Y \|u\|_Y + \|v\|_Y \|u - v\|_Y) \\ &\leq \text{const} (\|u - v\|_Y \|u\|_X + \|v\|_X \|u - v\|_Y). \end{aligned}$$

Hence we obtain the *key inequality*

$$\|Au - Av\|_X \leq \text{const} \cdot r \|u - v\|_Y. \quad (93)$$

This implies property (a), by (92).

Ad (b). Let (v_n) be a bounded sequence in X . Since the embedding $X \subseteq Y$ is compact, there exists a subsequence $(v_{n'})$ that converges in Y . Thus, $(v_{n'})$ is Cauchy in Y . By (93), the sequence $(Av_{n'})$ is Cauchy in X , and hence it is convergent in X .

Ad (c). Let $v \in X$. Since D is dense in X , there is a sequence (v_n) in D such that $v_n \rightarrow v$ in X as $n \rightarrow \infty$. By (H5),

$$(Av_n | v_n) = a(v_n, v_n, v_n) = 0 \quad \text{for all } n.$$

Letting $n \rightarrow \infty$, we obtain $(Av | v) = 0$, since the operator A is continuous.

Step 4: The equivalent operator equation. It follows from

$$\eta(v | \phi) + a(v, v, \phi) = K(\phi) \quad \text{for fixed } v \in X \text{ and all } \phi \in D \quad (94)$$

that

$$\eta(v | \phi) + (B(v, v) | \phi) = (K_* | \phi) \quad \text{for all } \phi \in D.$$

Recall that $Av = B(v, v)$. Then

$$(\eta v + Av - K_* | \phi) = 0 \quad \text{for all } \phi \in D.$$

Since the set D is dense in the Hilbert space X , we obtain the operator equation

$$\eta v + Av - K_* = 0, \quad v \in X. \quad (95)$$

Conversely, the operator equation (95) implies (94). Thus, equation (95) is equivalent to (94).

Step 5: The crucial a priori estimates. Set $\mu := \eta^{-1}$. Consider the modified operator equation

$$v = -\mu t Av + \mu t K_*, \quad v \in X, \quad (96)$$

where $t \in [0, 1]$. Note that, for $t = 1$, equation (96) is identical to (95). Suppose that v is a solution to (96). Then

$$(v | v) = -\mu t(Av | v) + \mu t(K_* | v).$$

Since $(Av | v) = 0$, $(v | v) = \mu t(K_* | v)$. Hence $\|v\|_X^2 \leq \mu \|K_*\|_X \|v\|_X$. This implies the a priori estimates

$$\|v\|_X \leq \eta^{-1} \|K_*\|_X \quad (97)$$

for any solution v to problem (96).

Step 6: The existence proof via the Leray–Schauder principle. It follows from (97) that, for given K_* , equation (95) has a solution, for $t = 1$, by the Leray–Schauder principle (cf. Theorem 1.D in Section 1.18 of AMS Vol. 108).

In this connection, note that the operator $v \mapsto \mu Av + \mu K_*$ is compact on X because A is compact.

Step 7: The uniqueness proof via local Lipschitz continuity. Choose $K_* \in X$. Suppose that v and w are solutions of the operator equation (95). By the a priori estimates (97),

$$\|v\|, \|w\| \leq r,$$

where $r := \eta^{-1} \|K_*\|_X$. From (95) we obtain

$$\eta(v - w) + Av - Aw = 0.$$

By the local Lipschitz continuity of the operator $A: X \rightarrow X$,

$$\eta\|v - w\|_X \leq \text{const} \cdot r\|v - w\|_X.$$

If the quantity $\eta^{-1}r = \eta^{-2} \|K_*\|_X$ is sufficiently small, then $v = w$.

Since $\|K_*\|_X = \|K\|_{X^*}$ and $\|K\|_{X^*} \leq \text{const} \|K\|_{\mathcal{H}^*}$, the quantity

$$\eta^{-2} \|K\|_{\mathcal{H}^*}$$

has to be sufficiently small. □

Proof of Proposition 11(iii). We will use the Smale principle.

Step 8: Further properties of the operator $A: X \rightarrow X$. We will show that

- (a) A is C^∞ .
- (b) For any $\eta > 0$, the operator $\eta I + A: X \rightarrow X$ is a nonlinear Fredholm operator of index zero.
- (c) The operator $\eta I + A: X \rightarrow X$ is proper.

Let us prove this.

Ad (a). Recall that $Av := B(v, v)$ on X . Since the operator $B: X \times X \rightarrow X$ is bilinear and bounded, it follows from Example 5 in Section 4.2 that A is C^∞ .

Ad (b). Since the operator $A: X \rightarrow X$ is compact, the F -derivative $A'(u): X \rightarrow X$ is also compact, for each $u \in X$ (cf. Problem 5.2). Set $C := \eta I + A$. Then the F -derivative of C at the point $u \in X$ is given by $C'(u) := \eta I + A'(u)$. Thus, $C'(u)$ represents a compact perturbation of the Fredholm operator $\eta I: X \rightarrow X$. Since this operator is bijective, $\text{ind}(\eta I) = 0$. By Theorem 5.E in Section 5.8, $\eta I + A$ is Fredholm of index zero.

Ad (c). Let M be a compact set in X . We have to show that the preimage $N := (\eta I + A)^{-1}(M)$ is compact.

The set M is bounded in X . By the a priori estimate in (97), the set N is bounded in X . Let (v_n) be a sequence in N . Since the operator $A: X \rightarrow X$ is compact, the set $A(M)$ is relatively compact. Thus, there exists a subsequence $(v_{n'})$ such that $Av_{n'} \rightarrow w$ as $n' \rightarrow \infty$. Set

$$b_{n'} := \eta v_{n'} + Av_{n'}. \quad (98)$$

Since the sequence $(b_{n'})$ lives in the compact set M , there is a convergent subsequence $b_{n''} \rightarrow b$ as $n'' \rightarrow \infty$. In addition, $b \in M$. By (98),

$$v_{n''} \rightarrow v \text{ as } n'' \rightarrow \infty,$$

where $\eta v = b - w$. Finally, since the operator $A: X \rightarrow X$ is continuous, it follows from (98) that

$$b = \eta v + Av.$$

This shows that $v \in N$. Thus, the set N is compact.

Step 9: Generic finiteness via the Smale principle. This principle tells us that there is an open, dense subset X_0 of X such that, for given $K_* \in X_0$, equation (88) has only a finite number of solutions (cf. Theorem 5.I in Section 5.15).

Define $X_0^* := J(X_0)$. By Step 3, the set X_0^* is open and dense in X^* .

Step 10: Since X_0^* is open and dense in X^* , and since the embedding $\mathcal{H} \subseteq X$ is continuous, the intersection set $\mathcal{H}_0^* := X_0^* \cap \mathcal{H}^*$ is dense in \mathcal{H}^* , by Proposition 9 in Section 5.16.

Step 11: Since \mathcal{H}_0^* is open and dense in \mathcal{H}^* , and since the embedding $\mathcal{H} \subseteq Z$ is continuous, the intersection $\mathcal{H}_0^* \cap Z^*$ is dense and open in Z^* , again by Proposition 9 in Section 5.16.

Step 12: The intersection $\mathcal{H}_0^* \cap Z$ is open and dense in Z . To see that this follows from Step 11, let $v \in Z$. Then the duality map $\mathcal{J}: Z \rightarrow Z^*$ assigns a functional $v^* := \mathcal{J}(v)$ in Z^* to the point v such that

$$v^*(u) = (v | u)_Z \quad \text{for all } u \in Z.$$

The restriction of the linear continuous functional $v^*: Z \rightarrow \mathbb{R}$ to \mathcal{H} is a linear continuous functional $v^*: \mathcal{H} \rightarrow \mathbb{R}$, that is,

$$v^* \in \mathcal{H}^*.$$

Similarly, we write

$$v \in \mathcal{H}^*.$$

By Propositions 8 and 9 in Section 5.16, this corresponds to

$$Z^* \subseteq \mathcal{H}^* \quad \text{and} \quad Z \subseteq \mathcal{H}^*.$$

This way, we identify $\mathcal{H}_0^* \cap Z$ with $\mathcal{H}_0^* \cap Z^*$.

Finally, since the duality map is a normisomorphism between Z and Z^* , it sends open and dense sets in Z onto open and dense sets in Z^* . \square

The proof of Proposition 11 has been finished.

5.17.6 Proof of Theorem 5.J

We have to show only that the assumptions (H1) through (H5) of Proposition 11 are satisfied. However, this can be easily done. To this end, choose

$$Y := L_4(G) \times L_4(G) \times L_4(G)$$

along with the norm

$$\|v\|_Y := \sum_{j=1}^3 \|u_j\|_4 = \sum_{j=1}^3 \left(\int_G u_j^4 dx \right)^{\frac{1}{4}}.$$

We will use the following two simple *key observations*:

(a) Let $u, v \in L_4(G)$, and $w \in \overset{\circ}{W}_2^1(G)$. Then, $\partial_k w \in L_2(G)$. By the Hölder inequality for three factors,

$$\left| \int_G uv \partial_k w dx \right| \leq \left(\int_G u^4 dx \right)^{\frac{1}{4}} \left(\int_G v^4 dx \right)^{\frac{1}{4}} \left(\int_G (\partial_k w)^2 dx \right)^{\frac{1}{2}} \quad (99)$$

(cf. Problem 5.9b).

(b) If $v_j \in C_0^\infty(G)$ and $\partial_j v_j = 0$, then integration by parts yields

$$2 \int_G v_j v_m \partial_j v_m dx = \int_G v_j \partial_j (v_m^2) dx = - \int_G v_m^2 \partial_j v_j dx = 0.$$

Ad (H1), (H2), and (H3). This follows from Section 5.16.

Ad (H4). Recall that

$$\|v\|_X = \left(\int_G (\partial_j v_j)^2 dx \right)^{\frac{1}{2}}.$$

By the key observation (a),

$$|a(u, v, w)| = \left| \int_G u_j v_m \partial_j w_m dx \right| \leq \text{const} \|u\|_Y \|v\|_Y \|w\|_X$$

for all $u, v, w \in X$.

Ad (H5). Observe that assumption (H5) coincides with (b).

The proof of Theorem 5.J has been finished. \square

5.17.7 A Result from Modern Vector Calculus

In the following two sections, let us use the language of distribution theory. As before, G denotes a nonempty, bounded, open, and connected set in \mathbb{R}^3 . Let U be a distribution, that is, $U \in \mathcal{D}'(G)$. Recall from Section 2.8.3 in AMS Vol. 108 that the derivative $\partial_j U$ of U is defined through

$$(\partial_j U)(\phi) = -U(\partial_j \phi) \quad \text{for all test functions } \phi \in C_0^\infty(G).$$

In contrast to classical functions, distributions are mathematical objects that possess derivatives of arbitrary order. If the function $u: G \rightarrow \mathbb{R}$ is locally integrable, then we set

$$U(\phi) := \int_G u \phi dx \quad \text{for all } \phi \in C_0^\infty(G).$$

In this sense, locally integrable functions can be regarded as distributions. It is convenient to denote both the distribution U and the function by the same symbol u .

We now consider the basic equation

$$-\nabla p = \mathbf{P} \quad \text{on } G \tag{100*}$$

from vector calculus. In hydrodynamics,²⁸ \mathbf{P} is the force density generated by the pressure p . Using components in a Cartesian coordinate system, equation (100*) is equivalent to the following equation:

$$-\partial_m p = P_m \quad \text{on } G, \quad m = 1, 2, 3. \tag{100}$$

²⁸In mechanics, p denotes the potential to the force \mathbf{P} . Moreover, in electrodynamics, \mathbf{P} is the electric field vector to the electrostatic potential p .

This equation was considered from a classical point of view in Section 5.17.2. Let us now investigate equation (100) from the distribution theory point of view. First we formulate a simple necessary solvability condition. Recall that we denote the dual space $\overset{\circ}{W}_2^1(G)^*$ by $W_2^{-1}(G)$.

Proposition 13 (Necessary solvability condition). *Let $p, P_m \in \mathcal{D}'(G)$ for $m = 1, 2, 3$. Suppose that equation (100) is fulfilled. Then the following are true:*

(i) *For all $\phi \in D$,*

$$P_m(\phi_m) = 0. \quad (101)$$

(ii) *If $p \in L_2(G)$, then $P_m \in W_2^{-1}(G)$ for all m .*

More precisely, statement (ii) means that P_m can be uniquely extended to a functional $P_m \in W_2^{-1}(G)$.

Proof. Ad (i). Recall that $\phi \in D$ implies $\partial_m \phi_m = 0$. Thus, from (100) we obtain

$$P_m(\phi_m) = p(\partial_m \phi_m) = 0.$$

Ad (ii). It follows from (100) that, for all $\psi \in C_0^\infty(G)$,

$$\begin{aligned} |P_m(\psi)| &= |p(\partial_m \psi)| = \left| \int_G p \partial_m \psi dx \right| \\ &\leq \left(\int_G p^2 dx \right)^{\frac{1}{2}} \left(\int_G (\partial_m \psi)^2 dx \right)^{\frac{1}{2}}, \end{aligned}$$

by the Schwarz inequality. Hence

$$|P_m(\psi)| \leq \text{const} \|\psi\|_{1,2} \quad \text{for all } \psi \in C_0^\infty(G).$$

Observe that $C_0^\infty(G)$ is dense in the Hilbert space $H := \overset{\circ}{W}_2^1(G)$. Therefore, P_m can be extended to a linear continuous map $P_m: H \rightarrow \mathbb{R}$, by the extension principle (cf. Section 3.6 in AMS Vol. 108). \square

It is remarkable that the simple conditions from Proposition 13 are strong enough to ensure the existence of a pressure function p .

Proposition 14 (Sufficient solvability condition). *Suppose that the boundary ∂G of the region G is sufficiently regular.²⁹ If we are given functionals $P_j \in W_2^{-1}(G)$, $j = 1, 2, 3$, condition (101) is satisfied.*

²⁹For example, suppose that ∂G is a two-dimensional C^1 -manifold, where G lies locally on one side of ∂G .

Then equation (100) has a solution $L_2(G)$ that is unique up to a constant.

The proof will be given in Problem 5.18b and will be based on the closed range theorem.

5.17.8 Determination of the Pressure

Let us again consider the equation of motion for a viscous fluid

$$-\eta\Delta\mathbf{v} + \nabla(\mathbf{v} \otimes \mathbf{v}) = \mathbf{K} - \nabla p \quad \text{on } G.$$

Using components in a Cartesian coordinate system, this means that

$$-\eta\partial_j\partial_j v_m + \partial_j(v_j v_m) = K_m - \partial_m p \quad \text{on } G, \quad (102)$$

where $m = 1, 2, 3$. Suppose that v_j , $v_j v_m$, K_m , and p are distributions.³⁰ Then equation (102) tells us that

$$-\eta(\partial_j\partial_j v_m)(\psi) + (\partial_j(v_j v_m))(\psi) = K_m(\psi) - (\partial_m p)(\psi)$$

for all $\psi \in C_0^\infty(G)$. Replacing ψ with ϕ_m and summing over m , we obtain

$$-\eta(\partial_j\partial_j v_m)(\phi_m) + (\partial_j(v_j v_m))(\phi_m) = K_m(\phi_m) - (\partial_m p)(\phi_m) \quad (103)$$

for all $\phi_m \in C_0^\infty(G)$. Hence

$$\eta(\partial_j v_m)(\partial_j \phi_m) - (v_m v_j)(\partial_j \phi_m) = K_m(\phi_m) + p(\partial_m \phi_m) \quad (104)$$

for all $\phi_m \in C_0^\infty(G)$. Recall that $\partial_m \phi_m = 0$ if $\phi \in D$. Thus,

$$\eta(\partial_j v_m)(\partial_j \phi_m) - (v_j v_m)(\partial_j \phi_m) = K_m(\phi_m) \quad (105)$$

for all $\phi \in D$. This is precisely the generalized problem from Definition 2.

We now want to reverse the preceding argument. Observe that the following proof resembles the classical argument used in Section 5.17.2.

Proof of Corollary 6. We are given the force functional $K \in \mathcal{H}^*$. Let the velocity field $v \in X$ be a generalized solution to the equation of motion, in the sense of Theorem 5.J.

Recall that $\mathcal{H} = \overset{\circ}{W}_2^1(G) \times \overset{\circ}{W}_2^1(G) \times \overset{\circ}{W}_2^1(G)$ and that $X \subseteq \mathcal{H}$. Hence $\mathcal{H}^* \subseteq X^*$. Denote the restriction of K to the m th factor $W_2^1(G)$ of \mathcal{H} by K_m . Then

$$K(\phi) = K_m(\phi_m) \quad \text{for all } \phi \in \mathcal{H}. \quad (106)$$

³⁰Observe that a general product for distributions does not exist (cf. Oberguggenberger (1992)). For example, this fact is responsible for serious mathematical difficulties arising in quantum field theory. However, in the present case we can assume that v_j and $v_j v_m$ are locally integrable functions.

Since $v \in X$, $v_m \in \overset{\circ}{W}_2^1(G)$. Hence $v_m \in L_2(G)$. Thus, the product $v_j v_m$ is integrable over G , by the Schwarz inequality. Consequently, $v_m v_j$ is a distribution.

Step 1: Existence of the pressure p . Define

$$P_m := -\eta \partial_j \partial_j v_m + \partial_j(v_j v_m) - K_m. \quad (107)$$

Then, P_m is a distribution, that is, $P_m \in \mathcal{D}'(G)$. Consider now the pressure equation

$$-\partial_m p = P_m, \quad m = 1, 2, 3. \quad (108)$$

Since v is a generalized solution to the equation of motion, equation (105) holds true. Hence

$$P_m(\phi_m) = 0 \quad \text{for all } \phi \in D.$$

We want to show that the following conditions are met:

- (a) $K_m \in W_2^{-1}(G)$,
- (b) $\partial_j \partial_j v_m \in W_2^{-1}(G)$,
- (c) $\partial_j(v_j v_m) \in W_2^{-1}(G)$.

Ad (a). By (106),

$$|K_m(\psi)| \leq \text{const} \|\psi\|_{1,2} \quad \text{for all } \psi \in \overset{\circ}{W}_2^1(G).$$

Ad (b). For all $\psi \in C_0^\infty(G)$,

$$\begin{aligned} |(\partial_j \partial_j v_m)(\psi)| &= |v_m(\partial_j \partial_j \psi)| = \int_G v_m \partial_j \partial_j \psi dx \\ &= \int_G \partial_j v_m \partial_j \psi dx \\ &\leq \left(\int_G (\partial_j v_m)^2 dx \right)^{\frac{1}{2}} \left(\int_G (\partial_j \psi)^2 dx \right)^{\frac{1}{2}} \\ &\leq \text{const} \|\psi\|_{1,2}. \end{aligned}$$

Therefore, $\partial_j \partial_j v_m$ can be uniquely extended to a linear continuous functional on $\overset{\circ}{W}_2^1(G)$, by the same argument as in the proof of Proposition 13.

Ad (c). For all $\psi \in C_0^\infty(G)$,

$$\begin{aligned} |(\partial_j(v_j v_m))(\psi)| &= |(v_j v_m)(\partial_j \psi)| = \left| \int_G v_j v_m \partial_j \psi dx \right| \\ &\leq \|v_j\|_4 \|v_m\|_4 \|\phi\|_{1,2} \\ &\leq \text{const} \|\phi\|_{1,2}, \end{aligned}$$

by (99). This proves (c).

It follows from (a) through (c) that $P_m \in W_2^{-1}(G)$. By Proposition 14, $p \in L_2(G)$.

The remaining statements of Corollary 6 follow from Proposition 5. \square

Problems

5.1. A special C^1 -function. Prove that the function

$$H(s, w, p) := \begin{cases} s^{-1}F(sb + sw, p) & \text{if } s \neq 0 \\ F_u(0, p)(b + w) & \text{if } s = 0 \end{cases} \quad (109)$$

is C^1 on an open neighborhood of $(0, 0, 0)$ provided F is C^2 on an open neighborhood of $(0, 0)$, where F satisfies the assumptions of Theorem 5.H with $u_0 = 0$, $p_0 = 0$.

Solution: Recall that $F(0, p) \equiv 0$. By the Taylor theorem from Section 4.5,

$$\begin{aligned} F(sb + sw, p) &= F(sb + sw, p) - F(0, p) \\ &= sF_u(0, p)(b + w) + \frac{s^2}{2} \int_0^1 F_{uu}(\tau(sb + sw), p)(b + w)^2 d\tau \\ &= sF_u(0, p)(b + w) + \frac{s^2}{2} F_{uu}(0, p)(b + w)^2 + o(s^2), \quad s \rightarrow 0, \end{aligned} \quad (110)$$

and

$$\begin{aligned} F_u(sb + sw, p)(b + w) &= F_u(0, p)(b + w) + s \int_0^1 F_{uu}(\tau(sb + sw), p)(b + w)^2 d\tau \\ &= F_u(0, p)(b + w) + sF_{uu}(0, p)(b + w)^2 + o(s), \quad s \rightarrow 0, \end{aligned} \quad (111)$$

as well as

$$\begin{aligned} F_p(sb + sw, p) &= F_p(0, p) + s \int_0^1 F_{up}(\tau(sb + sw), p)(b + w) d\tau \\ &= sF_{up}(0, p)(b + w) + o(s), \quad s \rightarrow 0. \end{aligned} \quad (112)$$

It follows from (110) that H is continuous at $(0, w, p)$ and

$$H_s(0, w, p) = \lim_{s \rightarrow 0} s^{-1}(H(s, w, p) - H(0, w, p)) = 2^{-1}F_{uu}(0, p)(b + w)^2.$$

By (109),

$$H_s(s, w, p) = -s^{-2}F(sb + sw, p) + s^{-1}F_u(sb + sw, p)(b + w) \quad \text{if } s \neq 0.$$

Thus, it follows from (110) and (111) that, for $s \neq 0$,

$$H_s(s, w, p) = 2^{-1} F_{uu}(0, p) + o(1), \quad s \rightarrow 0.$$

Hence

$$H_s(s, w, p) \rightarrow H_s(0, w, p) \quad \text{as } s \rightarrow 0.$$

By (109),

$$H_w(s, w, p) = F_u(sb + sw, p)w \quad \text{for small } |s|.$$

Finally,

$$H_p(s, w, p) = \begin{cases} s^{-1} F_p(sb + sw, p) & \text{if } s \neq 0 \\ F_{pu}(0, p)(b + w) & \text{if } s = 0. \end{cases}$$

According to Problem 4.10, $F''(u, p)(h, k) = F''(u, p)(k, h)$. Hence $F_{up}(u, p) = F_{pu}(u, p)$. Thus, by (112),

$$H_p(s, w, p) \rightarrow H_p(0, w, p) \quad \text{as } s \rightarrow 0.$$

Summarizing, we find that the partial derivatives H_s , H_w , and H_p are continuous on a neighborhood of $(0, 0, 0)$. By Problem 4.11, H is C^1 on a neighborhood of $(0, 0, 0)$.

5.2. Compact operators. Let $A: X \rightarrow X$ be a compact C^1 -operator on the Banach space X over \mathbb{K} . Show that, for each $u \in X$, the Fréchet derivative $A'(u): X \rightarrow X$ is compact.

Hint: Cf. Zeidler (1986), Vol. 1, Proposition 7.33.

5.3. Nonlinear Fredholm operators. Show that the operator

$$B + C: X \rightarrow X$$

is a nonlinear Fredholm operator of index zero on the Banach space X over \mathbb{K} provided the following conditions are met:

- (i) The operator $B: X \rightarrow X$ is linear, continuous, and *bijective*.
- (ii) The operator $C: X \rightarrow X$ is *compact* and C^1 .

Hint: Use the same argument as in the proof of Proposition 11 in Section 5.17.5.

5.4.* The generalized Jordan normal form. Let

$$C: X \rightarrow X$$

be a linear *compact* operator on the Banach space X over \mathbb{K} . We want to decompose the space X into *invariant subspaces*, with respect to C , that are as *small* as possible. We are given the number $\lambda \in \mathbb{K}$, $\lambda \neq 0$. Set

$$A := \lambda I - C.$$

We already know that if λ is not an eigenvalue of C , then the operator

$$\lambda I - C: X \rightarrow X$$

is *bijective*, and the inverse operator $(\lambda I - C)^{-1}: X \rightarrow X$ is continuous.

Now suppose that λ is an *eigenvalue* of A . Show that

- (i) There exists a natural number n such that³¹

$$N(A^1) \subset N(A^2) \subset \cdots \subset N(A^n) = N(A^{n+1}) = N(A^{n+2}) = \cdots$$

and

$$R(A^1) \supset R(A^2) \supset \cdots \supset R(A^n) = R(A^{n+1}) = R(A^{n+2}) = \cdots.$$

- (ii) The space X decomposes into the direct sum

$$X = N(A^n) \oplus R(A^n),$$

where $\dim N(A^n) < \infty$. The closed linear subspaces $N(A^n)$ and $R(A^n)$ of X are *invariant* under the operator C .

- (iii) The operator

$$\lambda I - C: R(A^n) \rightarrow R(A^n)$$

is *bijective*, and the corresponding inverse operator is continuous on $R(A^n)$.

- (iv) There exist linear subspaces L_1, \dots, L_r of $N(A^n)$ such that we have the direct sum decomposition

$$N(A^n) = L_1 \oplus \cdots \oplus L_r.$$

Each of the spaces L_j is *invariant* under C . Moreover, for each L_j , there exists a basis u_1, \dots, u_m such that

$$\begin{aligned} Cu_s &= \lambda u_s + u_{s+1}, & s &= 1, \dots, m-1, \\ Cu_m &= \lambda u_m. \end{aligned}$$

This is called the *Jordan normal form* for the operator C . Note that if $\dim L_j = m = 1$, then L_j is a one-dimensional eigenspace of the operator C .

Hint: Cf. Riesz and Nagy (1955), Sections 80 and 89.

Next we want to study the following three important inequalities:

$$\begin{aligned} \text{Young inequality} &\Rightarrow \text{Hölder inequality} \\ &\Rightarrow \text{Minkowski inequality}. \end{aligned}$$

³¹Recall our convention that $K \subset M$ means both $K \subseteq M$ and $K \neq M$.

5.5. The Young inequality. Show that

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q} \quad \text{for all } a, b \geq 0, \quad (113)$$

where

$$\frac{1}{p} + \frac{1}{q} = 1, \quad 1 < p, q < \infty. \quad (114)$$

Solution: If $p = q = 2$, then (113) follows from $(a - b)^2 \geq 0$. In the general case, we consider the function

$$F(a) = \frac{a^p}{p} + \frac{b^q}{q} - ab,$$

for fixed $b > 0$. Note that

$$F(0) > 0, \quad F(b^{\frac{q}{p}}) = 0, \quad \text{and} \quad \lim_{a \rightarrow +\infty} F(a) = +\infty.$$

Hence F has a minimum on $[0, \infty[$. Thus, there exists a number $a_0 > 0$ such that

$$F(a) \geq F(a_0) \quad \text{for all } a \in [0, \infty[.$$

From $F'(a_0) = 0$ it follows that $a_0 = b^{\frac{q}{p}}$, and hence $F(a_0) = 0$. This is (113).

5.6. The Hölder inequality in \mathbb{R}^N . Assume (114). Show that

$$\left| \sum_{j=1}^N \xi_j \eta_j \right| \leq \left(\sum_{j=1}^N |\xi_j|^p \right)^{\frac{1}{p}} \left(\sum_{j=1}^N |\eta_j|^q \right)^{\frac{1}{q}} \quad (115)$$

for all $\xi_j, \eta_j \in \mathbb{C}$, $j = 1, \dots, N$.

Solution: It follows from the Young inequality (113) that

$$\frac{|\xi_j| |\eta_j|}{(\sum_j |\xi_j|^p)^{\frac{1}{p}} (\sum_j |\eta_j|^q)^{\frac{1}{q}}} \leq \frac{|\xi_j|^p}{p \sum_j |\xi_j|^p} + \frac{|\eta_j|^q}{q \sum_j |\eta_j|^q}.$$

Summing over $j = 1, \dots, N$ and using (114), we get (115).

5.7. The Minkowski inequality. Set

$$\|x\|_p := \left(\sum_{j=1}^N |\xi_j|^p \right)^{\frac{1}{p}},$$

where $x = (\xi_1, \dots, \xi_N)$. Show that

$$\|x + y\|_p \leq \|x\|_p + \|y\|_p \quad \text{for all } x, y \in \mathbb{C}^N, \quad (116)$$

where $1 \leq p < \infty$.

Solution: Let $p > 1$. By the Hölder inequality (115),

$$\begin{aligned} \sum_{j=1}^N |\xi_j + \eta_j|^p &\leq \sum_{j=1}^N |\xi_j + \eta_j|^{p-1} (|\xi_j| + |\eta_j|) \\ &\leq \left(\sum_{j=1}^N |\xi_j + \eta_j|^{(p-1)q} \right)^{\frac{1}{q}} \left(\sum_{j=1}^N |\xi_j|^p \right)^{\frac{1}{p}} \\ &\quad + \left(\sum_{j=1}^N |\xi_j + \eta_j|^{(p-1)q} \right)^{\frac{1}{q}} \left(\sum_{j=1}^N |\eta_j|^p \right)^{\frac{1}{p}}. \end{aligned}$$

Since $(p-1)q = p$ and $\frac{1}{q} = 1 - \frac{1}{p}$, we get (116).

5.8. The Banach space $\ell_p^{\mathbb{K}}$. Let \mathbb{K}^∞ denote the linear space of all the sequences $(u_n)_{n \geq 1}$, where $u_n \in \mathbb{K}$ for all $n \in \mathbb{N}$ (cf. Problem 1.5 in AMS Vol. 108). Moreover, let $\ell_p^{\mathbb{K}}$ denote the set of all $(u_n) \in \mathbb{K}^\infty$ such that

$$\|(u_n)\|_p := \left(\sum_{j=1}^{\infty} |u_n|^p \right)^{\frac{1}{p}} < \infty,$$

where $1 \leq p < \infty$. The space $\ell_\infty^{\mathbb{K}}$ has been introduced in Problem 1.5 in AMS Vol. 108. Using Problem 5.7, show that

- (i) $\ell_p^{\mathbb{K}}$ is a separable Banach space over \mathbb{K} if $1 \leq p < \infty$.
- (ii*) For $1 < p, q < \infty$ and $p^{-1} + q^{-1} = 1$,

$$(\ell_p^{\mathbb{K}})^* \cong \ell_q^{\mathbb{K}}.$$

That is, the dual space $(\ell_p^{\mathbb{K}})_\mathbb{K}^*$ is *normisomorphic*³² to $\ell_q^{\mathbb{K}}$. More precisely, let $y \in \ell_q^{\mathbb{K}}$. Setting

$$f(x) := \sum_{j=1}^{\infty} \eta_j \xi_j \quad \text{for all } x \in \ell_p^{\mathbb{K}}, \quad (117)$$

³²Recall that the Banach space X over \mathbb{K} is *normisomorphic* to the Banach space Y over \mathbb{K} iff there exists a linear bijective map $\phi: X \rightarrow Y$ such that $\|\phi(u)\| = \|u\|$ for all $u \in X$.

we get a linear continuous functional f on $\ell_p^{\mathbb{K}}$ (i.e., $f \in (\ell_p^{\mathbb{K}})^*$) and

$$\|f\| = \|y\|_q. \quad (118)$$

Each functional $f \in (\ell_q^{\mathbb{K}})^*$ can be obtained this way, where y is uniquely determined by f .

(iii) $\ell_p^{\mathbb{K}}$ is *reflexive* if $1 < p < \infty$.

(iv*) $(\ell_1^{\mathbb{K}})^* \cong \ell_{\infty}^{\mathbb{K}}$, in the sense of (ii) with $p = 1$ and $q = \infty$.

(v*) $\ell_1^{\mathbb{K}}$ and $\ell_{\infty}^{\mathbb{K}}$ are *not* reflexive.

(vi*) $\ell_{\infty}^{\mathbb{K}}$ is *not* separable.

Hint: Cf. Köthe (1960), Vol. 1, Section 14.

5.9. *The fundamental Banach space $L_p^{\mathbb{K}}(G)$, $1 \leq p < \infty$.* Let G be a nonempty open subset of \mathbb{R}^N , $N \geq 1$. Set

$$\|u\|_p := \left(\int_G |u(x)|^p dx \right)^{\frac{1}{p}}, \quad 1 \leq p < \infty.$$

Let $L_p^{\mathbb{K}}(G)$ denote the set of all measurable functions $u: G \rightarrow \mathbb{K}$ such that³³ $\|u\|_p < \infty$.

5.9a. Basic properties. Show that

(i) If $1 < p, q < \infty$ and $p^{-1} + q^{-1} = 1$, then we have the *Hölder inequality*:

$$\left| \int_G uv dx \right| \leq \|u\|_p \|v\|_q, \quad (119)$$

for all $u \in L_p^{\mathbb{K}}(G)$ and $v \in L_p^{\mathbb{K}}(G)$.

(ii) For all $u, v \in L_p^{\mathbb{K}}(G)$ with $1 \leq p < \infty$, we have the *Minkowski inequality*:

$$\|u + v\|_p \leq \|u\|_p + \|v\|_p. \quad (120)$$

(iii) Let $1 \leq p < \infty$. Then $L_p^{\mathbb{K}}(G)$ becomes a separable *Banach space* over \mathbb{K} with the norm $\|\cdot\|_p$ once we identify any two functions that differ only on a set of measure zero on G . Moreover, $C_0^\infty(G)_{\mathbb{K}}$ is *dense* in $L_p^{\mathbb{K}}(G)$.

³³For brevity we write $L_p(G)$ instead of $L_p^{\mathbb{K}}(G)$ in the case where $\mathbb{K} = \mathbb{R}$.

- (iv) Assume $1 \leq p < \infty$. Let (u_n) be a sequence in $L_p(G)$ such that $u_n \rightarrow u$ in $L_p^{\mathbb{K}}(G)$ as $n \rightarrow \infty$. Then there are a subsequence $(u_{n'})$ and a function $v \in L_p(G)$ such that, for almost all $x \in G$,

$$u_{n'}(x) \rightarrow u(x) \quad \text{as } n' \rightarrow \infty,$$

and $\sup_{n'} |u_{n'}(x)| \leq v(x)$.

Solution: Ad (i). We may assume that $\|u\|_p = \|v\|_p = 1$. Otherwise we replace u and v with λu and μv , respectively. Integration of the *Young inequality*,

$$|uv| \leq \frac{|u|^p}{p} + \frac{|v|^q}{q},$$

over G yields $\int_G |uv| dx \leq 1$. This is (119).

Ad (ii). Use a similar argument as in Problem 5.7.

Ad (iii). Use the same arguments as for the space $L_2^{\mathbb{K}}(G)$ in Section 2.2 of AMS Vol. 108 .

Ad (iv). Use the same argument as in the proof given in Section 2.2.1 of AMS Vol. 108. For the construction of the function v , see Kufner et al. (1977), Section 2.8.

5.9b. The Hölder inequality for three factors (special case). Show that

$$\left| \int_G uvwdx \right| \leq \left(\int_G u^4 dx \right)^{\frac{1}{4}} \left(\int_G v^4 dx \right)^{\frac{1}{4}} \left(\int_G w^2 dx \right)^{\frac{1}{2}} \quad (121)$$

for all $u, v \in L_4^{\mathbb{K}}(G)$ and $w \in L_2^{\mathbb{K}}(G)$.

This inequality plays an important role in the existence proof for the stationary Navier–Stokes equations (cf. Section 5.17.6).

Solution: Since $|u|^2, |v|^2 \in L_2^{\mathbb{K}}(G)$, it follows from the Hölder inequality for two factors from (119) that

$$\int_G |u|^2 |v|^2 dx \leq \left(\int_G |u|^4 dx \right)^{\frac{1}{2}} \left(\int_G |v|^4 dx \right)^{\frac{1}{2}}. \quad (122)$$

Hence $uv, w \in L_2^{\mathbb{K}}(G)$. Thus, again by the Hölder inequality for two factors,

$$\left| \int_G (uv)wdx \right| \leq \left(\int_G |uv|^2 dx \right)^{\frac{1}{2}} \left(\int_G |w|^2 dx \right)^{\frac{1}{2}}.$$

Using (122), we obtain (121).

5.9c. The Hölder inequality for n factors. Let $1 < p_1, \dots, p_n < 1$, where

$$\sum_{j=1}^n \frac{1}{p_j} = 1,$$

and $n = 2, 3, \dots$. Show that

$$\left| \int_G \prod_{j=1}^n u_j dx \right| \leq \prod_{j=1}^n \left(\int_G |u_j|^{p_j} dx \right)^{\frac{1}{p_j}}$$

for all $u_j \in L_{p_j}^{\mathbb{K}}(G)$, $j = 1, \dots, n$.

Hint: Use the same argument as in Problem 5.9b.

5.10.* The dual space. Let $1 < p, q < \infty$ and $p^{-1} + q^{-1} = 1$. Then

$$L_p^{\mathbb{K}}(G)^* \cong L_q^{\mathbb{K}}(G),$$

that is, the dual space $(L_p^{\mathbb{K}}(G))^*$ is *normisomorphic* to $L_q^{\mathbb{K}}(G)$. More precisely, if $v \in L_q^{\mathbb{K}}(G)$, then

$$F(u) := \int_G u(x)v(x)dx \quad \text{for all } u \in L_p^{\mathbb{K}}(G)$$

defines a linear continuous functional F on $L_p^{\mathbb{K}}(G)$ with

$$\|F\| = \|v\|_q.$$

Each $F \in L_p^{\mathbb{K}}(G)^*$ can be obtained this way, where $v \in L_q^{\mathbb{K}}(G)$ is uniquely determined by F .

Hint: Study the proof in Kufner et al. (1977).

5.11. Reflexivity. If $1 < p < \infty$, then the space $L_p^{\mathbb{K}}(G)$ is reflexive.

Hint: Use Problem 5.10.

5.12. The Sobolev space $W_p^m(G)_{\mathbb{K}}$. Let G be a nonempty open subset of \mathbb{R}^N , let $1 \leq p < \infty$, and let $m = 1, 2, \dots$. Set

$$\|u\|_{m,p} := \left(\sum_{0 \leq |\alpha| \leq m} (\|\partial^\alpha u\|_p)^p \right)^{\frac{1}{p}},$$

meaning that we sum over all the partial derivatives of u up to order m .

By definition, the space $W_p^m(G)_{\mathbb{K}}$ consists of all the functions

$$u \in L_p^{\mathbb{K}}(G)$$

with

$$\partial^\alpha u \in L_p^{\mathbb{K}}(G) \quad \text{for all } \alpha: 0 < |\alpha| \leq m,$$

where the partial derivatives are to be understood in the sense of generalized functions.

Explicitly, this means the following. We have $u \in W_p^m(G)_{\mathbb{K}}$ iff $u \in L_p^{\mathbb{K}}(G)$, and for each $\alpha: 0 < |\alpha| \leq m$ there exists a function denoted by $\partial^\alpha u$ such that $\partial^\alpha u \in L_p^{\mathbb{K}}(G)$ and

$$\int_G u \partial^\alpha \phi \, dx = (-1)^{|\alpha|} \int_G (\partial^\alpha u) \phi \, dx,$$

for all $\phi \in C_0^\infty(G)$. Show that

- (i) $W_p^m(G)_{\mathbb{K}}$ becomes a Banach space over \mathbb{K} with the norm $\|\cdot\|_{m,p}$ once we identify any two functions that differ only on a set of measure zero on G .
- (ii) $W_p^m(G)_{\mathbb{K}}$ is *reflexive* if $1 < p < \infty$.

Hints: Ad (i). Use the same arguments as for $W_2^1(G)$ in Section 2.2 in AMS Vol. 108.

Ad (ii). Observe that $W_p^m(G)_{\mathbb{K}}$ is *normisomorphic* to a closed linear subspace of the product space

$$L_p^{\mathbb{K}}(G) \times \cdots \times L_p^{\mathbb{K}}(G),$$

by means of the map

$$u \mapsto (\partial^\alpha u)_{|\alpha| \leq m}.$$

Furthermore, use Problem 5.11 and the following two facts:

- (a) Products of reflexive Banach spaces are again reflexive.
- (b) Closed linear subspaces of reflexive Banach spaces are again reflexive.

The Sobolev spaces $W_p^m(G)_{\mathbb{K}}$ represent the basic tool for the modern theory of linear and nonlinear partial differential equations.

This can be found in Zeidler (1986), Vols. 2ff.

5.13. Approximation of compact operators. Let $C_n, C: X \rightarrow Y$ be linear continuous operators, where X and Y are Banach spaces over \mathbb{K} . Show that if

$$\lim_{n \rightarrow \infty} \|C_n - C\| = 0$$

and C_n is compact for all n , then C is also *compact*.

5.14. Integral operators. Let $-\infty < a < b < \infty$. Define

$$(Au)(x) := \int_a^b \mathcal{A}(x, y)u(y)dy \quad \text{for all } x \in]a, b[,$$

where the function $\mathcal{A}:]a, b[\times]a, b[\rightarrow \mathbb{R}$ is measurable and

$$\int_a^b \int_a^b |\mathcal{A}(x, y)|^p dx dy < \infty \quad (123)$$

for fixed $p: 1 < p < \infty$. Set

$$X := L_q(a, b),$$

where $p^{-1} + q^{-1} = 1$. By Problem 5.10,

$$X^* = L_p(a, b),$$

in the sense of a normisomorphism. Hence $X^{**} = X$. Show that

(i) The operator $A: X \rightarrow X^*$ is linear and continuous with

$$\|A\| \leq \left(\int_a^b \int_a^b |\mathcal{A}(x, y)|^p dx dy \right)^{\frac{1}{p}}. \quad (124)$$

(ii) The operator $A: X \rightarrow X^*$ is compact.

(iii) The dual operator $A^T: X \rightarrow X^*$ is given by $A^T = B$, where

$$(Bv)(x) := \int_a^b \mathcal{A}(y, x)v(y)dy \quad \text{for all } x \in]a, b[. \quad (125)$$

(iv) If $p = 2$, then the adjoint operator $A^*: X \rightarrow X$ is given by $A^* = B$.

Solution: Ad (i). By the *Fubini theorem* (cf. the appendix to AMS Vol. 108), it follows from (123) that

$$\int_a^b |\mathcal{A}(x, y)|^p dy < \infty$$

for almost all $x \in]a, b[$. Let $u \in X$. By the Hölder inequality,

$$|(Au)(x)| \leq \left(\int_a^b |\mathcal{A}(x, y)|^p dy \right)^{\frac{1}{p}} \left(\int_a^b |u(y)|^q dy \right)^{\frac{1}{q}}.$$

Hence

$$\|Au\|_p = \left(\int_a^b |(Au)(x)|^p dx \right)^{\frac{1}{p}} \leq \left(\int_a^b \int_a^b |\mathcal{A}(x, y)|^p dx dy \right)^{\frac{1}{p}} \|u\|_q.$$

Ad (ii). Set $G :=]a, b[\times]a, b[$. Since $C_0^\infty(G)$ is dense in $L_p(G)$, for each $n \in \mathbb{N}$ there is a function $\mathcal{A}_n \in C_0^\infty(G)$ such that

$$\|\mathcal{A}_n - \mathcal{A}\| \leq \left(\int_a^b \int_a^b |\mathcal{A}_n(x, y) - \mathcal{A}(x, y)|^p dx dy \right)^{\frac{1}{p}} < n^{-1}.$$

It follows as in the proof of Lemma 3 in Section 4.4 of AMS Vol. 108 that the operator $A_n: X \rightarrow X^*$ corresponding to the kernel \mathcal{A}_n is compact. By Problem 5.13, the operator $A: X \rightarrow X^*$ is compact, too.

Ad (iii). Observe that

$$\langle w, u \rangle = \int_a^b w(x)u(x)dx \quad \text{for all } w \in X^*, u \in X,$$

by Problem 5.10. According to the *Tonelli theorem* from the appendix to AMS Vol. 108, for all $u, v \in X$, we get

$$\begin{aligned} \langle Bv, u \rangle &= \int_a^b \left(\int_a^b \mathcal{A}(y, x)v(y)dy \right) u(x)dx \\ &= \int_a^b \left(\int_a^b \mathcal{A}(y, x)u(x)dx \right) v(y)dy = \langle v, Au \rangle. \end{aligned}$$

Ad (iv). Observe that $(u | v) = \langle u, v \rangle$ if $p = 2$, that is, $X = L_2(a, b)$.

5.15. Applications to integral equations. Show that Proposition 1 in Section 5.3 remains valid if the function $\mathcal{A}:]a, b[\times]a, b[\rightarrow \mathbb{R}$ is measurable with

$$\int_a^b \int_a^b |\mathcal{A}(x, y)|^2 dx dy < \infty.$$

5.16. The zoo of function spaces. Let G be a nonempty, open subset of \mathbb{R}^N . In order to solve linear or nonlinear partial differential equations by the methods of functional analysis, one has to choose the appropriate function spaces. There exist two important types of function spaces, namely,

- (i) the Lebesgue spaces $L_p^{\mathbb{C}}(G)$ and the Sobolev spaces $W_p^m(G)_{\mathbb{C}}$ of complex functions $f: G \rightarrow \mathbb{C}$, where $1 \leq p \leq \infty$ and $m = 1, 2, \dots$ (cf. Problems 5.10 and 5.12), and
- (ii) the Hölder spaces $C^{m,\alpha}(\overline{G})$ of complex functions $f: G \rightarrow \mathbb{C}$, where $m = 0, 1, 2, \dots$ and $0 < \alpha < 1$ (cf. Problem 1.8 in AMS Vol. 108).

5.16a. Show that the embedding $C^{m,\alpha}(\overline{G}) \subseteq C^{k,\beta}(\overline{G})$ is compact provided that G is bounded and $k < m$ or $k = m$ and $0 < \beta < \alpha$.

Hint: Use the Arzelà–Ascoli theorem from Section 1.11 of AMS Vol. 108.

There exist many other important classes of function spaces (e.g., fractional Sobolev spaces, Zygmund spaces, Hardy spaces, Morrey–Campanato spaces, spaces of bounded variation (and of generalized bounded variation), spaces of bounded mean oscillations, and so on). It turns out that there are two scales of spaces, namely the *Besov spaces* B_{pq}^s and the *Triebel–Lizorkin spaces* F_{pq}^s , which play a fundamental role in organizing the zoo of function spaces. In this connection, the *Fourier transformation* plays the decisive role. Let us briefly discuss this.

5.16b. *Dyadic partition of unity for \mathbb{R}^N .* By definition, such a partition is a family $\{\phi_j\}$ of C^∞ -functions $\phi_j: \mathbb{R}^N \rightarrow \mathbb{C}$ which have the following properties:

- (i) $\sum_{j=0}^{\infty} \phi_j(x) = 1$ for all $x \in \mathbb{R}^N$.
- (ii) $\phi_0(x) = 0$ if $|x| > 2$.
- (iii) $\phi_j(x) = 0$ if $|x| < \frac{1}{2^{j-1}}$ or $\frac{1}{2^{j+1}} < |x|$, where $j = 1, 2, \dots$

5.16c. *The definition of the basic scales of function spaces via Fourier transformation.* Let $\{\phi_j\}$ be a dyadic partition of unity for \mathbb{R}^N . Suppose that $1 \leq p, q \leq \infty$ and $s \in \mathbb{R}$. We define

$$B_{pq}^s(\mathbb{R}^N) := \{f \in \mathcal{S}' : \mathcal{N}_q(\|2^{sj} f_j\|_p) < \infty\},$$

and, for $p < \infty$,

$$F_{pq}^s(\mathbb{R}^N) := \{f \in \mathcal{S}' : \|\mathcal{N}_q(2^{sj} f_j)\|_p\} < \infty.$$

The Besov spaces B_{pq}^s are related to Hölder spaces, whereas the Triebel–Lizorkin spaces F_{pq}^s are related to Sobolev spaces and fractional Sobolev spaces (cf. Problem 5.16f).

Discussion of notation. Let us explain the notation used in the preceding definitions. Recall from Section 3.8 in AMS Vol. 108 that \mathcal{S}' denotes the space of tempered distributions and that the classic Fourier transformation can be extended to a linear bijective operator $F: \mathcal{S}' \rightarrow \mathcal{S}'$. By definition

$$f_j := F^{-1}(\phi_j F f), \quad j = 1, 2, \dots$$

Let the distribution $f \in \mathcal{S}'$ be given. Then the distribution $\phi_j F f$ represents a localization of the Fourier transform $F f$ of f . The inverse Fourier transformation applied to ϕ_j yields the function f_j . More precisely, we obtain

the decomposition³⁴

$$f = \sum_{j=0}^{\infty} f_j,$$

where f_j is an *entire analytic function* for all j , by the Payley–Wiener–Schwartz theorem (cf. Schwartz (1966)).

The norm $\|\cdot\|_p$ is the norm on the Lebesgue space $L_p^{\mathbb{C}}(\mathbb{R}^N)$ (cf. Problem 5.9), whereas $\mathcal{N}_q(a_j)$ is the norm on the space $l_q^{\mathbb{C}}$. That is,

$$\mathcal{N}_q(a_j) := \begin{cases} \left(\sum_{j=0}^{\infty} |a_j|^q \right)^{\frac{1}{q}} & \text{if } 1 \leq q < \infty, \\ \sup_j |a_j| & \text{if } q = \infty. \end{cases}$$

5.16d. *Banach spaces.* Show that³⁵

- (i) $B_{pq}^s(\mathbb{R}^N)$ is a complex Banach space with respect to the norm $\mathcal{N}_q(\|2^{sj} f_j\|_p)$.
- (ii) $F_{pq}^s(\mathbb{R}^N)$ is a complex Banach space with respect to the norm $\|\mathcal{N}_q(2^{sj} f_j)\|_p$.

5.16e. *Function spaces on bounded domains.* Let G be a nonempty, bounded, open subset of \mathbb{R}^N with smooth boundary.³⁶ Define

$$B_{pq}^s(G) := \text{restriction of the elements from } B_{pq}^s(\mathbb{R}^N) \text{ to } G,$$

and

$$F_{pq}^s(G) := \text{restriction of the elements from } F_{pq}^s(\mathbb{R}^N) \text{ to } G.$$

This means the following. Let $f \in B_{pq}^s(\mathbb{R}^N)$. Then f is a tempered distribution. The restriction f_* of f to the set G is defined through

$$f_*(\phi) := f(\phi) \quad \text{for all } \phi \in C_0^\infty(G). \quad (126)$$

Show that

³⁴This is to be understood in the sense of distributions, that is,

$$f(\phi) = \sum_{j=0}^{\infty} f_j(\phi) \quad \text{for all } \phi \in \mathcal{S},$$

where $f_j(\phi) = \int_{\mathbb{R}^N} f_j(x) \phi(x) dx$.

³⁵It turns out that the definition of the spaces B_{pq}^s and F_{pq}^s does not depend on the choice of the dyadic partition of unity. However, changing this partition leads to equivalent norms.

³⁶This means that the boundary ∂G of G is an $(N - 1)$ -dimensional C^∞ -manifold such that G lies locally on one side of ∂G .

- (i) $B_{pq}^s(G)$ is a complex Banach space with respect to the norm

$$\|f_*\| := \inf \|f\|.$$

Here $\|f\|$ denotes the norm on $B_{pq}^s(\mathbb{R}^N)$. and the infimum is taken over all elements f of $B_{pq}^s(\mathbb{R}^N)$ for which relation (126) holds true.

- (ii) $F_{pq}^s(G)$ is a complex Banach space with respect to the norm (126) replacing $B_{pq}^s(\mathbb{R}^N)$ with $F_{pq}^s(\mathbb{R}^N)$.

5.16.f* *Important special cases.* Let $G = \mathbb{R}^N$ or let G be given as in Problem 5.16e. Then

- (i) $B_{\infty,\infty}^{m,\alpha}(G) = C^{m,\alpha}(\overline{G})$ (Hölder spaces) if $m = 0, 1, \dots$ and $0 < \alpha < 1$.
- (ii) $F_{p2}^0(G) = L_p^{\mathbb{C}}(G)$ (Lebesgue spaces) if $1 < p < \infty$.
- (iii) $F_{p2}^m(G) = W_p^m(G)_{\mathbb{C}}$ (Sobolev spaces) if $m = 1, 2, \dots$ and $1 < p < \infty$.

Generally, if s is an arbitrary real number and $1 < p < \infty$, then $F_{p2}^s(G)$ is called a *fractional Sobolev space* denoted by $W_p^s(G)$.

Hint: Cf. Triebel (1992).

5.16g. Characterization of fractional Sobolev spaces. Let $s \in \mathbb{R}$ and $1 < p < \infty$. Then $W_p^s(\mathbb{R}^N)_{\mathbb{C}}$ consists of all tempered distributions f such that

$$F^{-1}(\psi_s F f) \in L_p^{\mathbb{C}}(\mathbb{R}^N),$$

where $\psi_s(x) := (1 + |x|^2)^{\frac{s}{2}}$. Show that $W_p^s(\mathbb{R}^N)_{\mathbb{C}}$ is a complex Banach space with respect to the norm

$$\|f\| := \|F^{-1}(\psi_s F f)\|_p.$$

Hint: Cf. Zeidler (1986), Vol. 2A, Section 21.20 and Triebel (1992).

Historical Remark. Hölder and Ljapunov introduced the class of Hölder continuous functions at the end of the nineteenth century in order to describe subtle properties of potentials caused by mass distributions. Sobolev spaces $W_p^m(G)_{\mathbb{C}}$ for $m = 1, 2, \dots$ emerged in the 1930s. These two classes of spaces play a fundamental role in the modern theory of partial differential equations (e.g., see Zeidler (1986), Vols. 1–5). Since the 1950s, many attempts were made to generalize these two classes of function spaces. The spaces B_{pq}^s and F_{pq}^s were introduced in the 1970s. A detailed study of these spaces and of their relations to other function spaces (along with valuable historical remarks) can be found in the monograph by Triebel (1992).

As an elementary introduction to function spaces, we recommend the textbooks by Kufner, John, and Fučík (1977) and Zeidler (1986), Vol. 2A, Chapters 21 and 22. A summary of important results about function spaces

and about their relation to *interpolation theory* can be found in the extensive appendix to Zeidler (1986), Vol. 2B. Interpolation theory emerged in the 1960s and represents an important tool in order to organize the zoo of function spaces and to obtain properties of operators between function spaces in an intelligent and very effective way.

5.17. The generalized Riesz theorem for functionals. Let X be a real Banach space, and let Y be a real Hilbert space. Suppose that the linear continuous operator $A: X \rightarrow Y$ has closed range. We are given a functional $F \in X^*$ that vanishes on the null space $N(A)$ of the operator A . Show that there exists an element p of Y such that

$$F(v) = (p \mid Av)_X \quad \text{for all } v \in X.$$

Solution: Consider the dual operator $A^T: Y^* \rightarrow X^*$. The closed range theorem from Section 3.12 tells us that

$$N(A)^\perp = R(A^T).$$

By assumption, $F \in N(A)^\perp$. Thus, there is a functional $f \in Y^*$ such that

$$F = A^T f.$$

By the Riesz theorem, there exists a point $p \in Y$ such that

$$f(w) = (p \mid w) \quad \text{for all } w \in Y.$$

Thus, for all $v \in X$,

$$F(v) = (A^T f)(v) = f(Av) = (p \mid Av). \quad \square$$

This result will be used in Problem 5.18b in order to prove the existence of a pressure function p in a fluid.

5.18. Modern vector calculus and its physical interpretation. Let G be a nonempty, bounded, open, connected set in \mathbb{R}^3 such that the boundary ∂G is sufficiently smooth.³⁷

5.18a.* The compressibility equation. Recall from Section 5.17.3 that

$$\mathcal{H} = \overset{\circ}{W}_2^1(G) \times \overset{\circ}{W}_2^1(G) \times \overset{\circ}{W}_2^1(G)$$

and

$$Z = L_2(G) \times L_2(G) \times L_2(G).$$

³⁷For example, suppose that the boundary ∂G is a two-dimensional C^1 -manifold, where G lies locally on one side of ∂G .

Furthermore,

$$D = \{(v_1, v_2, v_3) : v_j \in C_0^\infty(G) \text{ for all } j, \text{ and } \partial_j v_j = 0\},$$

and X denotes the closure of D in the Hilbert space \mathcal{H} . In what follows we sum over two equal indices from 1 to 3.

Consider now the problem

$$\begin{aligned} \nabla \mathbf{v} &= \mu && \text{on } G, \\ \mathbf{v} &= 0 && \text{on } \partial G. \end{aligned} \quad (127)$$

Recall that $\nabla \mathbf{v} = \operatorname{div} \mathbf{v}$. For given μ , we are looking for a velocity field \mathbf{v} of a fluid on G .

Here, μ measures the compressibility of the fluid. More precisely, the quantity $\operatorname{div} \mathbf{v}$ measures the relative change in volume of the flow (in first-order approximation). In particular, $\operatorname{div} \mathbf{v} \equiv 0$ is equivalent to the fact that the flow is volume preserving, that is, the flow is incompressible (cf. Zeidler (1986), Vol. 4, Section 70.5).

Using the velocity components with respect to a Cartesian coordinate system, we obtain the following equivalent problem:

$$\begin{aligned} \partial_j v_j &= \mu && \text{on } G, \\ v_j &= 0 && \text{on } \partial G, \quad j = 1, 2, 3. \end{aligned} \quad (128)$$

We are looking for a solution $v \in \mathcal{H}$. The following are met:

- (i) If $v \in \mathcal{H}$ is a solution of (128), then $\mu \in Z$ and

$$\int_G \mu dx = 0. \quad (129)$$

- (ii) For given $\mu \in Z$ with (128), the original problem (128) has a solution $v_* \in \mathcal{H}$. The general solution to (128) is given by

$$v = v_* + w, \quad w \in X. \quad (130)$$

Hint: Relation (129) follows from the Gauss theorem $\int_G \mu dx = \int_G \operatorname{div} \mathbf{v} dx = \int_{\partial G} \mathbf{v} \cdot \mathbf{n} dS = 0$.

Study the proof to (ii) in the monograph by Galdi (1994), Vol. 1, Sections 3.3 and 3.4.1.

Remark: Let $\mu = 0$. If G is an unbounded domain, then each velocity field $v \in X$ is a solution to equation (128). Unfortunately, if G is poorly shaped, then it may be that other solutions to (128) are not living in the space X . This fact, discovered by Heywood in 1976, complicates the investigations of the nonstationary Navier–Stokes equation (cf. Galdi (1994), Vols. 3 and 4).

5.18b. *The weak pressure equation.* Consider the equation

$$\begin{aligned} -\nabla p &= \mathbf{P} && \text{on } G, \\ \int_G p dx &= 0. \end{aligned} \quad (131)$$

Recall that $\nabla p = \text{grad } p$. Here, p can be regarded as the pressure in a fluid. For given outer force density³⁸ \mathbf{P} , we are looking for a pressure function p normalized by the second equation in (131). Using a Cartesian coordinate system, problem (131) is equivalent to the following problem:

$$\begin{aligned} -\partial_j p &= P_j && \text{on } G, \quad j = 1, 2, 3, \\ \int_G p dx &= 0. \end{aligned} \quad (132)$$

We are given the functionals $P_j \in W_2^{-1}(G)$, $j = 1, 2, 3$, such that

$$P_j(\phi_j) = 0 \quad \text{for all } \phi \in D. \quad (133)$$

Show that problem (132) has a unique solution $p \in L_2(G)$.

Use Problems 5.17 and 5.18a.

Solution: *Step 1: Existence.* Let $Y := \{p \in L_2(G) : \int_G p dx = 0\}$. The linear continuous operator

$$\nabla: \mathcal{H} \rightarrow Y$$

is surjective, and it has the null space $N(\nabla) = X$, by Problem 5.18a. Define

$$P(\phi) := P_j(\phi_j) \quad \text{for all } \phi \in \mathcal{H}.$$

Then the functional $P: \mathcal{H} \rightarrow \mathbb{R}$ is linear and continuous, and it vanishes on the null space $N(\nabla)$, by (133). Thus, Problem 5.17 tells us that there exists a $p \in Y$ such that

$$P(\phi) = (p \mid \nabla \phi)_Y = \int_G p \nabla \phi dx \quad \text{for all } \phi \in \mathcal{H}.$$

In particular, this implies

$$P_j(\psi) = \int_G p \partial_j \psi dx \quad \text{for all } \psi \in C_0^\infty(G).$$

Therefore, p satisfies equation (132), in the sense of distribution theory.

Step 2: Suppose that we are given a function $p \in L_2(G)$ along with

$$\partial_j p = 0 \quad \text{on } G, \quad j = 1, 2, 3.$$

Then, $p \in W_2^1(G)$. Let's use Friedrich's mollification from Problem 2.12 in AMS Vol. 108. Set

$$p_\varepsilon(x) := \int_G \phi_\varepsilon(x - y) p(y) dy \quad \text{on } G \text{ for all } \varepsilon > 0.$$

³⁸The pressure p generates the force $\int_H \mathbf{P} dx$ acting on each open subset H of G .

Let H be a connected open subset of G . Then, for all $\varepsilon < \text{distance}(\partial G, H)$,

$$\partial_j p_\varepsilon(x) = \int_G \partial_j^x \phi_\varepsilon(x - y) p(y) dy = - \int_G \partial_j^y \phi_\varepsilon(x - y) p(y) dy.$$

Integration by parts yields

$$\partial_j p_\varepsilon(x) = \int_G \phi_\varepsilon(x - y) \partial_j p(y) dy = 0 \quad \text{on } H \text{ for all } j.$$

Since the function p_ε is smooth, it is a constant on H for all $\varepsilon < \text{distance}(\partial G, H)$.

From $p_\varepsilon \rightarrow p$ in $L_2(G)$ as $\varepsilon \rightarrow 0$, it follows that there is a subsequence such that

$$p_{\varepsilon_n}(x) \rightarrow p(x) \quad \text{as } n \rightarrow \infty \quad \text{for almost all } x \in G,$$

by Problem 5.9a. Thus, $p(x) = \text{const}$ for almost all $x \in G$. The normalization condition $\int_G p dx = 0$ enforces $p(x) = 0$ for almost all $x \in G$.

5.18c.* *The strong pressure equation.* We want to solve the pressure equation (132) in the case where P_1 , P_2 , and P_3 are functions. Let $P_j \in L_2(G)$, $j = 1, 2, 3$. Then problem (132) has a solution $p \in W_2^1(G)$ iff

$$\int_G P_j v_j dx = 0 \quad \text{for all } v \in D. \quad (134)$$

This tells us that there is a duality between the velocity fields $v \in X$ of a viscous incompressible fluid and the outer force densities $P \in Z$ generated by a pressure p .

Hint: Study the elegant proof given in the monograph by Galdi (1994), Vol. 1, p. 103.

5.18d. *The famous Helmholtz–Weyl decomposition of vector fields.* Let us reformulate the theorem from the preceding problem in terms of Hilbert space theory. Consider the Hilbert space

$$Z = L_2(G) \times L_2(G) \times L_2(G),$$

along with the inner product $(P \mid v) = \int_G P_j v_j dx$. Define

$$Z_1 := \text{closure of } D \text{ in } Z,$$

$$Z_2 := \{P \in Z : P = -\text{grad } p \text{ for some } p \in W_2^1(G)\}.$$

Recall that D consists of all $C_0^\infty(G)$ -vector fields v with $\text{div } v = 0$. Obviously, Z_1 and Z_2 are closed linear subspaces of the Hilbert space Z .

Problem 5.18c is now equivalent to saying that

$$Z_1^\perp = Z_2 \quad \text{in } Z,$$

that is, Z_2 represents the orthogonal complement to Z_1 in the Hilbert space Z . Therefore, we have the famous orthogonal decomposition

$$Z = Z_1 \otimes Z_2. \quad (135)$$

This means that, for each vector field $w \in Z$, there exists the unique decomposition

$$w = v + P, \quad v \in Z_1, \quad P \in Z_2.$$

In particular, we have

$$\operatorname{div} v = 0 \quad \text{on } G$$

and

$$\operatorname{curl} P = 0 \quad \text{on } G, \quad (136)$$

in the sense of distribution theory. Observe that $\operatorname{curl} P = -\operatorname{curl} \operatorname{grad} p = -\nabla \times (\nabla p) = 0$ for smooth functions p . Then a passage to the limit yields (136).

The decomposition of vector fields w into the sum of a divergence-free field v and a curl-free field P was used by Hermann von Helmholtz in 1870. The orthogonal decomposition (135) dates back to a famous paper written by Hermann Weyl in 1940.

5.18e.** *The very weak pressure equation.* Let G be a nonempty open subset of \mathbb{R}^n , $n \geq 1$. Consider the pressure equation

$$-\partial_j p = P_j \quad \text{on } G, \quad j = 1, \dots, n. \quad (137)$$

We are given the distributions $P_j \in \mathcal{D}'(G)$ for all j . Then, a distribution $p \in \mathcal{D}'(G)$ is a solution of (137) iff

$$P_j(\phi_j) = 0 \quad \text{for all } \phi_j \in C_0^\infty(G) \text{ with } \partial_j \phi_j = 0. \quad (138)$$

Here we sum over j from 1 to n .

Hint: Obviously, relation (137) implies (138) (cf. the proof of Proposition 13 in Section 5.17). The converse statement represents the special case of a profound theorem on differential forms due to de Rham (cf. Temam (1977), p. 14).

The philosophy of the de Rham theorem is that, for equations in terms of differential forms, the natural necessary solvability conditions are also sufficient. This was also the general philosophy of the present chapter.

5.19. *Bifurcation and formation of patterns in nature.* Bifurcation describes the change of the qualitative behavior of systems in nature produced by a loss of stability under external influences. From a mathematical point of view, the following problems can be solved by using Theorem 5.J.

5.19a.* *The Bénard problem.* Consider a viscous fluid between two plates, as shown in Figure 5.5, where the temperature T_0 of the lower plate and the temperature T_1 of the upper plate satisfy the condition

$$T_0 > T_1.$$

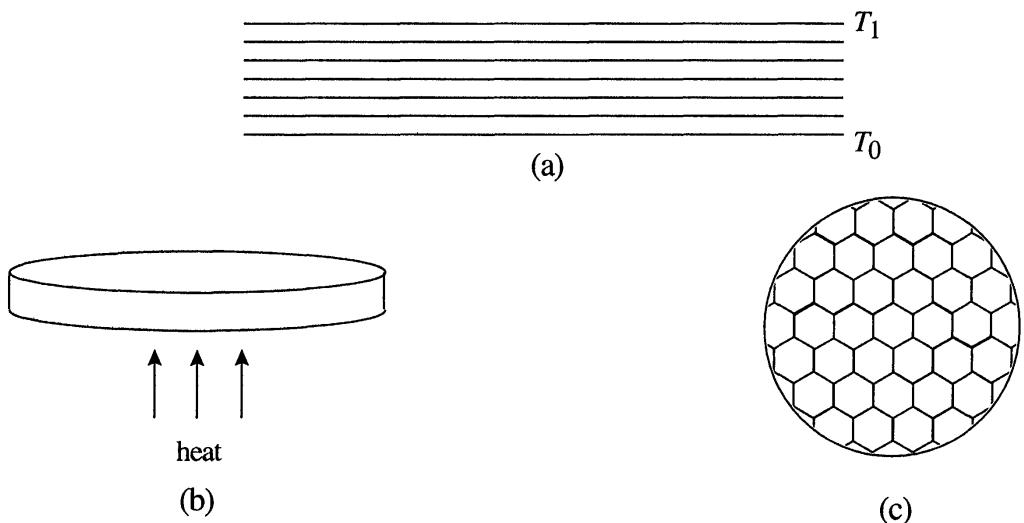


FIGURE 5.5.

If the temperature difference $T_0 - T_1$ is sufficiently small, then the fluid is at rest. If the temperature difference is increased, then at a critical value, Bénard cells appear in the fluid. These cells have a hexagonal structure. This phenomenon was discovered experimentally by Bénard in 1901.

In experiments a pan with silicon oil is heated with hot water from underneath it. The fluid flow is made visible through small, equally distributed aluminium pieces. After reaching the critical temperature difference, hexagonal cells appear in the pan, which are shown from above in Figure 5.5(c).

Bénard cells correspond to a bifurcation phenomenon. Physically, they arise by combining the gravitational force and the heat convection flow. During the past twenty years, physicists, chemists, and biologists have shown a great deal of interest in these Bénard cells, because one observes the formation of a complicated structure. This process frequently occurs in the evolution of life.

From a mathematical point of view, one can apply Theorem 5.J to the Navier–Stokes equations combined with the equations for heat conduction.

Study the proof in Zeidler (1986), Vol. 4, Section 72.9.

5.19b.* The Taylor problem. As in Figure 5.6(a) we consider a viscous fluid between two concentric cylinders, whereby the outer cylinder is at rest and the inner cylinder rotates counterclockwise around the z -axis with angular velocity ω . Let the cylinder radii be r and R , respectively, with $r < R$. The Reynolds number Re is important. We set

$$\text{Re} = \frac{\rho \omega r^2}{\eta}.$$

Here, ρ and η denote the density and the viscosity of the fluid, respectively. In experiments one observes a critical Reynolds number Re_0 with the following properties:

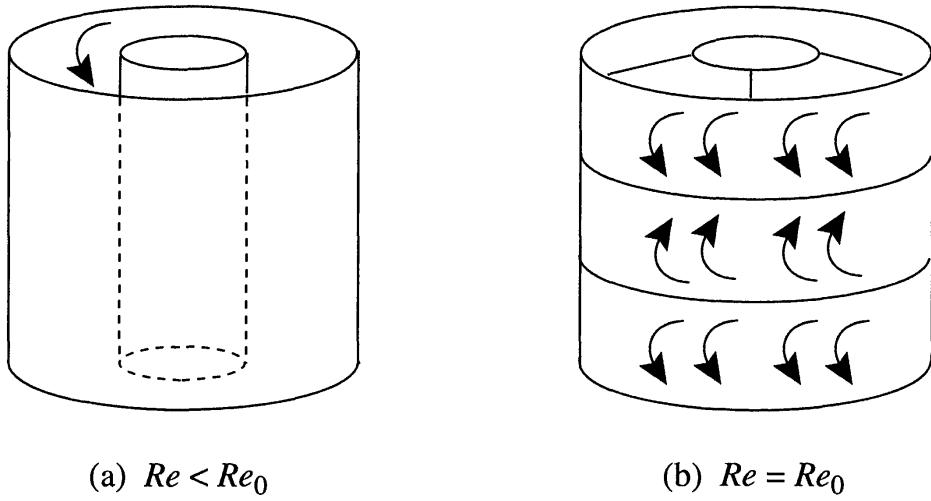


FIGURE 5.6.

- (i) For $Re < Re_0$, that is, for small angular velocities ω there exists an axisymmetric flow that does not depend on the z -coordinate. This is the Couette flow.
- (ii) For $Re = Re_0$, Taylor vortices occur, which are periodic in z (see Figure 5.6(b)). These Taylor vortices were discovered experimentally by Taylor in 1922.
- (iii) If the angular velocity ω gets larger and larger, that is, for increasing Reynolds numbers, one obtains more and more complicated flow pictures until at a certain Re_{crit} turbulence occurs.

From a mathematical point of view, one can apply Theorem 5.J to the Navier–Stokes equations. Study the proof to (ii) in Zeidler (1986), Vol. 4, Section 72.7. A detailed discussion of the Couette–Taylor flow can be found in the monograph by Chossat and Iooss (1994).

5.19c.* Hopf bifurcation. Consider a finite-dimensional or infinite-dimensional dynamical system that is in an equilibrium state. If an external influence acts on the system, then it may happen that the equilibrium state loses its stability and the system starts oscillations. This important phenomenon, called Hopf bifurcation, was discovered by Eberhard Hopf in 1942.

From a mathematical point of view, this bifurcation problem can be solved by using Theorem 5.J.

Study the proof in Zeidler (1986), Vol. 4, Section 79.9.

5.19d.* Water waves and bifurcation. Consider a parallel water flow in a channel. If the velocity c of the flow becomes critical, then permanent water waves occur (cf. Figure 5.7).

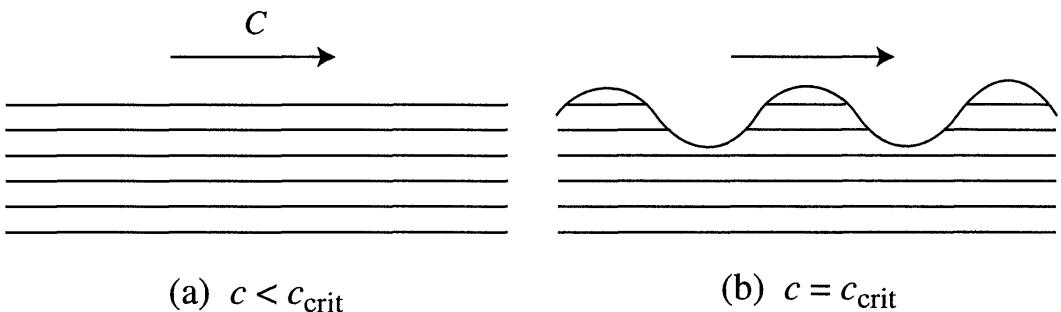


FIGURE 5.7.

From a mathematical point of view, one can use Theorem 5.J. Study the proof for permanent gravitational water waves in Zeidler (1986), Vol. 4, Chapter 71.

The rigorous treatment of permanent water waves represented a famous open problem in the nineteenth century. A detailed mathematical and physical discussion of a broad class of permanent waves (including capillary-gravity waves and tidal waves) along with historical remarks can be found in the monograph by Zeidler (1972). See also the survey article by Zeidler (1977).

5.20. The buckling of beams and plates, and bifurcation for problems that have a variational structure. Under the influence of critical external forces, buckling of beams and plates occurs as pictured in Figure 5.3. Such problems can frequently be solved by using Theorem 5.J. However, there is a general bifurcation theorem for problems that have a variational structure. Roughly speaking, this theorem says that each eigenvalue of the linearized problem is also a bifurcation parameter.

- (i) Study the main theorem of bifurcation theory for Fredholm operators of variational type in Zeidler (1986), Vol. 2B, Section 29.18.
- (ii) Study applications of this theorem to general variational problems in Zeidler (1986), Vol. 2B, Section 29.19ff.
- (iii) Study the buckling of beams in Zeidler (1986), Vol. 2B, Section 29.13.
- (iv) Study the buckling of plates in Zeidler (1986), Vol. 4, Chapter 65 (the von Kármán equations).

References

Additional references along with hints for further reading can be found in AMS Vol. 108.

- Abraham, R., Marsden, J., and Ratiu, T. (1983): *Manifolds, Tensor Analysis, and Applications*. Addison-Wesley, Reading, MA.
- Abraham, R. and Robbin, J. (1967): *Transversal Mappings and Flows*. Benjamin, New York.
- Albers, D., Alexanderson, G., and Reid, C. (1987): *International Mathematical Congresses*. Springer-Verlag, New York.
- Alt, H. (1992): *Lineare Funktionalanalysis: eine anwendungsorientierte Einführung*. 2nd edition. Springer-Verlag, Berlin, Heidelberg.
- Amann, H. (1990): *Ordinary Differential Equations: An Introduction to Nonlinear Analysis*. De Gruyter, Berlin.
- Amann, H. (1995): *Linear and Quasilinear Parabolic Problems*, Vol. 1. Birkhäuser, Basel.
- Ambrosetti, A. (1993): *A Primer of Nonlinear Analysis*. Cambridge University Press, Cambridge, UK.
- Ambrosetti, A. and Coti-Zelati, V. (1993): *Periodic Solutions of Singular Lagrangian Systems*. Birkhäuser, Basel.
- Antman, S. (1995): *Nonlinear Elasticity*. Springer-Verlag, New York.
- Appell, J. and Zabrejko, P. (1990): *Nonlinear Superposition Operators*. Cambridge University Press, Cambridge, UK.
- Aubin, J. (1977): *Applied Functional Analysis*. Wiley, New York.

- Aubin, J. (1993): *Optima and Equilibria: An Introduction to Nonlinear Analysis*. Springer-Verlag, Berlin, Heidelberg (translation from French).
- Aubin, J. and Ekeland, I. (1983): *Applied Nonlinear Functional Analysis*. Wiley, New York.
- Bagger, J. and Wess, J. (1991): *Supersymmetry and Supergravity*. 2nd expanded edition. Princeton University Press, Princeton, NJ.
- Baggett, L. (1992): *Functional Analysis: A Primer*. Marcel Dekker, New York.
- Bakelman, I. (1994): *Convex Analysis and Nonlinear Geometric Elliptic Equations*. Springer-Verlag, Berlin, Heidelberg.
- Banach, S. (1932): *Théorie des opérations linéaires*. Warszawa. (English edition: *Theory of Linear Operations*. North-Holland, Amsterdam, 1987.)
- Banks, R. (1994): *Growth and Diffusion Phenomena*. Springer-Verlag, Berlin, Heidelberg.
- Barton, G. (1989): *Elements of Green's Functions and Propagation: Potentials, Diffusion, and Waves*. Clarendon Press, Oxford.
- Bartsch, T. (1993): *Topological Methods for Variational Problems with Symmetry*. Springer-Verlag, Berlin, Heidelberg.
- Beauzamy, B. (1988): *Introduction to Operator Theory and Invariant Subspaces*. North-Holland, Amsterdam.
- Berberian, S. (1974): *Lectures in Functional Analysis and Operator Theory*. Springer-Verlag, New York.
- Berezin, F. (1987): *Introduction to Superanalysis*. Reidel, Dordrecht.
- Berezin, F. and Shubin, M. (1991): *The Schrödinger Equation*. Kluwer, Dordrecht.
- Berger, M. (1977): *Nonlinearity and Functional Analysis*. Academic Press, New York.
- Bethuel, F., Brézis, H., and Hélein, F. (1994): *Ginzburg-Landau Vortices*. Birkhäuser, Basel.
- Boccara, N. (1990): *Functional Analysis*. Academic Press, New York.
- Bogoljubov, N. (1967): *Lectures on Quantum Statistics*, Vols. 1, 2. Gordon and Breach, New York (translation from Russian).
- Booss, B. and Bleecker, D. (1985): *Topology and Analysis*. Springer-Verlag, New York.
- Bourgignon, J. (1995): *Variational Calculus*. Springer-Verlag, Berlin, Heidelberg.
- Bratteli, C. and Robinson, D. (1979): *Operator Algebras and Quantum Statistical Mechanics*, Vols. 1, 2. Springer-Verlag, New York.
- Brézis, H. (1983): *Analyse fonctionnelle et applications*. Masson, Paris.

- Brody, F. and Vamos, T. (eds.) (1994): *Neumann Compendium* (selected papers by John von Neumann). World Scientific, Singapore.
- Brokate, M. and Sprekels, J. (1995): *Hysteresis Phenomena in Phase Transitions*. Springer-Verlag, Berlin, Heidelberg (to appear).
- Browder, F. (ed.) (1992): *Nonlinear and Global Analysis*. Reprints from the Bulletin of the American Mathematical Society. Providence, RI.
- Caroll, R. (1988): *Mathematical Physics*. North-Holland, Amsterdam.
- Chang, K. (1966): *Critical Point Theory and Its Applications*. Springer-Verlag, Berlin, Heidelberg (to appear).
- Choquet-Bruhat, Y., DeWitt-Morette, and Dillard-Bleick, M. (1988): *Analysis, Manifolds, and Physics*. Vols. 1, 2. North-Holland, Amsterdam.
- Chorin, J. (1975): *Lectures on Turbulence Theory*. Publish or Perish, Boston, MA.
- Chorin, A. (1994): *Vorticity and Turbulence*. Springer-Verlag, New York.
- Chossat, P. and Iooss, G. (1994): *The Couette-Taylor Flow*. Springer-Verlag, New York.
- Chow, S. and Hale, J. (1982): *Methods of Bifurcation Theory*. Springer-Verlag, Berlin, Heidelberg.
- Ciarlet, P. (1977): *Numerical Analysis of the Finite Element Method for Elliptic Boundary-Value Problems*. North-Holland, Amsterdam.
- Ciarlet, P. (1983): *Lectures on Three-Dimensional Elasticity*. Springer-Verlag, New York.
- Collet, P. and Eckmann, J. (1990): *Instabilities and Fronts in Extended Systems*. Princeton University Press, Princeton, NJ.
- Collins, J. (1984): *Renormalization*. Cambridge University Press, Cambridge, UK.
- Colombeau, J. (1992): *Multiplication of Distributions*. Lecture Notes in Mathematics, Vol. 1532. Springer-Verlag, Berlin, Heidelberg.
- Connes, A. (1994): *Noncommutative Geometry*. Academic Press, New York.
- Constantinescu, F. and de Groote, H. (1994): *Geometrische und algebraische Methoden der Physik: Supermannigfaltigkeiten und Virasoro-Algebren*. Teubner-Verlag, Stuttgart.
- Conway, J. (1990): *A Course in Functional Analysis*. Springer-Verlag, New York.
- Cornwell, J. (1989): *Group Theory in Physics*. Vol. 1: *Fundamental Concepts*; Vol. 2: *Lie Groups and Their Applications*; Vol. 3: *Supersymmetries and Infinite-Dimensional Algebras*. Academic Press, New York.
- Courant, R. and Hilbert, D. (1937): *Die Methoden der Mathematischen Physik*, Vols. 1, 2. (English edition: *Methods of Mathematical Physics*, Vols. 1, 2. Wiley, New York, 1989).

- Creutz, M. (1983): *Quarks, Gluons, and Lattices*. Cambridge University Press, Cambridge, UK.
- Cycon, R., Froese, R., Kirsch, W., and Simon, B. (1986): *Schrödinger Operators*. Springer-Verlag, New York.
- Dacarogna, B. (1989): *Direct Methods in the Calculus of Variations*. Springer-Verlag, Berlin, Heidelberg.
- Dal Maso, G. (1993): *An Introduction to Γ -Convergence*. Birkhäuser, Basel.
- Dautray, D. and Lions, J. (1990): *Mathematical Analysis and Numerical Methods for Science and Technology*; Vol. 1: *Physical Origins and Classical Methods*; Vol. 2: *Functional and Variational Methods*; Vol. 3: *Spectral Theory and Applications*; Vol. 4: *Integral Equations and Numerical Methods*; Vol. 5: *Evolution Problems I*; Vol. 6: *Evolution Problems II – the Navier–Stokes Equations, the Transport Equations, and Numerical Methods*. Springer-Verlag, Berlin, Heidelberg (translation from French).
- Davies, P. (ed.) (1989): *The New Physics*. Cambridge University Press, Cambridge, UK.
- Deimling, K. (1985): *Nonlinear Functional Analysis*. Springer-Verlag, New York.
- Deimling, K. (1992): *Multivalued Differential Equations*. De Gruyter, Berlin.
- Deuflhard, P. and Hohmann, A. (1993): *Numerische Mathematik I*. De Gruyter, Berlin. (English edition: *Numerical Analysis: A First Course in Scientific Computation*. De Gruyter, Berlin, 1994.)
- Deuflhard, P. and Bornemann, F. (1994): *Numerische Mathematik II: Integration gewöhnlicher Differentialgleichungen*. De Gruyter, Berlin (English edition in preparation).
- DeVito, C. (1990): *Functional Analysis and Linear Operator Theory*. Addison-Wesley, Reading, MA.
- Diekman, O., Lunel, S., van Gils, A., and Walther, H. (1995): *Delay Equations: Functional Analysis, Complex Analysis, and Nonlinear Analysis*. Springer-Verlag, Berlin, Heidelberg (to appear).
- Dierkes, U., Hildebrandt, S., Küster, A., and Wohlrab, O. (1992): *Minimal Surfaces*, Vols. 1, 2. Springer-Verlag, Berlin, Heidelberg.
- Dieudonné, J. (1969): *Foundations of Modern Analysis*. Academic Press, New York.
- Dieudonné, J. (1981): *History of Functional Analysis*. North-Holland, Amsterdam.
- Donoghue, J., Golowich, M., and Holstein, B. (1992): *The Dynamics of the Standard Model*. Cambridge University Press, Cambridge, UK.
- Dubrovin, B., Fomenko, A., and Novikov, S. (1992): *Modern Geometry: Methods and Applications*, Vols. 1–3. Springer-Verlag, New York (translation from Russian).

- Dunford, N. and Schwartz, J. (1988): *Linear Operators*, Vols. 1–3. Wiley, New York.
- Economou, E. (1988): *Green's Functions in Quantum Physics*. Springer-Verlag, New York.
- Edwards, R. (1994): *Functional Analysis*. Dover, New York.
- Ekeland, I. and Temam, R. (1974): *Analyse convexe et problèmes variationnels*. Dunod, Paris. (English edition: *Convex Analysis and Variational Problems*. North-Holland, New York, 1976).
- Ekeland, I. (1979): *Eléments d'économie mathématique*. Hermann, Paris.
- Ekeland, I. (1990): *Convexity Methods in Hamiltonian Mechanics*. Springer-Verlag, New York.
- Emch, G. (1986): *Mathematical and Conceptual Foundations of 20th-Century Physics*. North-Holland, Amsterdam.
- Erickson, J. and Kinderlehrer, D. (eds.) (1988): *Theory and Applications of Liquid Crystals*. Springer-Verlag, New York.
- Euler, L. (1911ff): *Opera Omnia (Collected Papers)*. Leipzig-Berlin, later Basel-Zürich, Vols. 1–72.
- Evans, L. (1994): *Partial Differential Equations*. Berkeley Mathematics Lecture Notes, Vols. 3A and 3B. University of Berkeley, CA.
- Farkas, M. (1994): *Periodic Motions*. Springer-Verlag, Berlin, Heidelberg.
- Fenyö, S. and Stolle, H. (1982): *Theorie und Praxis der linearen Integralgleichungen*, Vols. 1–4. Deutscher Verlag der Wissenschaften, Berlin.
- Feynman, R., Leighton, R., and Sands, M. (1963): *The Feynman Lectures in Physics*. Addison-Wesley, Reading, MA.
- Finn, R. (1985): *Equilibrium Capillary Surfaces*. Springer-Verlag, Berlin, Heidelberg.
- Foias, C., Sell, G., and Temam, R. (1993): *Turbulence in Fluid Flows: A Dynamical Systems Approach*. Springer-Verlag, New York.
- Friedman, A. (1982): *Variational Principles and Free Boundary-Value Problems*. Wiley, New York.
- Friedman, A. (1989/94): *Mathematics in Industrial Problems*, Vols. 1–6. Springer-Verlag, New York.
- Gajewski, H., Gröger, K., and Zacharias, K. (1974): *Nichtlineare Operatorgleichungen*. Akademie-Verlag, Berlin.
- Galdi, G. (1994): *An Introduction to the Mathematical Theory of the Navier-Stokes Equations*, Vols. 1–4. Springer-Verlag, Berlin, Heidelberg (Vols. 3 and 4 to appear).
- Gelfand, I. and Shilov, E. (1964): *Generalized Functions*, Vols. 1–5. Academic Press, New York (translation from Russian).
- Gell-Mann, M. (1994): *The Quark and the Jaguar: Adventures in the Simple and the Complex*. Freeman, San Francisco, CA.

- Giaquinta, M. (1993): *Introduction to Regularity Theory for Nonlinear Elliptic Systems*. Birkhäuser, Basel.
- Giaquinta, M. and Hildebrandt, S. (1995): *Calculus of Variations*, Vols. 1, 2. Springer-Verlag, New York.
- Gilbarg, D. and Trudinger, N. (1994): *Elliptic Partial Differential Equations of Second Order*. 2nd edition. Springer-Verlag, New York.
- Gilkey, P. (1984): *Invariance Theory, the Heat Equation, and the Atiyah-Singer Index Theorem*. Publish or Perish, Boston, MA.
- Girvin, S. and Prange, R. (1990): *The Quantum Hall Effect*. 2nd edition. Springer-Verlag, New York.
- Green, M., Schwarz, J., and Witten, E. (1987): *Superstrings*, Vols. 1, 2. University Press, Cambridge, UK.
- Greiner, W. and Müller, B. (1994): *Quantum Mechanics: Symmetries*. Springer-Verlag, Berlin, Heidelberg.
- Greiner, W. and Reinhardt, J. (1994): *Quantum Electrodynamics*. Springer-Verlag, Berlin, Heidelberg.
- Greiner, W. (1993): *Gauge Theory of Weak Interactions*. Springer-Verlag, Berlin, Heidelberg.
- Greiner, W. and Schäfer, A. (1994): *Quantum Chromodynamics*. Springer-Verlag, Berlin, Heidelberg.
- Grosche, G., Ziegler, D., Ziegler, V., and Zeidler, E. (eds.) (1995): *Teubner-Taschenbuch der Mathematik II (Handbook of Advanced Mathematics)*. Teubner-Verlag, Stuttgart, Leipzig (English edition in preparation).
- Grosse, H. (1995): *Models in Statistical Physics and Quantum Field Theory*. Springer-Verlag, Berlin, Heidelberg (to appear).
- Gruber, P. and Wills, J. (1993): *Handbook of Convex Geometry*, Vols. 1, 2. North-Holland, Amsterdam.
- Guillemin, V. and Pollack, A. (1974): *Differential Topology*. Prentice-Hall, Englewood Cliffs, NJ.
- Guillemin, V. and Sternberg, S. (1990): *Symplectic Techniques in Physics*. Cambridge University Press, Cambridge, UK.
- Giusti, E. (1984): *Minimal Surfaces and Functions of Bounded Variation*. Birkhäuser, Basel.
- Gurtin, M. (1993): *Thermomechanics of Evolving Phase*. Clarendon Press, Oxford.
- Haag, R. (1993): *Local Quantum Physics: Fields, Particles, Algebras*. Springer-Verlag, Berlin, Heidelberg.
- Hackbusch, W. (1992): *Elliptic Differential Equations: Theory and Numerical Treatment*. Springer-Verlag, New York (translation from German).
- Hackbusch, W. (1994): *Iterative Solution of Large Sparse Systems of Equations*. Springer-Verlag, New York (translation from German).

- Hale, J. and Koçak, H. (1991): *Dynamics of Bifurcations*. Springer-Verlag, Berlin, Heidelberg (cf. also Koçak (1989)).
- Hatfield, B. (1992): *Quantum Field Theory of Point Particles and Strings*. Addison-Wesley, Redwood City, CA.
- Henneaux, M. and Teitelboim, C. (1993): *Quantization of Gauge Systems*. Princeton University Press, Princeton, NJ.
- Henry, D. (1981): *Geometric Theory of Semilinear Parabolic Equations*. Lecture Notes in Mathematics, Vol. 840. Springer-Verlag, New York.
- Hermann, C. and Sapoval, B. (1994): *Physics of Semiconductors*. Springer-Verlag, New York.
- Heuser, H. (1975): *Funktionalanalysis*. Teubner-Verlag, Stuttgart. (English edition: *Functional Analysis*, Wiley, New York, 1986).
- Hilbert, D. (1912): *Grundzüge einer allgemeinen Theorie der Integralgleichungen*. Teubner-Verlag, Leipzig.
- Hilbert, D. (1932): *Gesammelte Werke (Collected Works)*, Vols. 1–3. Springer-Verlag, Berlin.
- Hildebrandt, S. and Tromba, T. (1985): *Mathematics and Optimal Form*. Scientific American Library, Freeman, New York.
- Hiriart-Urruty, J. and Lemarchal, C. (1993): *Convex Analysis and Minimization Algorithms*, Vols. 1, 2. Springer-Verlag, Berlin, Heidelberg.
- Hirzebruch, F. and Scharlau, W. (1971): *Einführung in die Funktionalanalysis*. Bibliographisches Institut, Mannheim.
- Hofer, H. and Zehnder, E. (1994): *Symplectic Invariants and Hamiltonian Dynamics*. Birkhäuser, Basel.
- Holmes, R. (1975): *Geometrical Functional Analysis and Its Applications*. Springer-Verlag, New York.
- Honerkamp, J. and Römer, H. (1993): *Theoretical Physics: A Classical Approach*. Springer-Verlag, New York.
- Hoppenstaedt, F. and Peskin, C. (1994): *Mathematics in Medicine and the Life Sciences*. Springer-Verlag, New York.
- Hörmander, L. (1983): *The Analysis of Linear Partial Differential Operators*; Vol. 1: *Distribution Theory and Fourier Analysis*; Vol. 2: *Differential Operators with Constant Coefficients*; Vol. 3: *Pseudo-Differential Operators*; Vol. 4: *Fourier Integral Operators*. Springer-Verlag, New York.
- Hörmander, L. (1994): *Notions of Convexity*. Birkhäuser, Basel.
- Huang, K. (1992): *Quarks, Leptons, and Gauge Fields*. 2nd edition. World Scientific, Singapore.
- Isham, C. (1989): *Modern Differential Geometry for Physicists*. World Scientific, Singapore.

- Jost, J. (1984): *Harmonic Maps between Surfaces*. Springer-Verlag, Berlin, Heidelberg.
- Jost, J. (1991): *Two-Dimensional Geometric Variational Problems*. Wiley, New York.
- Jost, J. (1994): *Differentialgeometrie und Minimalflächen*. Springer-Verlag, Berlin, Heidelberg.
- Jost, J. (1994a): *Riemannian Geometry and Geometric Analysis*. Springer-Verlag, Berlin, Heidelberg.
- Jost, J. (1996): *Postmodern Analysis*. Springer-Verlag, Berlin, Heidelberg (to appear).
- Kadison, R. and Ringrose, J. (1983): *Fundamentals of the Theory of Operator Algebras*, Vols. 1–4. Academic Press, New York.
- Kaku, M. (1987): *Introduction to Superstring Theory*. Springer-Verlag, New York.
- Kaku, M. and Trainer, J. (1987): *Beyond Einstein: The Cosmic Quest for the Theory of the Universe*. Bantam Books, New York.
- Kaku, M. (1991): *Strings, Conformal Fields, and Topology*. Springer-Verlag, New York.
- Kaku, M. (1993): *Quantum Field Theory*. Oxford University Press, Oxford.
- Kantorovich, L. and Akilov, G. (1964): *Functional Analysis in Normed Spaces*. Pergamon Press, Oxford (translation from Russian).
- Kato, T. (1976): *Perturbation Theory for Linear Operators*. Springer-Verlag, Berlin.
- Kelley, J. (1955): *General Topology*. Van Nostrand, New York.
- Kevasan, S. (1989): *Topics in Functional Analysis and Applications*. Wiley, New York.
- Kirillov, A. and Gvishiani, A. (1982): *Theory and Problems in Functional Analysis*. Springer-Verlag, New York.
- Koçak, H. (1989): *Differential and Difference Equations Through Computer Experiments*. With Diskettes. Springer-Verlag, New York (cf. also Hale and Koçak (1991)).
- Kleinert, V. (1989): *Gauge Fields in Condensed Matter*, Vols. 1, 2. World Scientific, Singapore.
- Kolb, E. and Turner, M. (1990): *The Early Universe*. Addison-Wesley, Redwood City, CA.
- Kolmogorov, A., Fomin, S., and Silverman, R. (1975): *Introductory Real Analysis*. Dover, New York (enlarged translation from the Russian).
- Köthe, G. (1960): *Topologische lineare Räume*. Springer-Verlag, Berlin.
- Krasnoselskii, M. and Zabreiko, P. (1984): *Geometrical Methods in Nonlinear Analysis*. Springer-Verlag, New York (translation from Russian).

- Kress, R. (1989): *Linear Integral Equations*. Springer-Verlag, New York.
- Kreyszig, E. (1989): *Introductory Functional Analysis with Applications*. Wiley, New York.
- Kufner, A., John, O., and Fučík, S. (1977): *Function Spaces*. Academia, Prague.
- Kuksin, S. (1993): *Nearly Integrable Infinite-Dimensional Systems*. Lecture Notes in Mathematics, Vol. 1556. Springer-Verlag, Berlin.
- Kuperschmidt, B. (1992): *The Variational Principles of Dynamics*. World Scientific, Singapore.
- Ladyzhenskaya, O. (1969): *The Mathematical Theory of Viscous Incompressible Flows*. Gordon and Breach, New York.
- Landau, L. and Lifšic, E. (1982): *Course of Theoretical Physics*, Vols. 1–10. Elsevier, New York.
- Lang, S. (1993): *Real Analysis*. 3rd edition. Springer-Verlag, New York.
- Lawson, H. and Michelsohn, M. (1989): *Spin Geometry*. Princeton University Press, Princeton, NJ.
- Lazutkin, V. (1993): *KAM-Theory and Semiclassical Approximations to Eigenfunctions*. Springer-Verlag, Berlin, Heidelberg.
- Leis, R. (1986): *Initial-Boundary Value Problems in Mathematical Physics*. Wiley, New York.
- Levitan, B. and Sargsjan, I. (1991): *Sturm-Liouville and Dirac Operators*. Kluwer, Boston, MA (translation from Russian).
- Lions, J. (1969): *Quelques méthodes de résolution des problèmes aux limites nonlinéaires*. Dunod, Paris.
- Lions, J. (1971): *Optimal Control of Systems Governed by Partial Differential Equations*. Springer-Verlag, Berlin (translation from French).
- Lions, J. and Magenes, E. (1972): *Inhomogeneous Boundary-Value Problems*, Vols. 1–3. Springer-Verlag, New York.
- Luenberger, D. (1969): *Optimization by Vector Space Methods*. Wiley, New York.
- Lüst, D. and Theissen, S. (1989): *Lectures on String Theory*. Lecture Notes in Physics, Vol. 346. Springer-Verlag, Berlin, Heidelberg.
- Mandl, F. and Shaw, G. (1989): *Quantum Field Theory*. Wiley, New York.
- Marathe, K. and Martucci, M. (1992): *The Mathematical Foundations of Gauge Theory*. North-Holland, Amsterdam.
- Marchioro, C. and Pulvirenti, M. (1994): *Mathematical Theory of Inviscid Fluids*. Springer-Verlag, New York.
- Markowich, P. (1990): *Semiconductor Equations*. Springer-Verlag, Berlin, Heidelberg.

- Marsden, J. (1974): *Applications of Global Analysis in Mathematical Physics*. Publish or Perish, Boston, MA.
- Marsden, J. (1992): *Lectures in Mechanics*. Cambridge University Press, Cambridge, UK.
- Marsden, J. and Tromba, A. (1976): *Vector Calculus*. Freeman, San Francisco, CA.
- Marsden, J. and Hughes, T. (1983): *Mathematical Foundations of Elasticity*. Prentice-Hall, Englewood Cliffs, NJ.
- Marsden, J. and Ratiu, T. (1994): *Introduction to Mechanics and Symmetry: A Basic Exposition of Classical Mechanical Systems*. Springer-Verlag, New York.
- Mawhin, J. and Willem, M. (1987): *Critical Point Theory and Hamiltonian Systems*. Springer-Verlag, New York.
- Maurin, K. (1972): *Methods of Hilbert Spaces*. Polish Scientific Publishers, Warsaw.
- Meirmanov, A. (1992): *The Stefan Problem*. De Gruyter, Berlin (translation from Russian).
- Meyer, K. and Hall, G. (1992): *Introduction to Hamiltonian Dynamical Systems and the N-Body Problem*. Springer-Verlag, New York.
- Mielke, A. (1991): *Hamiltonian and Lagrangian Flows on Center Manifolds with Applications to Elliptic Variational Problems*. Lecture Notes in Mathematics, Vol. 1489. Springer-Verlag, Berlin, Heidelberg.
- Milnor, J. (1969): *Topology from the Differentiable Point of View*. University of Virginia Press, Charlottesville, VA.
- Monastirsky, M. (1993): *Topology of Gauge Fields and Condensed Matter*. Plenum Press, New York.
- Müller, I. and Rugeri, T. (1993): *Extended Thermodynamics*. Springer-Verlag, Berlin, Heidelberg.
- Murray, J. (1989): *Mathematical Biology*. Springer-Verlag, Berlin, Heidelberg.
- Nakahara, M. (1990): *Geometry, Topology, and Physics*. Hilger, Bristol.
- Nečas, J. (1967): *Les méthodes directes en théorie des équations elliptiques*. Academia, Prague.
- Nishikawa, K. and Wakatani, M. (1993): *Plasma Physics: Basic Theory with Fusion Applications*. Springer-Verlag, Berlin, Heidelberg.
- Nobel Prizes (1954ff): *Nobel Lectures*. Edited by the Nobel Foundation, Stockholm.
- Novikov, S., Manakov, S., Pitajevskii, L., and Zakharov, V. (1984): *Theory of Solitons*. Plenum Press, New York (translation from Russian).
- Oberguggenberger, M. (1992): *Multiplication of Distributions and Applications to Partial Differential Equations*. Harlow, Longman, UK.

- Peierls, R. (1979): *Surprises in Theoretical Physics*. Princeton University Press, Princeton, NJ.
- Petryshyn, V. (1993): *Approximation-Solvability of Nonlinear Functional and Differential Equations*. Marcel Dekker, New York.
- Plakida, N. (1994): *High-Temperature Superconductivity: Experiment and Theory*. Springer-Verlag, New York.
- Polyakov, A. (1987): *Gauge Fields and Strings*. Academic Publishers, Harwood, NJ.
- Pressley, A. and Segal, G. (1988): *Loop Groups*. Oxford, Clarendon Press.
- Rabinowitz, P. (1986): *Methods in Critical Point Theory with Applications*. Amer. Math. Soc., Providence, RI.
- Racke, R. (1992): *Lectures on Evolution Equations*. Vieweg, Braunschweig.
- Raychaudhuri, A., Banerji, S. and Banerjee, A. (1993): *General Relativity, Astrophysics, and Cosmology*. Springer-Verlag, New York.
- Reed, M. and Simon, B. (1972): *Methods of Modern Mathematical Physics*. Vol. 1: *Functional Analysis*; Vol. 2: *Fourier Analysis, Self-Adjointness*; Vol. 3: *Scattering Theory*; Vol. 4: *Analysis of Operators*. Academic Press, New York.
- Renardy, M. and Rogers, R. (1993): *Introduction to Partial Differential Equations*. Springer-Verlag, New York.
- Riesz, F. and Nagy, B. (1955): *Leçons d'analyse fonctionnelle* (English edition: *Functional Analysis*). Frederick Ungar Publishing Company, New York, 1978).
- Rolewicz, S. (1972): *Metric Linear Spaces*. Polish Scientific Publishers, Warsaw.
- Rolnick, W. (1994): *Fundamental Particles and Their Interactions*. Addison-Wesley, Reading, MA.
- Rowlatt, P. (1966): *Group Theory and Elementary Particles*. Elsevier, New York.
- Rudin, W. (1966): *Real and Complex Analysis*. McGraw-Hill, New York.
- Rudin, W. (1973): *Functional Analysis*. McGraw-Hill, New York.
- Ruei, K. (1971): *Quantum Theory of Particles and Fields*, Vol. 1, 2. University Press, Taipei, Taiwan.
- Ruei, K. (1972): *Classical Theory of Particles and Fields*, Vols. 1, 2. University Press, Taipei, Taiwan.
- Sakai, A. (1991): *Operator Algebras*. Cambridge University Press, Cambridge, UK.
- Sattinger, D. and Weaver, O. (1993): *Lie Groups, Lie Algebras, and Their Representations*. Springer-Verlag, New York.
- Schechter, M. (1971): *Principles of Functional Analysis*. Wiley, New York.

- Schmutzer, E. (1989): *Grundlagen der theoretischen Physik*, Vols. 1, 2. Deutscher Verlag der Wissenschaften, Berlin.
- Schneider, P., Ehlers, J., and Falco, E. (1992): *Gravitational Lenses*. Springer-Verlag, New York.
- Schrieffer, J. (1964): *Theory of Superconductivity*. Benjamin, New York.
- Schwartz, L. (1966): *Théorie des distributions*. Hermann, Paris.
- Seydel, R. (1994): *Practical Bifurcation and Stability Analysis: From Equilibrium to Chaos*. Springer-Verlag, Berlin, Heidelberg.
- Shore, S. (1992): *An Introduction to Astrophysical Hydrodynamics*. Academic Press, San Diego, CA.
- Simon, B. (1993): *The Statistical Mechanics of Lattice Gases*. Princeton University Press, Princeton, NJ.
- Sirovich, L. (ed.) (1991): *New Perspectives in Turbulence*. Springer-Verlag, New York.
- Sirovich, L. (ed.) (1994): *Trends and Perspectives in Applied Mathematics*. Springer-Verlag, New York.
- Smale, S. (1965): An infinite-dimensional version of Sard's theorem. *Amer. J. Math.* **87**, 861–866.
- Smoller, J. (1994): *Shock Waves and Reaction-Diffusion Equations*. 2nd enlarged edition. Springer-Verlag, New York.
- Soper, D. (1975): *Classical Field Theory*. Wiley, New York.
- Spohn, H. (1991): *Large Scale Dynamics of Interacting Particles*. Springer-Verlag, Berlin, Heidelberg.
- Stephani, H. (1989): *Differential Equations: Their Solutions Using Symmetries*. Edited by MacCallum. Cambridge University Press, Cambridge, UK.
- Sterman, G. (1993): *An Introduction to Quantum Field Theory*. Cambridge University Press, Cambridge, UK.
- Straub, D. (1989): *Thermofluid Dynamics of Optimized Rocket Propulsions*. Birkhäuser, Basel.
- Struwe, M. (1988): *Plateau's Problem and the Calculus of Variations*. Princeton University Press, Princeton, NJ.
- Struwe, M. (1990): *Variational Methods*. Springer-Verlag, New York.
- Sunder, V. (1987): *An Invitation to von Neumann Algebras*. Springer-Verlag, New York.
- Ta-Pai Cheng and Ling-Fong Li (1984): *Gauge Theory of Elementary Particle Physics*. University Press, Oxford.
- Temam, R. (1977): *Navier-Stokes Equations: Theory and Numerical Analysis*. North-Holland, Amsterdam.

- Temam, R. (1988): *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*. Springer-Verlag, New York.
- ter Haar Romeny, B. (ed.) (1994): *Geometry-Driven Diffusion in Computer Vision*. Kluwer, Dordrecht.
- Thaller, B. (1992): *The Dirac Equation*. Springer-Verlag, Berlin, Heidelberg.
- Thirring, W. (1991): *A Course in Mathematical Physics*. Vol. 1: *Classical Dynamical Systems*; Vol. 2: *Classical Field Theory*; Vol. 3: *Quantum Mechanics of Atoms and Molecules*; Vol. 4: *Quantum Mechanics of Large Systems*. Springer-Verlag, New York (translation from German).
- Triebel, H. (1985): *Analysis and Mathematical Physics*. Teubner-Verlag, Leipzig.
- Triebel, H. (1992): *Theory of Function Spaces II*. Birkhäuser, Basel.
- Tromba, A. (1977): *On the Number of Simply Connected Minimal Surfaces Spanning a Curve*. Mem. Am. Math. Soc., Providence, RI.
- Vanhorn, W. (1994): *The Stokes Equation*. Akademie-Verlag, Berlin.
- Visintin, A. (1994): *Differentiable Models of Hysteresis*. Springer-Verlag, Berlin, Heidelberg.
- Wahl, W. von (1985): *The Equations of Navier-Stokes and Abstract Parabolic Equations*. Vieweg, Braunschweig.
- Waldschmidt, M., Moussa, P., Luck, J., and Itzykson, C. (1992). *From Number Theory to Physics*. Springer-Verlag, Berlin, Heidelberg.
- Weinberg, S. (1992): *Dreams of a Final Theory*. Pantheon Books, New York.
- Wendland, W. (1996): *Integral Equation Methods for Boundary-Value Problems*. Springer-Verlag, Berlin, Heidelberg (to appear).
- Wiedemann, H. (1993): *Particle Accelerator Physics*, Vols. 1, 2. Springer-Verlag, Berlin, Heidelberg.
- Wightman, A. and Velo, G. (1980): *Rigorous Atomic and Molecular Physics*. Plenum Press, New York.
- Wloka, J. (1971): *Funktionalanalysis und ihre Anwendungen*. De Gruyter, Berlin.
- Yosida, K. (1991): *Lectures on Differential and Integral Equations*. Dover, New York.
- Yosida, K. (1988): *Functional Analysis*. 5th edition. Springer-Verlag, New York.
- Zabczyk, J. (1992): *Optimal Control Theory*. Birkhäuser, Basel.
- Zeidler, E. (1972): *Beiträge zur Theorie und Praxis freier Randwertaufgaben*. Akademie-Verlag, Berlin.

- Zeidler, E. (1977): *Bifurcation Theory and Permanent Waves*. In: P. Rabinowitz (ed.), *Applications of Bifurcation Theory*, Academic Press, New York, pp. 203–224.
- Zeidler, E. (1986): *Nonlinear Functional Analysis and Its Applications*. Vol. 1: *Fixed-Point Theorems*; Vol. 2A: *Linear Monotone Operators*; Vol. 2B: *Nonlinear Monotone Operators*; Vol. 3: *Variational Methods and Optimization*; Vols. 4, 5: *Applications to Mathematical Physics*. Springer-Verlag, New York (second enlarged editions of Vols. 1 and 4, 1992 and 1995, Vol. 5 in preparation.)
- Zeidler, E. (1995): *Teubner-Taschenbuch der Mathematik II* (Handbook on Advanced Mathematics) (Chapters 3 through 15: Linear and nonlinear functional analysis; dynamical systems; nonlinear partial differential equations in mathematical physics; analysis on manifolds; Riemannian geometry and general relativity; Lie groups, Lie algebras, and elementary particles; algebraic and differential topology; fibre bundles, modern differential geometry and gauge field theory; characteristic classes and the Atiyah–Singer index theorem; the Riemann–Roch–Hirzebruch theorem). Cf. Grosche, Ziegler, and Zeidler (eds.) (1995) (English edition in preparation).
- Zeidler, E. (1995): *Applied Functional Analysis: Applications to Mathematical Physics*. Applied Mathematical Sciences Vol. 108. Springer-Verlag, New York.
- Zinn-Justin, J. (1993): *Quantum Field Theory and Critical Phenomena*. Oxford University Press, Oxford.
- Zuily, C. (1988): *Problems in Distributions and Partial Differential Equations*. North-Holland, Amsterdam.

List of Symbols

Science is a first class piece of furniture for the bel étage — as long as common sense reigns on the ground floor.

Oliver Wendell Holmes

General Notation

$\mathcal{A} \Rightarrow \mathcal{B}$	\mathcal{A} implies* \mathcal{B}
iff	if and only if
$\mathcal{A} \Leftrightarrow \mathcal{B}$	\mathcal{A} iff \mathcal{B} (i.e., $\mathcal{A} \Rightarrow \mathcal{B}$ and $\mathcal{B} \Rightarrow \mathcal{A}$)
$f(x) := 2x$	$f(x) = 2x$ by definition
$x \in S$	x is an element of the set S
$x \notin S$	x is <i>not</i> an element of the set S
$\{x: \dots\}$	set of all elements x with the property \dots
$S \subseteq T$	the set S is contained in the set T
$S \subset T$	$S \subseteq T$ and $S \neq T$ (the set S is <i>properly</i> contained in T)
$S \cup T$	the union of the sets S and T (the set of all elements that live in S <i>or</i> T)
$S \cap T$	the intersection of the sets S and T (the set of all elements that live in S <i>and</i> T)

*One says that

- (i) condition \mathcal{A} is *sufficient* for \mathcal{B} , and
- (ii) condition \mathcal{B} is *necessary* for \mathcal{A} .

$S - T$	the difference set (the set of all elements that live in S and <i>not</i> in T)
\emptyset	empty set
2^S	set of all subsets of S (the power set of S)
$S \times T$	product set $\{(x, y) : x \in S \text{ and } y \in T\}$
$\{p\}$	set of the single point p
\mathbb{N}	set of the natural numbers $1, 2, \dots$
$\mathbb{R}, \mathbb{C}, \mathbb{Q}, \mathbb{Z}$	set of the real, complex, rational, integer numbers
\mathbb{K}	\mathbb{R} or \mathbb{C}
\mathbb{R}^N	set of all real N -tupels $x = (x_1, \dots, x_N)$ (i.e., $x_j \in \mathbb{R}$ for all j)
\mathbb{C}^N	set of all complex N -tupels (x_1, \dots, x_N) (i.e., $x_j \in \mathbb{C}$ for all j)
\mathbb{K}^N	\mathbb{R}^N or \mathbb{C}^N
$\operatorname{Re} z, \operatorname{Im} z$	real part of the complex number $z = x + yi$, imaginary part of z (i.e., $\operatorname{Re} z := x$, $\operatorname{Im} z := y$)
\bar{z}	conjugate complex number $\bar{z} := x - yi$,
$ z $	absolute value of the complex number z , $ z := \sqrt{x^2 + y^2}$
$[a, b]$	closed interval (the set $\{x \in \mathbb{R} : a \leq x \leq b\}$)
$]a, b[$	open interval (the set $\{x \in \mathbb{R} : a < x < b\}$)
$]a, b]$	half-open interval (the set $\{x \in \mathbb{R} : a < x \leq b\}$)
$[a, b[$	half-open interval (the set $\{x \in \mathbb{R} : a \leq x < b\}$)
$\operatorname{sgn} r$	signum of the real number r
δ_{jk}	Kronecker symbol, $\delta_{jk} := 1$ if $j = k$, and $\delta_{jk} := 0$ if $j \neq k$
$\inf S$	infimum of the set S of real numbers (the largest lower bound of S)
$\sup S$	supremum of the set S of real numbers (the smallest upper bound of S)
$\min S$	the minimum of the set S of real numbers (the smallest upper bound of S)
$\max S$	the maximum of the set S of real numbers (the largest element of S)
$\liminf_{n \rightarrow \infty} a_n$	lower limit of the real sequence (a_n) (see page 136)
$\limsup_{n \rightarrow \infty} a_n$	upper limit of the real sequence (a_n)

The Landau Symbols

$f(x) = O(g(x)),$ $x \rightarrow a$	$ f(x) \leq \operatorname{const} g(x) $ for all x in a neighborhood of the point a
$f(x) = o(g(x)),$ $x \rightarrow a$	$\lim_{x \rightarrow a} \frac{f(x)}{g(x)} = 0$

Norms and Inner Products

$\ x\ $	norm of x	I6*
$\lim_{n \rightarrow \infty} x_n = x$ (or $x_n \rightarrow x$ as $n \rightarrow \infty$)	the sequence (x_n) converges to the point x	I8
$\sum_{n=1}^{\infty} u_n$	infinite series in a Banach space	I75
$(x y)$	inner product	I103
$\langle x y \rangle$	Euclidean inner product, $\langle x y \rangle := \sum_{n=1}^N \bar{x}_n y_n$ (\bar{y}_j conjugate complex number to y_j)	I107
$ x $	Euclidean norm, $ x := \langle x x \rangle^{\frac{1}{2}} = \left(\sum_{n=1}^N x_n ^2 \right)^{\frac{1}{2}}$	I107
$ x _{\infty}$	special norm, $ x _{\infty} := \sup_n x_n $	I10
$(u v)_2$	inner product in the Lebesgue spaces $L_2(G)$ and $L_2^{\mathbb{C}}(G)$, $(u v)_2 := \int_G \overline{u(x)} v(x) dx$	I112
$\ u\ _2$	norm on the Lebesgue spaces $L_2(G)$ and $L_2^{\mathbb{C}}(G)$, $\ u\ _2 := (u u)^{\frac{1}{2}} = \left(\int_G u(x) ^2 dx \right)^{\frac{1}{2}}$	I112
$(u v)_{1,2}$	inner product on the Sobolev space $W_2^1(G)$, $(u v)_{1,2} := \int_G \left(uv + \sum_{j=1}^N \partial_j u \partial_j v \right) dx$	I118
$\ u\ _{1,2}$	norm on the Sobolev space $W_2^1(G)$, $\ u\ _{1,2} := (u u)_{1,2}^{\frac{1}{2}} = \left(\int_G \left(u^2 + \sum_{j=1}^N (\partial_j u)^2 \right) dx \right)^{\frac{1}{2}}$	I118
$(\cdot \cdot)_E$	energetic inner product	I271
$\ \cdot\ _E$	energetic norm	I271

Operators

$A: S \subseteq X \rightarrow Y$	operator from the set S into the set Y , where $S \subseteq X$	I16
$D(A)$ (or $\text{dom } A$)	domain of definition of the operator A	I17
$R(A)$ (or $\text{im } A$)	range (or image) of the operator A	I17
$N(A)$ (or $\ker A$)	null space (or kernel) of the operator A , $N(A) := \{x: Ax = 0\}$	I69

*If we add the symbol I, then the page number refers to AMS Vol. 108.

I (or id)	identical operator, $Ix := x$ for all x	I76
$A(S)$	image of the set S , $A(S) := \{Ax: x \in S\}$	
$A^{-1}(T)$	preimage of the set T , $A^{-1}(T) := \{x: Ax \in T\}$	I17
A^{-1}	inverse operator to A	I17
$G(A)$	graph of the operator A , $G(A) := \{(x, Ax): x \in D(A)\}$	I412
$\ A\ $	norm of the linear operator A	I69
$\ f\ $	norm of the functional f	I74
AB (or $A \circ B$)	the product of the operators A and B , $(AB)(u) := A(Bu)$	I28
$A \subseteq B$	the operator B is an extension of the operator A	I258
A^*	adjoint operator to the linear operator A	I261
A^T	dual operator to the linear operator A	I99
\bar{A}	closure of the linear operator A	I412
$\sigma(A)$	spectrum of the linear operator A	I82
$\rho(A)$	resolvent set of the linear operator A	I82
$r(A)$	spectral radius of the linear operator A	I93
rank A	rank of the linear operator A , rank $A := \dim R(A)$	195
ind A	index of the linear operator A , ind $A := \dim N(A) - \text{codim } R(A)$	292
det A	determinant of the matrix A	
tr A	trace of the $(N \times N)$ -matrix $A = (a_{jk})$, tr $A := a_{11} + \dots + a_{NN}$	
tr A	trace of the linear operator A in a Hilbert space	I45

Special Sets

\bar{S}	closure of the set S	I30
int S	interior of the set S	I30
ext S	exterior of the set S	I31
∂S	boundary of the set S	I31
$U_\varepsilon(p)$	ε -neighborhood of the point p in a normed space, $U_\varepsilon(p) := \{x \in X: \ x - p\ < \varepsilon\}$	I15
$U(p)$	neighborhood of the point p	I15
$\dim X$	dimension of the linear space X	I5
$X_{\mathbb{C}}$	complexification of the linear space X	I97
X/L	factor space	I84
codim L	codimension of the linear subspace L , codim $L := \dim(X/L)$	I91
L^\perp	orthogonal complement to the linear subspace L	I163
αS	the product $\alpha S := \{\alpha x: x \in S\}$, $\alpha \in \mathbb{R}, \mathbb{C}$	I6

$S + T$	the sum $S + T := \{x + y: x \in S \text{ and } y \in T\}$	I6
$M \oplus L$	orthogonal direct sum ($X = M \otimes L$, where $L = M^\perp$)	I163
$M \oplus L$	direct sum*	I88
$X \otimes Y$	tensor product	I222
X^*	dual space	I74
X_E	energetic space	I271
$\text{span } S$	linear hull of the set S	I30
$\overline{\text{span}} \, S$	closed linear hull of the set S	I30
$\text{co } S$	convex hull of the set S	I31
$\overline{\text{co }} S$	closed convex hull of the set S	
$\text{dist}(p, S)$	distance of the point p from the set S	I47
$\text{diam } S$	diameter of the set S	I46
$\text{meas } S$	measure of the set S	I429
$\delta(x)$	the Dirac delta function	I156
δ	the delta distribution	I159

Derivatives

$u'(t)$	derivative of an operator function $u = u(t)$ at time t	I79
$\partial_j f$	partial derivative $\frac{\partial f}{\partial x_j}$	
$\partial^\alpha f$	$\partial_1^{\alpha_1} \partial_2^{\alpha_2} \cdots \partial_N^{\alpha_N} f$, where $\alpha = (\alpha_1, \dots, \alpha_N)$ (the classical symbols are also used for the derivatives of generalized functions)	I157
$ \alpha $	the sum $\alpha_1 + \cdots + \alpha_N$	I157
$\frac{\partial}{\partial n}$	derivative in the direction of the exterior normal	I179
Δf	Laplacian, $\Delta f := \sum_{n=1}^N \partial_j^2 f$	I123
$\delta F(x; h)$	variation of the functional F at the point x in direction of h	43
$\delta^n F(x; h)$	n th variation of the functional F at the point x in the direction of h	43
$A'(x)$ (or $dA(x)$)	Fréchet-derivative of the operator A at the point x	228
$d^n A(x)(h_1, \dots, h_n)$	n th Fréchet-differential of the operator A at the point x in the direction of h_1, \dots, h_n	229

*To simplify notation, we will use the same symbol $M \oplus L$ for the direct sum, the topological direct sum, and the orthogonal direct sum. The text always refers to the momentary meaning.

Integral

$\int_G u(x)dx$	the Lebesgue integral that comprehends and generalizes the classical integral	I432
$\int_a^b F(x)dg(x)$	Lebesgue–Stieltjes integral	I439
$\int_{-\infty}^{\infty} F(\lambda)dE_{\lambda}$	operator-valued Lebesgue–Stieltjes integral (with respect to the spectral family $\{E_{\lambda}\}$)	I330

Spaces of Continuous Functions

$C[a, b]$, $C(\bar{G})$	I13, I114
$L(X, Y)$, $L_{\text{inv}}(X, Y)$	I72, I78

Spaces of Hölder Continuous Functions

$C^{\alpha}[a, b]$, $C^{k,\alpha}[a, b]$, $C^{\alpha}(\bar{G})$, $C^{k,\alpha}(\bar{G})$ ($C^{\alpha}(\bar{G}) = C^{0,\alpha}(\bar{G})$)	I95ff
-----------------------------------------------------------------------------------------------------------------------------------------------	-------

Spaces of Smooth Functions

$C^k[a, b]$, $C^k(G)$, $C^k(\bar{G})$, $C^{\infty}(G)$, $C^k(G)_{\mathbb{C}}$ ($C^0(\bar{G}) := C(\bar{G})$)	I95, I114
$C_0^{\infty}(G)$ (or $\mathcal{D}(G)$), \mathcal{S}	I114, I212

Spaces of Integrable Functions (Lebesgue Spaces)

$L_2(a, b)$, $L_2(G)$, $L_2^{\mathbb{K}}(G)$ ($L_2(G) := L_2^{\mathbb{K}}(G)$ if $\mathbb{K} = \mathbb{R}$)	I128, I112
$L_p(G)$	355

Sobolev Spaces

$W_2^1(G)$, $\overset{\circ}{W}_2^1(G)$	I128, I129
$W_p^m(G)$	357

Spaces of Sequences

\mathbb{K}^{∞} , $l_{\infty}^{\mathbb{K}}$, $l_2^{\mathbb{K}}$ ($l_2 := l_2^{\mathbb{K}}$ if $\mathbb{K} = \mathbb{R}$)	I94, I175
l_p	354

Spaces of Distributions

$\mathcal{D}'(G)$, \mathcal{S}'	I158, I217
$B_{pq}^s(\mathbb{R}^N)$, $F_{pq}^s(\mathbb{R}^N)$	361
$B_{pq}^s(G)$, $F_{pq}^s(G)$	362

List of Theorems

Do not imagine that mathematics is hard and crabbed, and repulsive to common sense. It is merely the etherealization of common sense.

Lord William Kelvin (1824–1907)

Theorem 1.A	(The Hahn–Banach theorem for linear spaces)	2
Theorem 1.B	(The Hahn–Banach theorem for normed spaces)	4
Theorem 1.C	(Separation of convex sets)	8
Theorem 1.D	(The minimum norm problem and duality)	15
Theorem 2.A	(Necessary and sufficient conditions for local minima)	45
Theorem 2.B	(Lack of compactness in infinite-dimensional Banach spaces)	48
Theorem 2.C	(The existence of weakly convergent subsequences)	50
Theorem 2.D	(The existence theorem for convex minimum problems)	53
Theorem 2.E	(Variational inequalities)	66
Theorem 2.F	(Saddle points and duality)	73
Theorem 2.G	(Existence of a saddle point)	76
Theorem 2.H	(Existence of a quasi-minimal point)	84
Theorem 2.I	(The minimum principle via the Palais–Smale condition)	86
Theorem 2.J	(The mountain pass theorem)	88
Theorem 2.K	(The main theorem on monotone operators)	93

Theorem 2.L	(Symmetry and the Noether theorem on conservation laws)	101
Theorem 2.M	(The basic equations of gauge field theory)	125
Theorem 3.A	(The Baire theorem)	170
Theorem 3.B	(The uniform boundedness theorem)	172
Theorem 3.C	(The open mapping theorem)	178
Theorem 3.D	(The closed graph theorem)	183
Theorem 3.E	(The closed range theorem)	210
Theorem 4.A	(Differentiation of analytic operators)	236
Theorem 4.B	(The fundamental theorem of calculus)	242
Theorem 4.C	(The Taylor theorem)	244
Theorem 4.D	(The chain rule)	248
Theorem 4.E	(The implicit function theorem)	251
Theorem 4.F	(The inverse mapping theorem)	259
Theorem 4.G	(The linearization principle)	262
Theorem 4.H	(The surjective implicit function theorem)	269
Theorem 5.A	(Duality for linear compact operators)	284
Theorem 5.B	(The Riesz–Schauder theory on Hilbert spaces)	286
Theorem 5.C	(The Riesz–Schauder theory on Banach spaces)	295
Theorem 5.D	(The spectrum of linear compact operators)	296
Theorem 5.E	(Compact perturbations of Fredholm operators)	300
Theorem 5.F	(The product index theorem)	301
Theorem 5.G	(The Fredholm alternative)	304
Theorem 5.H	(The main theorem of bifurcation theory)	311
Theorem 5.I	(The Smale principle)	318
Theorem 5.J	(The stationary Navier–Stokes equations)	337

List of the Most Important Definitions

The collection of all our experiences consists of what we know and what we have forgotten.

Marie von Ebner-Eschenbach (1830–1916)

Spaces

linear space	I5*
dimension	I5
linear subspace	I30
Banach space	I10
norm	I6
separable	I83
reflexive	61
Hilbert space	I105
inner product	I103
orthogonal elements	I103
orthogonal projection	I163
complete orthonormal system	I198
Fock space (bosons or fermions)	I361
Lebesgue space	I112, 355
Sobolev space	I128, 357
energetic space	I271

*If we add the symbol I, then the page number refers to AMS Vol. 108.

Triebel–Lizorkin space	361
Hölder space	I95
Besov space	361
dual space	I74
metric space	34
topological space	28

Convergence

norm convergence (strong convergence)	I8
Cauchy sequence	I9
weak convergence	49
sequentially continuous	I26
sequentially compact	I33
relatively sequentially compact	I33

Operators

domain of definition	I17
range and preimage	I17
injective	I17
surjective	I17
bijection	I17
inverse operator	I17
compact	I39
continuous	I26
k -contraction	I19
Lipschitz continuous	I27
Hölder continuous	I95
homeomorphism	I28
diffeomorphism	259
submersion	262
subimmersion	262
immersion	262
analytic	236
monotone	93
coercive	93
weakly coercive	53
linear	I69
symmetric	I262
the Friedrichs extension	I277
adjoint	I261
dual	199
self-adjoint	I262
Hamiltonian	I326

orthogonal projection operator	I268
skew-adjoint	I262
unitary	I210
Fourier transformation	I214
trace class	I345
statistical state	I346
statistical operator	I348
Hilbert–Schmidt operator	I345
semigroup	I296
Green function (propagator)	I384
one-parameter group	I296
dynamics of a quantum system	I326
Fredholm alternative	288
linear Fredholm operator	292
index	292
nonlinear Fredholm operator	317
m -linear bounded	227

Functional

nonlinear	43
linear	I74
convex	53
strictly convex	53
concave	75
bilinear form	I18
bounded	I118
symmetric	I118
distribution (generalized function)	I158
tempered distribution	I217
Fourier transformation	I218
generalized eigenfunction	I342
Dirac delta distribution	I159
Green function	I158
fundamental solution	I181
Palais–Smale condition	86

Embedding

continuous	323
compact	323

Spectrum

eigenvalue and eigenvector	I82
----------------------------	-----

generalized eigenvector	I342
resolvent set	I82
resolvent operator	I82
essential spectrum	I83
spectral family	I331
measurements in quantum systems	I341

Set

open	I15
neighborhood	I15
interior	I30
closed	I15
closure	I30
boundary	I31
compact or relatively compact (in normed spaces)	I33
(in topological spaces)	30
dense	I83
nowhere dense	169
first and second Baire category	169
convex	I29
bounded	I33
countable	I84

Point

critical point	44
local minimum	44
local maximum	44
saddle point	72
bifurcation point	309
fixed point	I18

Operator Algebras

Banach algebra	I75
von Neumann algebra	I357
C^* -algebra	I355
observable	I357
state	I356
pure	I356
mixed	I356
KMS-state (thermodynamic equilibrium)	I358
$*$ -automorphism	I356
dynamics of a quantum system	I357

Derivative

time derivative	I79
n th-variation	43
Fréchet-differential	228
Fréchet derivative (F -derivative)	228
partial Fréchet-derivative (F -derivative)	232
generalized derivative of a function	I127
derivative of a distribution	I160

Integral

Riemann integral for vector-valued functions	239
Lebesgue integral	I432
Lebesgue measure	I427
integration by parts	I157
Lebesgue–Stieltjes integral	I439
Feynman path integral	I382

Subject Index

The reader should also consult the index to AMS Volume 108.

- adjoint equation 201
- adjoint representation 111
- algebraic complement 188
- algebraic projection 188
- analytic operators 233, 236
- angular-momentum conservation 159
- antilinear 61
- antiquarks 115, 119
- applications to cubature formulas 175
- Arzelà–Ascoli theorem 35
- associated vector bundle 130
- Atiyah–Singer index theorem 283

- Baire theorem 169
- Banach–Steinhaus theorem 173
- bang-bang controls 28
- baryon number 113
- baryons 114
- Bénard problem 368
- Besov spaces 360

- Bianchi identity 125
- bifurcation 367
- bifurcation point 309
- bifurcation theory 309
- bilinear operators 231
- biorthogonal systems 196
- bosonic string 158
- bosons 163
- boundary-value problems 315
- brachistochrone 132
- Brouwer fixed-point theorem 94
- buckling of beams 370

- C^* -algebras 36
- C^r -diffeomorphism 259
- calculus of variations 56
- canonical mapping 186
- Cantor's nested interval principle 169
- capillary surfaces 144
- chain rule 247
- characteristic number 314
- Čebyšev approximation 18

- Clifford algebra 124
 closed 29
 closed range theorem 210
 cluster point 137
 codimension 191
 coercive 53
 coerciveness condition 57
 coin game 81
 commutation relations 162
 compact embedding 323
 compactness 30
 complete 31
 completion principle 23
 compressibility equation 363
 concave 75
 conformal coordinates 150
 conformal gauge 160
 connection 130
 conservation of energy 99
 conservation law 98, 135, 159
 constraints 158
 continuity 30
 continuous inverse theorem 179
 continuous projection 188
 control equation 26
 control functional 26
 control restriction 26
 convergence 29
 convex approximation models
 142
 Cooper pairs 155
 coupling constant 124
 covariant derivatives 124
 covariant directional derivative
 132
 critical point 44
 curvature 130

 defects 155
 deformation 139
 deformation tensor 140
 density and duality 37
 density of the fluid 146
 density of the outer force 140
 diagonal procedure 51

 diffeomorphisms 259
 differential 228
 differential geometry 130
 differential operators 234
 Dirac equation 126
 direct sum 188
 Dirichlet problem 134
 distance 34
 dual equation 201
 dual operators 199
 dual pair 303
 dual representation 120
 dual space 5
 duality functor 203
 duality for linear compact
 operators 284
 duality theory 73
 dyadic partition of unity 360

 eigensolutions 291
 eigenvalue problems 59
 Ekeland principle 83
 elastic energy 140
 elasticity 139
 electromagnetic field tensor 126
 elementary particle physics 106
 elementary particles 112
 embeddings 221
 embedding operator 323
 energetic space 24
 energy-momentum conservation
 159
 equicontinuous 35
 equilibrium condition 141
 equilibrium equation 141
 equivalent maps 261
 Euler equation 41
 Euler-Lagrange equation 46
 Euler-Lagrange system 48
 exact Banach sequence 205
 exact sequences 205
 exterior product 279
 extreme point 36

 F -derivative 228

- factor space 184
 farmer's allocation problem 27
 fat 169
 fermions 163
 fiber 131
 finite intersection property 31
 first category 169
 first Piola–Kirchhoff stress tensor 140
 formation of patterns 367
 four-potential 126
 fractional Sobolev space 362
 Fréchet derivative 43, 228
 Fredholm alternative 288, 304
 Fredholm operator 292
 free boundary problems 155
 free surface 146
 function spaces 359
 functional 43
 fundamental theorem of calculus 242

G-derivative 275
 Galerkin equation 96
 game theory 81
 Gâteaux derivatives 43, 275
 gauge field theory 102, 122
 gauge invariance 103, 127
 gauge transformations 103, 127
 Gelfand theorem 37
 Gell–Mann–Nishijima formula 112
 generalized surface area 151
 ghost state 164
 global gauge transformation 103
 global inverse mapping theorem 277
 GNS-theorem 36
 graph-closed 182
 Grassmann algebra 278
 gravitational constant 138
 growth condition 57

 Haar condition 25
 Hahn–Banach theorem 2

 half-spaces 6
 hanging rope 133
 harmonic maps 151, 158
 Helmholtz–Weyl decomposition 366
 Higgs field 154
 Hilbert spaces 16
 Hölder inequality 352
 Hölder spaces 362
 Hopf bifurcation 369
 hypercharge 113
 hyperplane 6

 index 292
 immersion 262
 implicit function theorem 251
 incompressibility condition 330
 inertial system 123
 infinite-dimensional Lie algebra 162
 integration 239
 integral equations 291, 313
 integral operators 233
 interpolation inequalities 322
 inverse mapping theorem 251
 isospin 112
 iterated derivatives 244

 Jacobi's identity 107
 Jordan curve 150
 Jordan normal form 350

 kinetic energy 139
 Krein theorem 24
 Krein–Milman convexity theorem 36

 Ladyzhenskaya inequality 326
 Lagrange multiplier rule 143, 270
 Lagrangian 46
 Landau–Ginzburg model 152
 Lebesgue spaces 362
 Leray–Schauder principle 342
 Lie algebra 107
 Lie group 130

- Lie product 107
 linear optimization 36
 linearization 225
 linearization principle 261
 local boundedness 95
 local diffeomorphism 259
 local inverse mapping theorem 259
 local maximum 44
 local minimum 44
 locally convex spaces 222
 Lorentz group 160
 Lorentz transformation 159
 lower semicontinuous 55
 lower semicontinuous functionals 136
m-linear bounded operators 227
 mapping degree 153
 matrix equation 201
 Maxwell equations 126
 Maxwell–Dirac equation 127
 meager 169
 mean curvature 147
 mesons 119
 metric space 34
 minimal surfaces 134, 149
 minimax theorem 75
 minimum norm problem 15
 Minkowski functional 9
 Minkowski inequality 352
 mixed state of three quarks 113
 mixed strategy 82
 moment problem 13
 monotone operators 93
 monotonicity trick 94
 mountain pass theorem 87
 multilinear forms 278
 multilinearization 225, 230
N-body problem 138
 n th *F*-derivative 246
 NASA 147
 natural boundary condition 148
 Navier–Stokes equations 329
 neighborhood 29
 neutron 112
 Newtonian equation 138
 Noether theorem 98, 156, 159
 noncollision solutions 138
 nondifferential continuous function 171
 nonlinear elasticity 139
 nonlinear Fredholm operators 317, 350
 normal form 265
 normed spaces 4
 normisomorphic 353
 nowhere dense 169
 open 29
 open mapping theorem 178
 optimal control of rockets 20
 optimal strategy pair 81
 order parameter 154
 orthogonal complement 212
 Palais–Smale condition (PS) 86
 parallel transport 131
 parametrix 298
 partial *F*-derivative 232, 276
 phase transformations 103
 phase transition 154
 Plateau problem 150
 Poincaré transformations 159
 Poincaré–Friedrichs inequality 58
 polyconvex material 142
 Pontrjagin maximum principle 25
 potential energy 139
 power operator 233
 precompact 31
 pressure 336
 pressure equation 365
 principle fiber bundle 130
 principle of gauge invariance 106
 principle of minimal potential energy 71, 134
 principle of relativity 123

- principle of stationary action 48,
 165
 probabilistic coin game 83
 product index theorem 301
 product rule 250
 product space 180
 projections 188
 proper 277, 317
 proton 112
 pseudo-orthogonal complements
 197

 quadratic variational inequalities
 70
 quantization of the bosonic
 string 164
 quantum electrodynamics 127
 quantum field theory 107
 quantum numbers 112
 quarks 114, 119
 quasi-convex 55
 quasi-minimal points 83
 quasi-solutions 84

 rank 195
 rank theorem 263
 reflexive 61
 reflexivity 220
 regular point 318
 regular value 318
 regularity up to the boundary
 151
 regularized problem 84
 relative adhesion coefficient 146
 relatively open 29
 relativistic particles 135
 renormalization of energy 155
 representation 110
 Reynolds numbers 369
 Riesz theorem 41
 Riesz–Schauder theory 286, 295

 saddle point 72
 Sard’s theorem 318
 Sard–Smale theorem 321

 second differential 228
 secondary category 169
 seminorm 222
 separated 29
 separation of convex sets 6
 sequentially compact 31
 sequentially lower semicontinuous functionals
 137
 short exact Banach sequence 207
 side condition 145
 singular point 318
 skew-adjoint 108
 Smale principle 318
 Sobolev space 56, 356
 space shuttle 147
 space–time manifold 157
 spectrum 296
 splits 189
 splitting subspaces 196
 standard model 106
 stationary 47
 step function 239
 Stieltjes integral 10, 11
 Stone–Weierstrass approximation theorem
 35
 stored energy function 140
 stored potential energy 140
 strangeness 113
 strategy set 81
 stress force 141
 stress tensor 140
 strictly convex 53
 string theory 156
 strong convergence 49
 strong pressure equation 366
 subalgebra 107
 subimmersion 262
 submersion 262
 subsequences 219
 sum rule 232
 supercommutativity 279
 superconducting state 154
 superconductor 154

- supercooled Helium 154
- superfluidity 154
- supermathematics 163
- superposition operator 273
- supernumbers 162
- superstring theory 165
- surface tension 146
- surjective implicit function theorem 269
- symmetries 98

- Taylor problem 368
- Taylor theorem 243
- tensor algebra 278
- tensor representations 111
- Tietze–Urysohn extension theorem 36
- topology 29
- topological complement 189
- topological direct sum 189
- topological space 28
- traceless 109
- trapezoid formula 176
- Triebel–Lizorkin spaces 360
- turbulence 330

- uniform boundedness theorem 172

- value of the game 81
- variation 43, 135
- variational inequality 66

- variational problem 46
- vector calculus 363
- velocity fields 335
- very weak pressure equation 367
- Virasoro algebra 161
- Virasoro charges 161
- Virasoro constraints 161

- weak compactness 222
- weak convergence 49, 217
- weak topology 221
- weak* convergence 218
- weakly coercive 53
- weakly open 221
- weakly sequentially continuous 53
- weakly sequentially lower semicontinuous 53
- Weierstrass approximation theorem 35
- Weierstrass existence theorem 53
- Weierstrass theorem 33
- weight diagram 117
- well-posedness principle 180
- winding number 152
- world sheet 157

- Yang–Mills equation 125
- Young inequality 352

- Zorn lemma 3

This is the second part of an elementary textbook which combines linear functional analysis, nonlinear functional analysis, numerical functional analysis, and their substantial applications with each other. The book addresses undergraduate students and beginning graduate students of mathematics, physics, and engineering who want to learn how functional analysis elegantly solves mathematical problems which relate to our real world and which play an important role in the history of mathematics. The book's approach begins with the question "what are the most important applications" and proceeds to try to answer this question. The applications concern integral equations, differential equations, bifurcation theory, the moment problem, Čebyšev approximation, the optimal control of rockets, game theory, symmetries and conservation laws (the Noether theorem), the quark model, and gauge theory in elementary particle physics. The presentation is self-contained. As for prerequisites, the reader should be familiar with some basic facts of calculus.

The first part of this textbook has been published under the title *Applied Functional Analysis: Applications to Mathematical Physics*

ISBN 0-387-94422-2



A standard 1D barcode representing the ISBN number 0-387-94422-2. The barcode is composed of vertical black lines of varying widths on a white background.

9 780387 944227 >

ISBN 0-387-94422-2